Prediction of Orbital Parameters for Undiscovered Potentially Hazardous Asteroids Using Machine Learning

Vadym Pasko

Abstract The purpose of this study is to make a prediction of combinations of orbital parameters for yet undiscovered potentially hazardous asteroids (PHAs) with the use of machine learning algorithms. The proposed approach aims at outlining subgroups of all major groups of near-Earth asteroids (NEAs) with high concentration of PHAs in them. The approach is designed to obtain meaningful results and easy-understandable boundaries of the PHA subgroups in 2- and 3-dimensional subspaces of orbital parameters. Boundaries of these PHA subgroups were found mainly by the use of Support Vector Machines algorithm with RBF kernel. Additional datasets of virtual asteroids were generated to handle sufficient amount of training and test data, as well as to emulate undiscovered asteroids. This synthetic data helped in revealing 'XX'-shaped region with high concentration of PHAs in the (ω, q) plane. Boundaries of this region were used to split all NEAs into several domains. For each domain the subgroups of PHAs were outlined in different subspaces of orbital parameters. Extracted subgroups have high PHA purity (∼90%) and contain ∼90% of all real and virtual PHAs. Obtained results can be useful for planning future PHA discovery surveys or asteroid-hunting space missions.

1 Introduction

The increasing rate of asteroid discovery has a limit connected with constraints of using ground-based telescopes for near-Earth asteroid (NEA) observations [1, 2]. Discovery of asteroids with space-based telescopes [3–5], being a promising solution, is currently limited by the low number of hardware launched to date. Development of new asteroid-hunting tools, both ground-based and space-based, requires sophisticated analysis of asteroid trajectories, which can be used for

V. Pasko (✉)
Yuzhnoye State Design Office, Dnipro, Ukraine
e-mail: mail@vadym-pasko.com; keenon3d@gmail.com

M. Vasile et al. (eds.), *Stardust Final Conference*, Astrophysics and Space Science Proceedings 52, https://doi.org/10.1007/978-3-319-69956-1_3

efficient survey planning, softening requirements to space-based hardware, and thus reducing the total cost of future asteroid-hunting space missions.

A list of works aimed at revealing true size frequency and orbital distributions of all existing NEAs ([6–10] and others) is growing with an increasing pace. The techniques typically adopted to meet these goals include the characterization of the detection efficiency of a reference survey and subsequent simulated detection of a synthetic population, or the statistical tracking of NEAs from their source regions in the main belt to the inner solar system and subsequent comparison to the detections by a reference survey, or the combination of these two approaches [10].

While each new approach of finding best debiased estimate for the whole NEA population is essential for better understanding the amount of yet undiscovered NEAs of particular sizes [21], successful planning of asteroid surveys requires more thorough analyses, capable of predicting orbital distributions of the NEA subpopulations. These subpopulations or groups[1], namely Atiras, Atens, Apollos and Amors contain asteroids that cross Earth orbit at small distances (less than 0.05 AU) which makes them objects of the top interest, since they may evolve into potential impactors within the foreseeable future. Such objects are known as potentially hazardous asteroids (PHAs) and are defined as asteroids with an Earth Minimum Orbit Intersection Distance (MOID) of 0.05 AU or less and an absolute magnitude $H = 22.0$ ($\sim$140 m in diameter) or less. The estimation of PHA distributions and prediction of orbital parameters for undiscovered PHAs is placed in the center of the current research.

An effort of analyzing orbital distributions of PHAs has already been made by Mainzer et al. [9], where authors estimated the entire population of PHAs larger than 100 m. In this regard they extended the definition of PHA to include objects with diameters down to 100 m. In the current study the limit on asteroid size is omitted from the PHA definition. This is done with consideration that even small objects (30–50 m or even smaller) can cause major regional damage in the event of an Earth impact [5]. And since the current research is focused on the introduction of a different method of the PHA orbital distribution analysis, the issue of the survey biases that put the limit of 100 m in the work of Mainzer et al. [9] has been left behind the scene, but will be incorporated in the future work. Thus, here and after, any close-approaching asteroid with MOID <0.05 AU is referred as PHA.

While Mainzer et al. [9] treated PHAs as a separate subpopulation of NEAs, in the current research PHAs are examined with respect to each NEA group listed above. The approach of extracting and analyzing smaller subpopulations of NEAs allows to reveal hidden peculiarities in their size and orbital distributions specific to these particular subpopulations, and thus obtain new insights that can be used to build more accurate models. The examples of such analyses includes works of Granvik et al. [11] and Fedorets et al. [12], where authors calculated the population characteristics of the Earth's irregular natural satellites (NES) that are temporarily captured from the NEA population.

[1] http://neo.jpl.nasa.gov/neo/groups.html

Being devoted to the analysis of orbital distributions of PHAs inside groups of NEAs, the approach presented in the current paper differs from the common statistical treatment of NEAs and relies mainly on the application of machine learning techniques. It is aimed at revealing correlations between orbital parameters of PHAs that help to outline subgroups inside each group of NEAs with high concentration of PHAs in them. In order to obtain better insight into these correlations the approach incorporates generation of virtual asteroids using simplified models of orbital distributions of all known NEAs.

The majority of related works provide estimations of NEA orbital distributions with regard to reduced subsets of orbital parameters, considering uniform distribution of NEAs by the argument of perihelion and longitude of the ascending node. In the current work all 5 parameters that define heliocentric orbit were taken into consideration, which has been justified be revealing dependency of the PHA orbital distribution on the argument of perihelion and longitude of the ascending node for the Amors and non-uniform distribution of all observed NEAs by the last parameter. Thus, the correlation analyses were performed for pairs of 5 orbital parameters, namely: semi-major axis (a), perihelion distance (q), inclination (i), argument of perihelion (ω) and longitude of the ascending node (Ω).

Considering strict dependence of MOID on these 5 orbital parameters, we can assume that there exists a boundary in the space of these parameters that divides PHAs from non-hazardous asteroids (NHAs). This boundary is essentially a surface of the hypersolid that encompasses all possible combinations of orbital parameters that define PHA in our formulation. In order to obtain meaningful results the process of outlining populations of PHAs, presented here, is based on the application of several consequent cuts of this hypersolid by finding boundaries in low-dimensional projections of NEA orbital distributions (2D and 3D). This approach allows to obtain regions (or subgroups) of high PHA purity ($\sim$90%) that together contain $\sim$90% of all existing and hypothetical PHAs. The ensemble of these regions provides a unique insight into the possible residences of yet undiscovered PHAs, which, as believed by the author, can facilitate future discoveries of PHAs.

The structure of the paper is as follows. Section 2 is devoted to a brief overview of the machine learning and its applications in solving modern problems of astronomy. This section also contains brief descriptions of two machine learning algorithms used in the current work. In Sect. 3 we'll dive into the analysis of NEAs' orbital parameters and their correlations for PHAs. Here we will generate virtual asteroids and observe an interesting structure of PHAs in the (ω, q) projection. In the Sect. 4 we will split all NEAs into 4 domains. Section 5 is devoted to the divisions of PHAs from NHAs inside each domain and outlining PHA subgroups. At the end of this section we'll make a summary of the divisions' qualities and purities of the PHA subgroups.

2 Machine Learning in Astronomy

Since the advent of astrophotography and spectroscopy over a century ago, astronomers have faced the challenge of characterizing and understanding vast numbers of asteroids, stars, galaxies and other cosmic populations [13].

