

# TAXONOMIC EVIDENCE APPLYING ALGORITHMS OF INTELLIGENT DATA MINING. ASTEROIDS FAMILIES

Gregorio Perichinsky(1) Magdalena Servente(2) Arturo Carlos Servetto(1)  
Ramón García Martínez(2,3) Rosa Beatriz Orellana(5) Angel Luis Plastino (4)

<p>(1){aserve,gperi}@mara.fi.uba.ar Databases and Operating System Laboratory (2)rgm@mara.fi.uba.ar Intelligent System Laboratory Computer Science Department - 4 th Floor South Wing Faculty of Engineering - University of Buenos Aires Paseo Colón N° 850 - (1063) Buenos Aires - Argentina Phone: (54 11) 4343-1177 (int. 140/145) FAX: (54 1) 4331-0129 (3) rgm@itba.edu.ar Buenos Aires Institute of Technology Madero 399. (1106) Phone: (54-11) 4314-8181 Buenos Aires - Argentina Phone (54 221) 421-7308</p>	<p>(4)Plastino@venus.fisica.unlp.edu.ar PROTEM Laboratory Department of Physical Sciences Faculty of Sciences - University of La Plata C.C. 727 or (115 # 48/49) (1900) La Plata – Buenos Aires - Argentina Phone: (54 221) 483-9061 - (54 221) 425-0791 (ext. 247) (5) rorellan@fcaglp.fcaglp.unlp.edu.ar Mechanics Laboratory Celestial Mechanics Department Faculty of Astronomical and Geophysical Sciences University of La Plata - Paseo del Bosque (1900) La Plata - Buenos Aires - Argentina Phone: (54 221) 421-7308</p>
--	---

**KEYWORDS:** classification, cluster (family), spectrum, induction, divide and rule, entropy.

## ABSTRACT

Numerical Taxonomy aims to group in clusters, using so-called structure analysis of operational taxonomic units (OTUs or taxons or taxa) through numerical methods. Clusters that constitute families was the purpose of this series of last projects.

Structural analysis, based on their phenotypic characteristics, exhibits the relationships, in terms of degrees of similarity, between two or more OTUs.

Entities formed by dynamic domains of attributes, change according to taxonomical requirements:

### **Classification of objects to form families.**

Taxonomic objects are represented by semantics application of Dynamic Relational Database Model.

Families of OTUs are obtained employing as tools i) the Euclidean distance and ii) **nearest neighbor** techniques. Thus **taxonomic evidence** is gathered so as to quantify the similarity for each pair of OTUs (**pair-group method**) obtained from the basic data matrix.

**The main contribution up until now is to introduce the concept of spectrum of the OTUs**, based in the states of their characters. The concept of families' spectra emerges, if the superposition principle is applied to the spectra of the OTUs, and the groups are delimited through the maximum of the Bienaymé-Tchebycheff relation, that determines **Invariants (centroid, variance and radius)**.

A new taxonomic criterion is thereby formulated.

An astronomic application is worked out. The result is a new criterion for the classification of asteroids in the hyperspace of orbital proper elements.

Thus, a new approach to **Computational Taxonomy** is presented, that has been already employed with reference to **Data Mining**.

This paper analyses the application of Machine Learning techniques to Data Mining. We focused our interest on the TDIDT (Top Down Induction Trees) induction family from pre-classified data, and in particular to the ID3 and the C4.5 algorithms, created by Quinlan. We tried to determine the degree of efficiency achieved by the TDIDT family's algorithms when applied in data mining to generate valid models of the data in classification problems with the **Gain of Entropy**.

The **Informatics (Data Mining and Computational Taxonomy)**, is always the **original objective** of our researches.

## 1. Introduction

Taxonomic objects are here represented by the application of the semantics of the Dynamic Relational Database Model: **Classification of objects to form families or clusters**[1].

Families of OTUs are obtained employing as tools i) the Euclidean distance and ii) **nearest neighbor** techniques. Thus **taxonomic evidence** is gathered so as to quantify the similarity for each pair of OTUs (**pair-group method**) obtained from the basic data matrix[2][3][4].