Various mathematical methods were invented and applied for interpreting astronomical data. A long way from the least-squares and maximum likelihood to the inverse probability and Bayesian methods have led to a rapid expansion in the diversity of numerical methods that are used nowadays by astronomers for data analysis.

Since the middle of the last century statistical modeling has become a crucial component in building an inference from incomplete observational data. But the continuously growing size, sources and diversity in spectrum of the astronomical data has led to the need of applying different techniques. Over the past two decades a significant progress in data analysis and, particularly in image processing, has been achieved with the use of machine learning, which is essentially the study of software that learns from experience. The machine learning approach is rather new for astronomy but has a great potential in bringing brand new inference to old problems as well as in discovering new dependencies in structure and behavior of celestial objects.

Machine learning is a method of data analysis, aimed at finding hidden insights by the means of using algorithms that iteratively learn from data rather than being explicitly programmed where to look. Machine learning can appear in many guises. But two types of problems in machine learning—classification and clustering refer to the most common problems in astronomy.

2.1 Classification and Support Vector Machines

Classification (or supervised learning) is a classic problem that goes back a few decades. In supervised learning a training set of examples with the correct responses (targets) are provided and, based on this training set, the algorithm generalizes to respond correctly to all possible inputs. A wide range of supervised learning algorithms includes but not limited to k-Nearest Neighbors, Decision Trees, Random Forests and Support Vector Machines (SVM). A tangible difference between all of them is the shape of the decision boundary that the algorithm can learn (to split data of different classes). In a classification problem with two classes, a decision boundary or decision surface is a hypersurface that partitions the underlying vector space (feature space) into two sets, one for each class (Fig. 1).

In high-dimensional spaces, data can more easily be separated linearly and the simplicity of classifiers such as linear SVMs might lead to better generalization than is achieved by other classification algorithms [14]. On the other hand, in low-dimensional spaces smooth nonlinear decision boundaries can provide not only

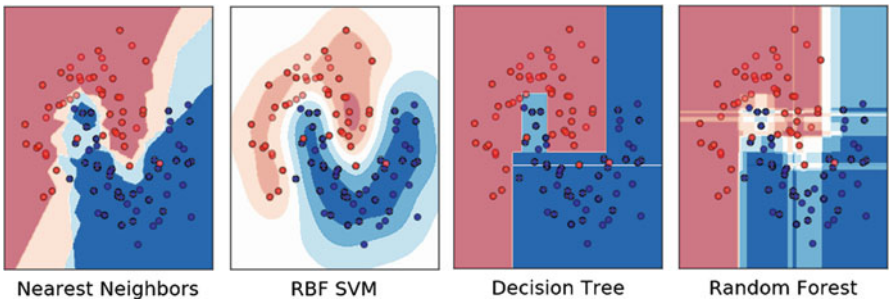

Fig. 1 Decision boundaries produced by different classification algorithms [14]

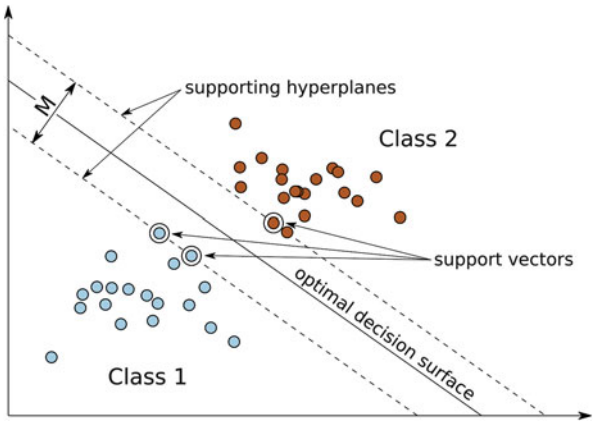

Fig. 2 Maximum margin separation of two linearly-separable classes

better classification accuracy, but a simpler insight. Nonlinear decision boundaries with controllable smoothness can be provided by Support Vector Machines algorithms that are usually applied in cases of overlapping classes. The core operational principle of all SVM algorithms is to find a hyperplane (or hypersurface in nonlinear case) that separates the feature space with the maximum margin (Fig. 2).

The data samples closest to the decision hyperplane are known as support vectors [15]. In the two-class problems the support vectors define two parallel supporting hyperplanes equidistant from the decision hyperplane. Distance between these planes is essentially the margin M that has to be minimized.

The SVMs with non-linear kernel functions produce nonlinear boundaries by constructing a linear boundary in a large, transformed version of the feature space. One of the widely-used nonlinear kernel functions is the Gaussian radial basis function (RBF). Particularly the SVM with RBF kernel is the main tool in the current study.

Selection of the appropriate kernel is not the only way of getting desired behavior of the decision boundary. In some cases there might be a need to push the decision boundary towards one of the data classes. Such a need may arise when one of the

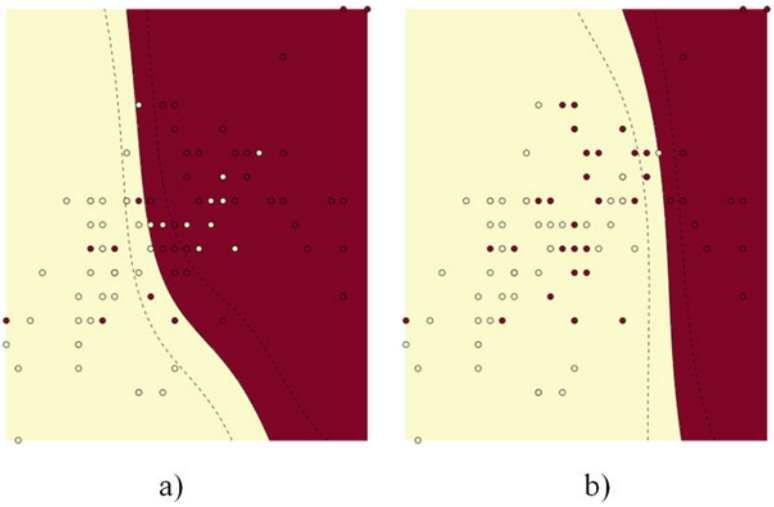

Fig. 3 Decision boundary produced by the RBF SVM for the same training data and different class weights: (**a**)—no class weights assigned, (**b**)—class weight of the yellow (light) samples is higher

classes is more important than the other, or if we want to increase the purity of one class sacrificing the purity of another. This can be achieved by manipulating the class weights (Fig. 3). We will use the trick with assigning different class weights several times in the current work in order to achieve desired classification purity.

Still there are much more options to tune SVM. That is the reason why SVM is a popular choice for classification.

The computational complexity of the SVM algorithms consists of the kernel complexity, which is $O(m^2\ n)$ for the RBF kernel, and the factorization complexity, which is $O(m^3)$ in general [16]. Here m is the number of data points (samples) and n is the dimensionality, or in other words—number of features. This is why the SVM is very expensive to use for large datasets. And this is exactly what we will do in the current work. Thus, some classification operations described below require significant computation time, and, depending on the processor and operating system may take up to 1 h or even more.

2.2 Clustering and DBSCAN Algorithm

When working with large datasets it is in most scenarios useful to be able to break data into several groups (clusters) and eventually, to do class identification. This objective can be efficiently achieved with the use of clustering techniques, which are applied to search for groupings of multivariate data points by proximity of objects or other criteria. This task is also known in the world of machine learning as unsupervised learning.

In the unsupervised learning correct responses are not provided, instead the algorithm tries to identify similarities between the inputs so that inputs that have something in common are categorized together in clusters [16].

The wide range of clustering algorithms includes k-means, k-medoids, SNN, MCLUST and many others [17]. In the current work, in order to perform efficient density-based clustering, we will use DBSCAN (Density Based Spatial Clustering of Applications with Noise) algorithm.

The DBSCAN algorithm can identify clusters by looking at the local density of data points [18]. DBSCAN can find clusters of arbitrary shape. However, clusters that lie close to each other tend to belong to the same class.

The central component of the DBSCAN is the concept of core samples, which are samples that are in areas of high density. A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples). There are two parameters to the algorithm, μ—minimal number of samples to form core and ε—radius of the core sample neighborhood. Higher μ or lower ε indicate higher density necessary to form a cluster [14].

More formally, core sample is a sample in the dataset such that there exist μ other samples within a distance of ε, which are defined as neighbors of the core sample. This ensures that the core sample is in a dense area of the vector space. A cluster is a set of core samples that can be built by recursively taking a core sample, finding all of its neighbors that are core samples, finding all of their neighbors that are core samples, and so on. A cluster also has a set of non-core samples, which are samples that are neighbors of a core sample in the cluster but are not themselves core samples. Intuitively, these samples are on the fringes of a cluster.

Any core sample is part of a cluster, by definition. Any sample that is not a core sample, and is distant from any core sample at least at ε, is considered an outlier by the algorithm (Fig. 4).

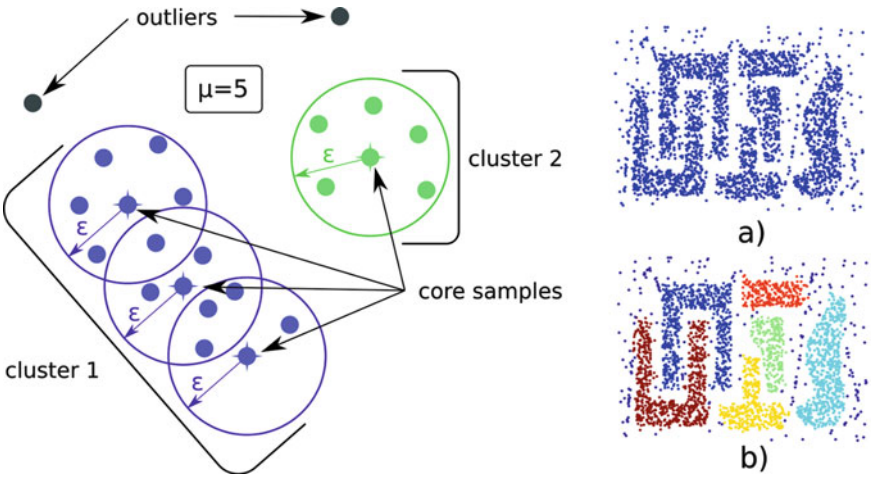

Fig. 4 Formation of clusters with DBSCAN. On the right: (**a**)—original points, (**b**)—clusters

3 Analysis of NEAs' Orbital Distributions

The typical application of machine learning algorithms in astronomy lies in the area of image processing. Automated classification tools have become increasingly useful with the growth of megadatasets in astronomy, often from wide-field surveys of the optical sky. Just as an example, the most elementary need is to discriminate galaxies, which are typically resolved blurry objects, from stars, which are unresolved [13].

Nevertheless, it is far not the only possible application of machine learning techniques in astronomy. Particularly, the interdisciplinary problem of detection, classification and orbit determination of near-Earth asteroids can be solved more efficiently using a unique inference that can be provided by the deep analysis of the existing asteroid database.

Asteroid database that we use in the current work counts over 600,000 asteroids including 14,858 NEAs, and was compiled from the Jet Propulsion Laboratory's Small Body Database and Minor Planet Center by Ian Webster—developer of the Asterank[2] (a web service for ranking asteroids by mining profit). All NEAs present in the database were split into two subsets—PHAs and NHAs by the threshold value of MOID (0.05 AU). Distributions of asteroids were analyzed for all possible combinations of two orbital parameters separately for PHAs and NHAs.

Particularly, orbital distribution of NEAs in the (ω, q) plane reveals correlation for PHAs, that gather into the M-shaped structure (Fig. 5). A similar structure has

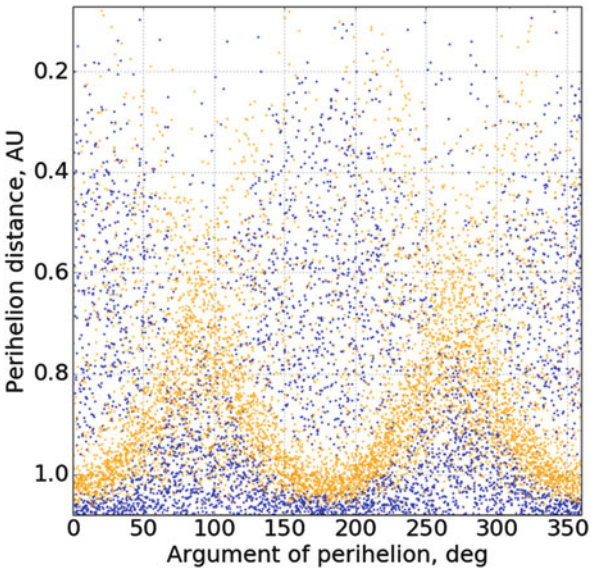

Fig. 5 Correlation between two orbital parameters for PHAs. On the figure orange (light) dots—PHAs, blue (dark) dots—NHAs

[2] http://www.asterank.com/

Table 1 Groups of NEAs

Group name	Definition	Population	Relative population (%)
Atiras	a < 1.0 AU, Q < 0.983 AU	16	0.1
Atens	a < 1.0 AU, Q > 0.983 AU	1087	7
Apollos	a > 1.0 AU, q < 1.017 AU	7968	54
Amors	a > 1.0 AU, 1.017 < q < 1.3 AU	5774	38.9

Q stands for aphelion distance

already been shown in the work of Gronchi and Valsecchi [19], where authors provided a nice explanation of the M-shaped structure for the faint asteroids and its dependency on the orbit distance.

Other pairs of orbital parameters don't provide more distinctive separation of PHAs from NHAs, so this pattern will serve us as a starting point for further divisions.

3.1 Virtual Asteroids

All NEAs, which are essentially asteroids with perihelion distance q < 1.3 AU, are represented by 4 groups[3] with different populations (Table 1). The quality of the analysis with machine learning depends on the number of samples and for small groups like Atiras and Atens it is highly desirable to get more data. More of that, even in the case of more numerous groups, additional data can reveal some yet unseen patterns.

Luckily our response parameter (0 for NHAs and 1 for PHAs) explicitly depends on the input parameters so, we can synthetically increase amount of data by generating virtual asteroids and computing MOID for them. In order to get better insight we will generate two additional datasets of asteroids with different distributions of orbital parameters: one with uniform distribution and another with distributions that approximate distributions of real NEAs. These additional datasets will be referred as uniform and non-uniform respectively.