**The main contribution of the series of papers presented until now was to introduce the concept of spectrum of the OTUs**, based in the states of their characters. The concept of families' spectra emerges, if the superposition principle is applied to the spectra of the OTUs, and the groups are delimited through the maximum of the Bienaymé-Tchebycheff relation, that determines Invariants (centroid, variance and radius) [1].

**Applying** the integrated, independent domain technique dynamically to compute the **Matrix of Similarity**, and, by recourse to an iterative algorithm, families or clusters are obtained.

A new taxonomic criterion was thereby formulated.

The considerable discrepancies among the incongruities and existing classifications of astrophysical study results have motivated an interdisciplinary program of research that notices a clustering of asteroids in stabilized families [5].

In our case, is worked in an interdisciplinary way in Celestial Mechanics[5], Theory of the Information[6][7], Neural Networks[8] and Dynamic Databases [1] and the Algorithmic of the Numerical Taxonomy [2] [4], to achieve the discovery of the depths of the structure formation of the Solar An astronomic application is worked out. The result is a new criterion for the classification of asteroids in the hyperspace of orbital proper elements.

Thus, a new approach to **Computational Taxonomy** is presented, that has been already employed with reference to **Data Mining**.

On the other hand: (i) the work of [1] has clarified subtle points concerning the dynamic evolution in the long-term of the asteroids orbits, whose modeling is an essential prerequisite for the proper elements deriving (for the classification in families); and (ii) the availability of physical data on sizes, shapes, numerical taxonomy and rotation velocity to many hundred asteroids has provoked new families analyses [1].

While the most populous families appear in both criteria in quite homogeneous form, the **criterion** of the composition and physical precedents and cosmochemical, is a criterion with more or less difficulty and the **criterion** which with less difficulty has identified families is that one which uses data from **celestial mechanics**.

We do not consider in the transformation of isotropic and homogeneous sets, changing the values of the eccentricity and the semiaxis to recompute the values of the zones of inter-gap of the asteroids belt into the velocities in average, or eliminating groups from 5 or fewer objects, all of which we consider are outside a Computational criterion.

## 1.1 Intelligent Data Mining Introduction

Machine Learning is the field dedicated to the development of computational methods underlying learning processes and to applying computer-based learning systems to practical problems. Data Mining tries to solve those problems related to the search of interesting patterns and important regularities in large databases [9] [[10]..[15]]. Data Mining uses methods

and strategies from other areas, including Machine Learning. When we apply Machine Learning techniques to solve a Data Mining problem, we refer to it as an Intelligent Data Mining.

This paper analyses the TDIDT (Top Down Induction Trees) induction family, and in particular to the C4.5 algorithm[13b][14]. We tried to determine the degree of efficiency achieved by the C4.5 algorithm when applied in data mining to generate valid models of the data in classification problems with the **Gain of Entropy**.

The C4.5 algorithm generate decision trees and decision rules from pre-classified data. The "divide and rule" method is used to build the decision trees. This method divides the input data in subsets according to some pre-established criteria. Then it works on each of these subsets dividing them again, until all the cases present in one subset belong to the same class.

## 2. Constructing the decision trees

### 2.1. ID3

The Induction Decision Trees algorithm was developed as a supervised learning method, for build decision trees from a set of examples. The examples must have a group of attributes and a class. The attributes and classes must be discrete, and the classes must be disjoint. The first versions of this algorithms allowed just two classes: positive and negative. This restriction was eliminated in later releases, but the disjoint classes restriction was preserved. The descriptions generated by ID3 cover each one of the examples in the training set.