Despite the constraints on the possible combinations of semi-major axis and perihelion distance for elliptical orbits, we will make an assumption that they are independent and will fix failed generated orbits (with negative eccentricity) by regenerating them. This simple iterative approach will help us preserve physical sense of generated orbits without increasing complexity of the process.

In the case of the non-uniform dataset, first we need to find continuous distributions that approximate distributions of real NEAs. Then virtual asteroids can be generated using these approximations.

[3] http://neo.jpl.nasa.gov/neo/groups.html

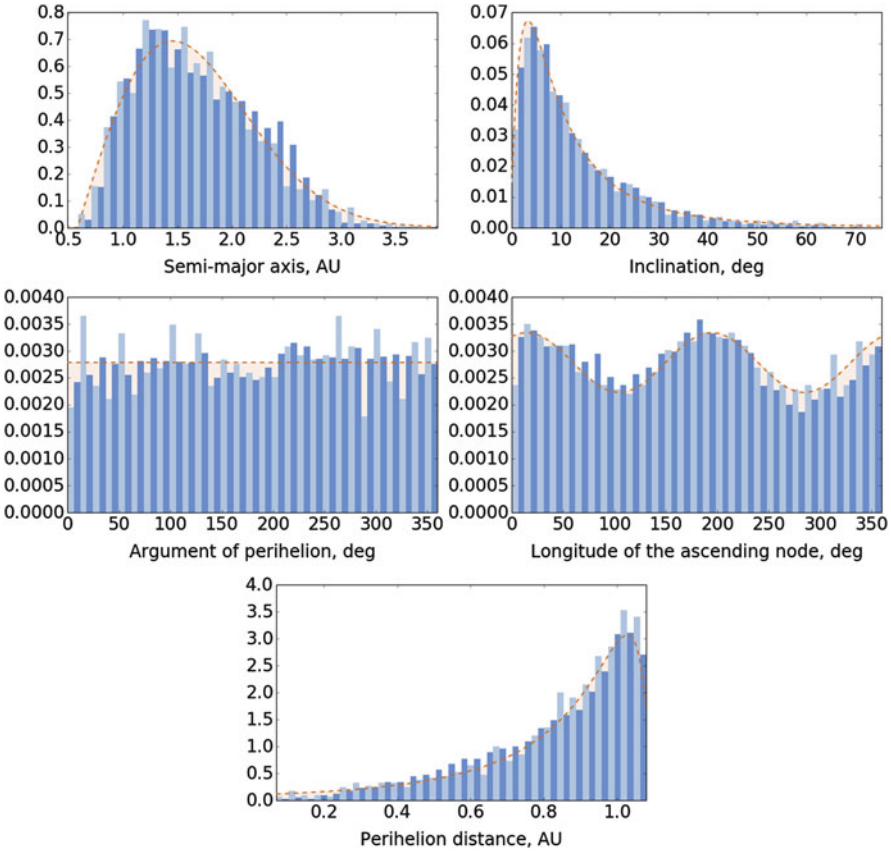

Fig. 6 Fitted continuous distributions of orbital parameters for NEAs. Dark bars—distribution density of real asteroids, light bars—same for virtual asteroids generated using continuous distributions (red dashed lines). Next models were used for approximations: semi-major axis—Rayleigh distribution; inclination—log-normal distribution; argument of perihelion—uniform distribution; longitude of the ascending node—harmonic distribution; perihelion distance—Johnson's SU-distribution

A rich set of probability distributions (81 continuous distributions) embedded in the SciPy library [20] along with the tools for curve-fitting enables a quick and efficient selection of the best models (Fig. 6).

So, by these means additional datasets of 30,000 uniform and 200,000 non-uniform virtual NEAs were generated. The amounts of virtual asteroids were selected from next considerations: to obtain uniformly filled space of orbital parameters without significant increase in density; and to obtain distributions similar to real asteroids but with more than 10 times higher density.

In order to find out PHAs amongst generated asteroids we need to calculate MOID for them. This can be achieved by the use of numerical optimization algorithms. We will use four initial guess points in the minimization problem:

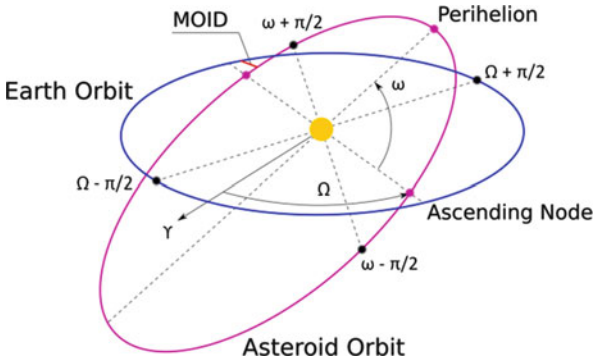

Fig. 7 Positions of the initial guess points on the Earth and asteroid orbits for calculating minimal distances

two opposite points on the Earth ellipse, shifted by $\pm\pi/2$ from the direction to the ascending node of the asteroid orbit; and two opposite points on the asteroid ellipse, shifted by $\pm\pi/2$ from the perihelion (Fig. 7). Therefore, we obtain 4 possible combinations for the pairs of initial points, and by iteratively altering their positions on the ellipses will find 4 minimal distances between them. Finally, by definition, MOID is a minimal of four obtained distances.

We use a downhill simplex optimization algorithm to find minimal distances between orbits. This method is commonly applied to find the minimum or maximum of an objective function in a multidimensional space and is embedded in the SciPy library [20] that we use. In our case it's a two-dimensional problem. Computation of MOID has taken 200 s for the uniform dataset and 1447 s for the non-uniform dataset, while running in parallel in 3 threads (CPU Intel Core i7-4510U 2.00GHz×2) on a 64-bit OS Linux Mint 17.2.

After calculating MOID for all virtual asteroids we can split them into the PHA and NHA datasets and take a better look on the distributions of asteroids in the (ω, q) plane. Distributions are shown on the Fig. 7 separately for PHAs (right) and NHAs (left) because of the high distribution density.

Now we are able to see new features of the M-shaped structure of PHAs in a (ω, q) plane. Particularly "upper branches" become visible for both datasets and 'M' morphs to a 'XX' shape, which is better distinguished for the uniform dataset. Areas of extreme PHA purity emerged in the neighborhood of $q = 1\ AU$.

The most obvious inference we get from these pictures is that a large part of PHAs is located in the thin 'XX'-shaped belt. Others are mixed with NHAs and, probably can be separated in other dimensions. According to this hypothesis we will try to split all NEAs into several domains neighboring to the 'XX' structure:

1st domain—what lies under the 'XX' structure;
2nd domain—what is on the left, right and in-between the 'XX' structure;
3rd domain—what is above the 'XX' structure;
4th domain—the 'XX' structure by itself.

4 Finding Boundaries of the NEA Domains

In order to preserve obvious symmetry of the asteroid distributions along the ω axis we will extend generated datasets with symmetric mirrors over vertical planes that cross $\omega = 90°$ and $\omega = 180°$. This operation will increase density of data points and, thus, significantly increase computations time.

As the shape of the 'XX' structure is more accurate for the uniform dataset (Fig. 8), we will use it as the basis for extraction of clusters that represent 4 defined domains. The 1st, 2nd and 3rd domains can be extracted from the dataset of the uniform NHAs by using DBSCAN clustering algorithm. The 4th domain can be

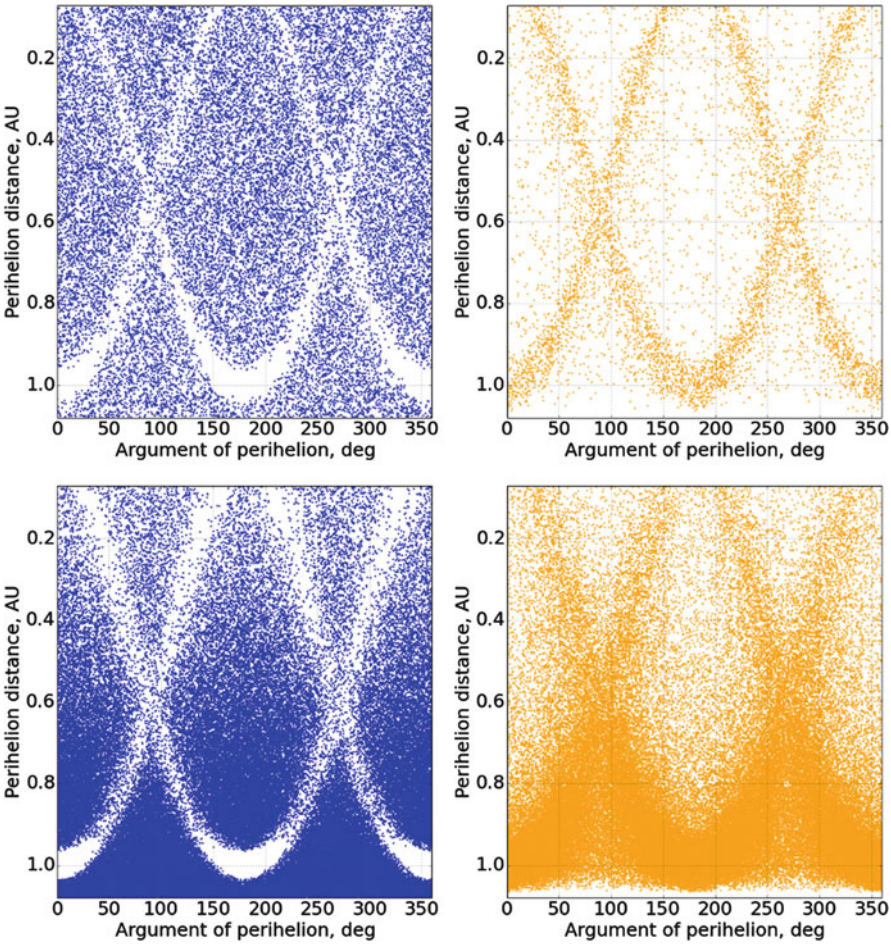

Fig. 8 Distribution of virtual asteroids in the (ω, q) plane. On the left (blue dots)—NHAs, on the right (orange dots)—PHAs. On the top—uniform virtual asteroids, on the bottom—non-uniform virtual asteroids

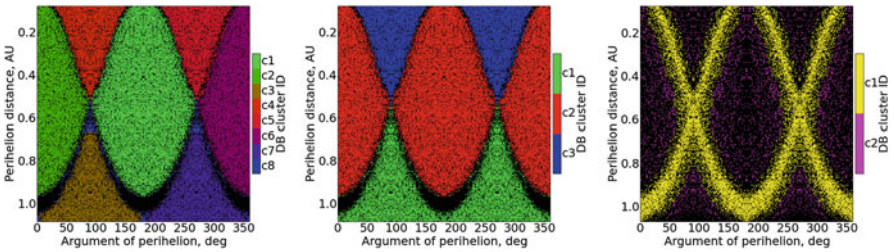

Fig. 9 Clusters found with DBSCAN. On the left—original 8 clusters (8th is a collection of outliers) found by DBSCAN in the uniform virtual dataset of NHAs; in the center—manually rearranged and merged clusters; on the right—original 'XX' cluster found by DBSCAN in the uniform virtual dataset of PHAs (second cluster is a collection of outliers)

easily extracted by the same means from the dataset of uniform PHAs. As we are using DBSCAN we can't explicitly control the number of generated clusters, what can be done with other clustering algorithms like k-means. So, selection of different values for μ and ε can lead to different number of clusters found.

The selected compromise values of $\mu = 105$ and $\varepsilon = 0.022$ provide us with 8 clusters of NHAs instead of desired 3 (Fig. 9 left). But this can be easily fixed by splitting and merging clusters (Fig. 9 center). What is more important is that we obtain clusters of the desired shape.

Three clusters from the NHA dataset overlap with the 'XX' cluster from the PHA dataset (Fig. 9 right). In order to find a smooth boundary between them we will apply SVM algorithm with RBF kernel. The SVM algorithm implemented in the scikit-learn package can handle multiclass classifications, so we will use cluster IDs as a class reference to find boundaries between 4 desired domains.

The implementation of RBF SVM in the scikit-learn uses two input parameters to control the decision surface shape: C and γ. Intuitively, the γ parameter defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. The γ parameters can be seen as the inverse of the radius of influence of samples selected by the model as support vectors. The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors [14].

So after a series of trials the SVM with values of $C = 10{,}000$ and $\gamma = 6$ has been selected. It produces smooth boundaries between domains (Fig. 10 right).

Now we can use trained SVM to 'predict' domain membership of any NEA. By passing dataset of real NEAs we can estimate PHA purities of the obtained domains (Fig. 10 right).

It is quite remarkable, that the estimated purity of the 4th domain for real NEAs reaches 0.93. Purities of other domains are significantly lower. In the next sections we will try to separate PHAs from NHAs in these domains in different dimensions and will use the non-uniform dataset of virtual asteroids as a training data for SVMs.

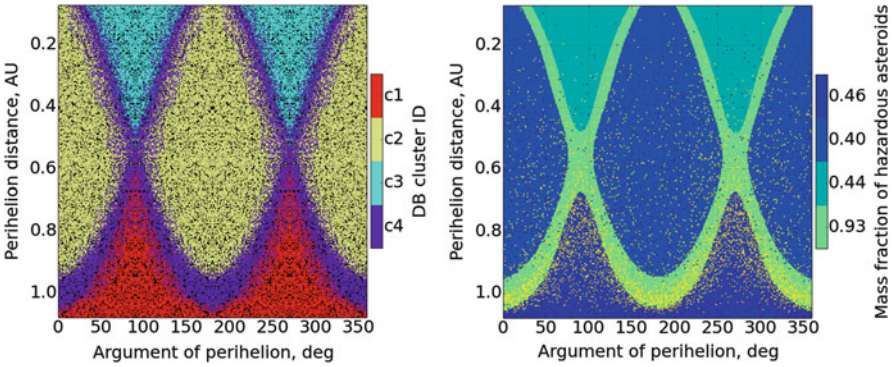

Fig. 10 Classification of NEAs by cluster IDs. On the left—clusters extracted with DBSCAN; on the right—domains outlined by RBF SVM and distribution of real NEAs plotted over (yellow dots—PHAs, blue dots—NHAs)

5 Extraction of PHA Subgroups From NEA Domains

In each domain we will find representatives of the NEA groups (Table 1) and work with them separately (except Atiras and Atens). This approach was proven to be the most successful by numerous failures in trying different strategies. Thus, we will extract subgroups with high PHA purity for each group of virtual non-uniform NEAs.

The representatives of the Amor group are present only in the 1st of 3 'XX'-neighboring domains. Asteroids of other groups are present in all 3 domains.

In some cases we will split PHAs from NHAs linearly, but in most cases we will use RBF SVM to find decision surfaces in 2- and 3-dimensional projections of asteroids' orbital distributions. The summary on the qualities of the divisions made in each domain is represented at the end of the section.

5.1 Domain #1

The Atiras & Atens in the 1st domain can be easily separated by the SVM with a linear kernel in the (a, i) plane (Fig. 11).

Two divisions were applied for the Amors. First—in the (ω, i) plane to separate most part of NHAs. Second decision surface was learned by the RBF SVM in the (ω, Ω, q) space for those classified as PHAs in the first division (Fig. 12). It separates only a half of PHAs with sufficient purity. The dependence of the PHA distribution on the longitude of the ascending node can be explained by the eccentricity of the Earth's orbit and its influence on the values of MOID for asteroids with outer orbits.

Two consequent divisions by RBF SVMs were applied for the Apollos: first in the (ω, q, i) space and the second in the (ω, q, a) space (Fig. 13). Second surface covers most part of PHAs left above the first surface. SVM parameters are presented in the Table 2.

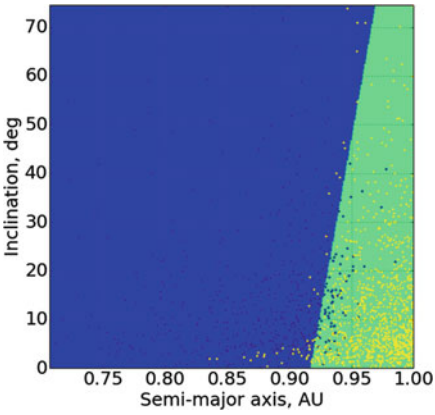

Fig. 11 Division of the Atiras & Atens into the PHA and NHA regions with linear SVM. Green (light) area—PHA region, blue (dark) area—NHA region; yellow dots—PHAs, blue dots—NHAs

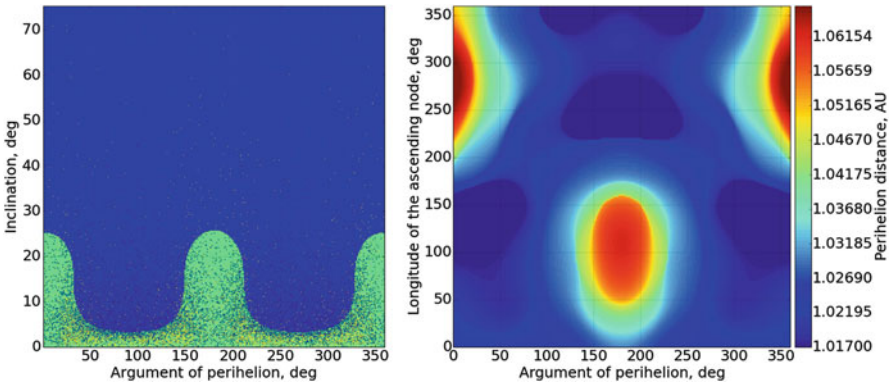

Fig. 12 Decision surfaces between PHAs and NHAs for the Amor group. The surface on the right covers approximately a half of PHAs from the PHA region (green) on the left

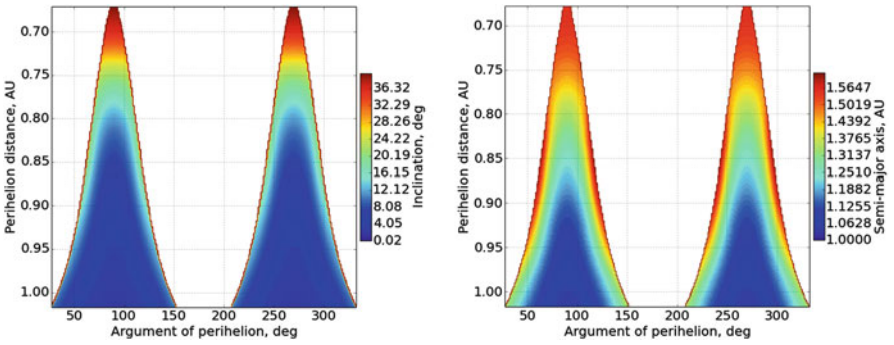

Fig. 13 Two decision surfaces for Apollos. PHA regions are below the surfaces. The second surface (on the right) covers PHAs that are above the first surface (on the left)

Table 2 SVM parameters for divisions in the 1st domain

NEA group	Division	Space	Kernel	γ	C	Class weight
Atiras & Atens	1	(a, i)	Linear	–	1	Equal
Amors	1	(ω, Ω, q)	RBF	20	8	NHA: 2.4
Apollos	1	(ω, q, i)	RBF	40	0.05	NHA: 1.2
	2	(ω, q, a)	RBF	40	0.1	NHA: 1.5

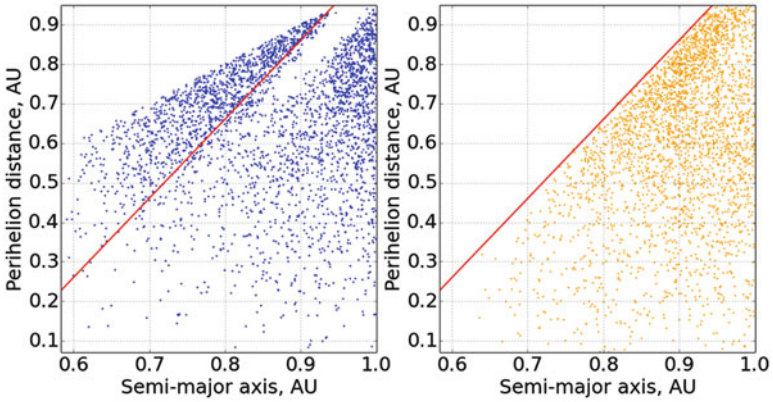

Fig. 14 Atiras & Atens of the 2nd domain. Red line—NHA division plane

5.2 Domain #2

A piece of NHAs of Atiras & Atens in the 2nd domain can be linearly separated from other asteroids. Despite the simplicity of such operation, it has turned out to be a challenging task for the linear SVM. So we make the section manually.

PHAs and NHAs below the red line on the Fig. 14 can be divided by a complex surface in the (ω, a, i) space, produced by the RBF SVM (Fig. 15 left). A small portion of asteroids misclassified by this division gather into a strap in the (a, q) plane. This strap can be outlined by another RBF SVM (Fig. 15 right).

The Apollo asteroids in the 2nd domain can be efficiently divided into PHA and NHA regions by two consequent splits with RBF SVMs—the first in the (ω, i) plane and the second for those left after the first split—in the (ω, q, i) space (Fig. 16). SVM parameters are presented in the Table 3.

5.3 Domain #3

The 3rd domain includes Atiras, Atens and Apollos. Atiras & Atens can be easily divided into PHAs and NHAs by applying single split in the (ω, i) plane with RBF SVM (Fig. 17 left). The boundary between PHAs and NHAs for the Apollo

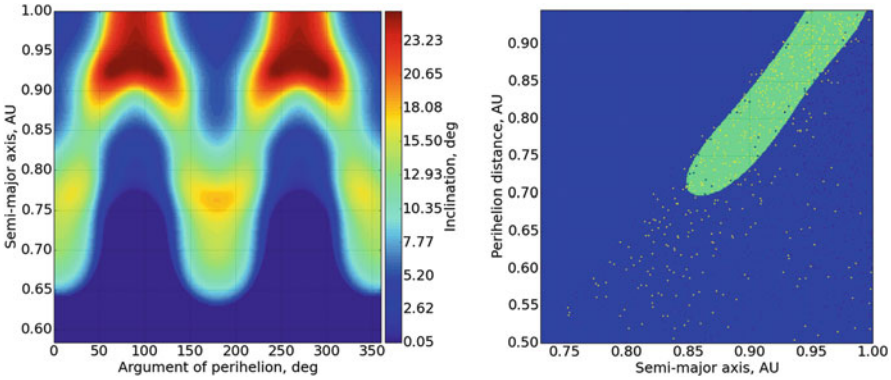

Fig. 15 PHA regions found by the RBF SVMs for Atiras & Atens

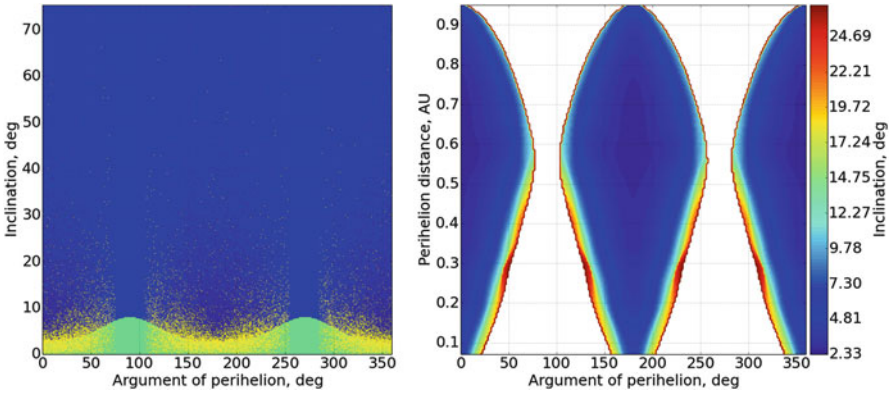

Fig. 16 Two divisions of Apollos. On the right—PHA region is filled with green (light), yellow dots are PHAs and blue dots are NHAs. The surfaces on the left cover PHAs left in the NHA region from the first split (blue area on the left)

Table 3 SVM parameters for divisions in the 2nd domain

NEA group	Division	Space	Kernel	γ	C	Class weight
Atiras & Atens	1	(ω, a, i)	RBF	80	0.1	NHA: 1.5
	2	(a, q)	RBF	8	1000	NHA: 1.5
Apollos	1	(ω, i)	RBF	30	0.1	NHA: 10
	2	(ω, q, i)	RBF	100	2	NHA: 1.5

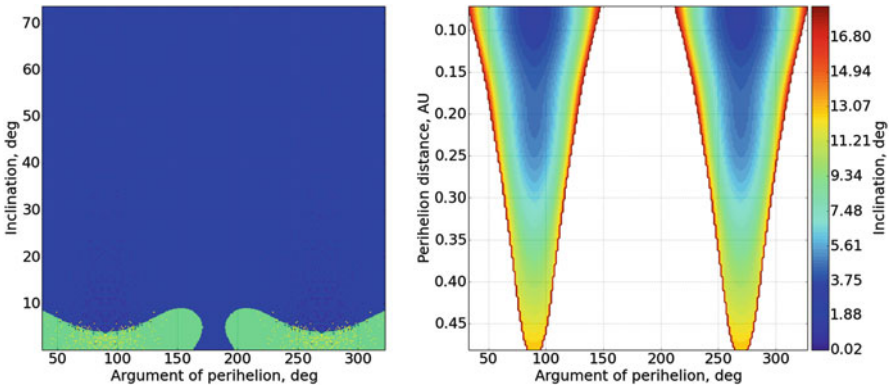

Fig. 17 PHA regions found by RBF SVMs for Atiras & Atens (green area on the left) and Apollos (space under the surface on the right)

Table 4 SVM parameters for divisions in the 3rd domain

NEA group	Division	Space	Kernel	γ	C	Class weight
Atiras & Atens	1	(ω, i)	RBF	80	0.4	NHA: 1.1
Apollos	1	(ω, q, i)	RBF	20	0.5	Equal

asteroids resides in the (ω, q, i) space (Fig. 17 right). SVM parameters are presented in the Table 4.

5.4 Assessment of the Divisions' Qualities

The worst-separable population of asteroids is the group of Amors in the 1st domain. Only 42% of virtual PHAs were separated by the surface in the (ω, Ω, q) space. Other hard-separable populations of asteroids belong to the 2nd domain. In other cases the fraction of correctly classified PHAs is close to 90%.

Most part of divisions was made by training SVM with RBF kernel, while in some cases a linear kernel was used, and once a manual linear separation was made. The non-uniform dataset of virtual asteroids was used for training SVMs. This dataset along with the dataset of real asteroids were used to estimate qualities of the divisions and purities of the PHA subgroups. The summary is depicted in the Table 5.

6 Conclusions

Generation of virtual asteroids and analysis of their orbital distributions revealed a new shape of the known 'M' structure of PHAs in the (ω, q) plane, which morphs into the 'XX' structure with the increase of samples. This 'XX' structure contains

Table 5 Divisions summary

Domain	Groups of NEAs	GW	N	PHA fraction	NHA fraction	PHA purity
1	Atiras & Atens	0.02 (0.01)	1	0.96 (0.97)	0.15 (0)	0.92 (1)
	Apollos	0.63 (0.69)	2	0.93 (0.97)	0.08 (0.12)	0.91 (0.93)
	Amors	0.35 (0.3)	1	0.42 (0.48)	0.01	0.88 (0.91)
	All NEAs	**1**	**4**	**0.86 (0.9)**	**0.05 (0.06)**	**0.91 (0.93)**
2	Atiras & Atens	0.09 (0.3)	3	0.86 (0.78)	0.1	0.9 (0.93)
	Apollos	0.91 (0.7)	2	0.88 (0.83)	0.02	0.95 (0.96)
	All NEAs	**1**	**5**	**0.88 (0.81)**	**0.03**	**0.94 (0.95)**
3	Atiras & Atens	0.11 (0.22)	1	0.87 (1)	0.03 (0)	0.9 (1)
	Apollos	0.89 (0.78)	1	0.92	0.05 (0.03)	0.94 (0.97)
	All NEAs	**1**	**2**	**0.92**	**0.05 (0.02)**	**0.94 (0.97)**
4	**All NEAs**	**1**	**0**	**1**	**1**	**0.9 (0.93)**
Total			**11**	**0.93**	**0.09 (0.1)**	**0.91 (0.93)**

GW stands for a group weight and N for the number of divisions. The fraction of correctly classified PHAs (PHA fraction) was calculated with regard to the total number of PHAs in a group or domain. Same is applied to the fraction of misclassified NHAs (NHA fraction). PHA purity represents a cumulative purity of all PHA regions outlined for a group or domain. Values without parentheses correspond to the virtual non-uniform NEAs and values in parentheses correspond to the real NEAs. If a value for real NEAs is not provided it is the same as for virtual NEAs

approximately a half of all real and virtual PHAs and its purity is around 0.9 (0.93 for real NEAs). The boundaries of the 'XX' structure, found with the application of the Support Vector Machines algorithm, divide all NEAs into 3 domains, while the 4-th domain is the region inside the 'XX'.

Other domains were analyzed in details and the effort of outlining 2- or 3-dimensional regions with high PHA concentration in them was more or less successful. The representatives of the main NEA groups were analyzed separately in each of 3 domains. In most cases separations of PHAs from NHAs were effective, producing PHA subgroups with high purity.

The analysis of PHAs in the Amor group revealed the dependency of PHA distribution density on the argument of perihelion and longitude of the ascending node. This may be explained by the eccentricity of the Earth's orbit and will be covered in more details in the future work.

The summary of all divisions (Table 5) shows that the proposed approach allows to group over 90% of all real and virtual PHAs into regions with ∼90% purity. This essentially means that dominant part of all yet undiscovered PHAs resides in these regions. While the knowledge of the shapes of PHA regions can be useful for planning future PHA discovery surveys and future asteroid-hunting space missions, yet some work has to be done to verify obtained results. Particularly, the original dataset of NEAs contains survey biases that may influence the shape of obtained PHA regions. This issue will be addressed in the future work and obtained results will be tested against debiased model of NEA orbital distribution ([7] or other). In the case of significant influence of the survey biases the similar method will be applied to the debiased data to correct the shapes of the PHA regions, preserving their purity.

Tools Used and the Source Code All computations carried out in the frame of the current work were made with the use of Python programming language and open tools for numeric computations (NumPy, SciPy), data analysis (Pandas, SciPy), machine learning (Scikit-learn) and data visualization (Matplotlib).

The code is organized as a collection of Python modules and Jupyter Notebooks as a separate open-source project named Asterion. The project was initiated at NASA Space Apps Challenge global hackathon in April 2016 and became a global finalist in the nomination "Best use of data". The code can be accessed at GitHub[4].

Acknowledgments The author is grateful to: the organizers of the global hackathon NASA Space Apps Challenge 2016 for offering awesome challenges, one of which inspired him for the current research; the officials of Kirovograd Flight Academy of the National Aviation University, who organized and hosted first-ever Space Apps Challenge in Ukraine, particularly—Alexey Izvalov and Sergey Nedelko; the members of the team Asterion—CYA, particularly—Eugene Scherbina and Andrij Blakitnij, who's hard labor and enthusiasm pushed the boundaries of impossible; Ian Webster—developer of Asterank for collecting and sharing the asteroid database and for collaborating with the author on including PHA ranking features to the service, Giovanni F. Gronchi from the University of Pisa for sharing his paper, Carrie R. Nugent from IPAC/Caltech for referring to important papers and pointing the need of testing obtained results on the debiased data, which will be the subject of the future work. The author would also like to show his gratitude to all contributors to the Scikit-learn project—open-source software for machine learning that provides easy access to sophisticated math and encourages experimenting with different algorithms.

References

1. Beeson, C.L., Elvis, M., Galache, J.L.: Scaling near Earth asteroid (NEA) characterization rates. Harv. Undergrad. Res. J. Astrophys. **6**(1), (2013). http://thurj.org/research/2013/05/4458/
2. Galache, J.L., Beeson, C.L., McLeod, K.K., Elvis, M.: The need for speed in near-Earth asteroid characterization. Planet. Space Sci. **111**, 155–166 (2015.) https://arxiv.org/pdf/1504.00712.pdf
3. Mainzer, A., Grav, T., Bauer, J., Conrow, T., Cutri, R.M., Dailey, J., Fowler, J., Giorgini, J., Jarrett, T., Masiero, J., Spahr, T., Statler, T., Wright, E.L.: Survey simulations of a new near-Earth asteroid detection system. Astron. J. **149**(5), 172 (2015.) 17pp
4. Myhrvold, N.: Comparing NEO search telescopes. Publ. Astron. Soc. Pac. **128**(962), 045004 (2016.) http://iopscience.iop.org/article/10.1088/1538-3873/128/962/045004/pdf
5. Shao, M., Turyshev, S.G., Spangelo, S., Werne, T., Zhai, C.: A constellation of SmallSats with synthetic tracking cameras to search for 90% of potentially hazardous near-Earth objects. Astron. Astrophys. **603**, A126 (2017)
6. Bottke, W.F., Morbidelli, A., Jedicke, R., Petit, J.M., Levison, H.F., Michel, P., Metcalfe, T.S.: Debiased orbital and absolute magnitude distribution of the near-Earth objects. Icarus. **156**(2), 399–433 (2002)
7. Granvik, M., Morbidelli, A., Jedicke, R., Bolin, B., Bottke, W.F., Beshore, E., Vokrouhlický, D., Delbò, M., Michel, P.: Super-catastrophic disruption of asteroids at small perihelion distances. Nature. **530**, 303–306 (2016)

[4] https://github.com/nomad-vagabond/asterion

8. Jedicke, R., Bolin, B., Granvik, M., Beshore, E.: A fast method for quantifying observational selection effects in asteroid surveys. Icarus. **266**, 173–188 (2016)
9. Mainzer, A., Grav, T., Masiero, J., Bauer, J., McMillan, R.S., Giorgini, J., Spahr, T., Cutri, R.M., Tholen, D.J., Jedicke, R., Walker, R., Wright, E., Nugent, C.R.: Characterizing subpopulations within the near-Earth objects with NEOWISE: preliminary results. Astrophys. J. **752**(2), 110 (2012.), 16pp
10. Tricarico, P.: The near-Earth asteroid population from two decades of observations. Icarus. **284**, 416–423 (2017)
11. Granvik, M., Vaubaillon, J., Jedicke, R.: The population of natural Earth satellites. Icarus. **218**, 262–277 (2012)
12. Fedorets, G., Granvik, M., Jedicke, R.: Orbit and size distributions for asteroids temporarily captured by the Earth-Moon system. Icarus. **285**, 83–94 (2017)
13. Feigelson, E.D., Babu, G.J.: Modern Statistical Methods for Astronomy: With R Applications. Cambridge University Press, New York (2012)
14. Scikit-learn official website. http://scikit-learn.org/stable/
15. Harrington, P.: Machine Learning in Action. Manning, Greenwich (2012)
16. Marsland, S.: Machine learning: an algorithmic perspective Machine Learning & Pattern Recognition, 2nd edn. Chapman & Hall/CRC, Boca Raton (2014)
17. Madhulatha, T.S.: An overview on clustering methods. IOSR J Eng. **2**(4), 719–725 (2012.) https://arxiv.org/ftp/arxiv/papers/1205/1205.1117.pdf
18. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, pp. 226–231. (1996). http://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf
19. Gronchi, G.F., Valsecchi, G.B.: On the possible values of the orbit distance between a near-Earth asteroid and the Earth. Mon. Not. R. Astron. Soc. **429**, 2687–2699 (2013)
20. Blanco-Silva, F.J.: Learning SciPy for Numerical and Scientific Computing. Packt Publishing. www.packtpub.com (2013)
21. Schunová-Lilly, E., Jedicke, R., Vereš, P., Denneau, L., Wainscoat, R.J.: The size-frequency distribution of H > 13 NEOs and ARM target candidates detected by Pan-STARRS1. Icarus. **284**, 114–125 (2017)