### 2.2. C4.5

The C4.5 algorithm is a descendant of the ID3 algorithm, and solves many of its predecessor's limitations. For example, the C4.5 works with continuous attributes, by dividing the possible results in two branches: one for those values  $A_i \leq N$  and another one for  $A_i > N$ . Moreover, the trees are less bushy because each leaf covers a distribution of classes and not one class in particular as the ID3 trees, this makes trees less profound and more understandable[13b][14]. C4.5 generates a decision tree partitioning the data recursively, according to the depth-first strategy. Before making each partition, the system analyses all the possible tests that can divide the data set and selects the test with the higher information gain or the higher gain ratio. For discrete attributes, it considers a test with  $n$  possible outcomes,  $n$  being the amount of possible values that the attribute can take. For continuous attribute, a binary test is performed on each of the values that the attribute can take.

### 2.3. Decision trees

The trees TDIDT, to those which belong generated them by the ID3 and post C4.5, are built from method of Hunt. The ID3 and C4.5 algorithms use the "divide and rule" strategy to build the initial decision tree from the training data [16].

The form of this method to build a decision tree as of a set  $T$  of training data, divides the data in each step according to the values of the “best” attribute. Any test that divides  $T$  in a non trivial manner, as long as two different  $\{T_i\}$  are not empty, is very simple. They will be the classes  $\{C_1, C_2, \dots, C_k\}$ .  $T$  contains cases belonging to several classes, in this case, the idea is to refine  $T$  in subsets of cases that tend, or seem to tend toward a collection of cases belonging to an only class. It is chosen a test based on an only attribute, that has one or more resulted, mutually excluding  $\{O_1, O_2, \dots, O_n\}$ .  $T$  is partition of the subsets  $T_1, T_2, \dots, T_n$  where  $T_i$  contains all the cases of  $T$  that have the result  $O_i$  for the elected test. The decision tree for  $T$  consists in a node of decision identifying the test, with a branch for each possible result. The construction mechanism of the tree is applied recursively to each subset of training data, so that the  $i$ -th branch carry to the decision tree built by the subset  $T_i$  of training data.

Still, the ultimate objective behind the process of constructing the decision tree isn't just to find any decision tree, but to find a decision tree that reveals a certain structure of the domain, that is to say, a tree with predictive power. That is the reason why each leave must cover a large number of cases, and why each partition must have the smallest possible number of classes. In an ideal case, we would like to choose in each step the test that generates the smallest decision tree.

Basically, what we are looking for is a small decision tree consistent with the training data. We could explore and analyze all the possible decision trees and choose the simplest one. However, the searching and hypothesis space has an exponential number of trees that would have to be explored. The problem of finding the smallest decision tree consistent with the training data has NP-complexity.

To calculate which is the “best” attribute to divide the data in each step, both the information gain and the gain ratio were used. Moreover, the trees generated with the C4.5 algorithm were pruned according to the method, this post-pruning was made in order to avoid the overfitting of the data.

## 2.4. Transforming decision trees to decision rules

Decision trees that are too big or too bushy are somewhat difficult to read and understand because each node must be interpreted in the context defined by the previous branches. In any decision tree, the conditions that must be satisfied when classifying a case can be found following a trail from the root to the leave to which that case belongs. If that trail was transformed directly into a production rule, the antecedent of the rule would be the conjunction of all the tests in the nodes that must be traversed to reach the leaf. All the antecedents of the rules built this way are mutually exclusive and exhaustive.

To transform a tree to decision rules, the C4.5 algorithm traverses the decision tree in preorder (from the root to the leaves, from left to right) and constructs a rule for each path from the root to the

leaves. The rule's antecedent is the conjunction of the value tests belonging to each of the visited nodes, and the class is the one corresponding to the leaf reached.

## 2.5. Evaluation of the TDIDT family

We used a crossed-validation approach to evaluate the decision trees and the production rules obtained. Each dataset was divided into two sets with proportions 2:3 and 1:3. We used two thirds of the original data as a training set and one third to evaluate the results. We expressed the results of these tests in a confusion matrix, where each class had two values associated to it: the number of examples classified correctly and the number of examples classified as belonging to another class.

## 3. Requirements engineering.

### 3.1. Hirayama

Examining the distribution of the asteroids with respect to their orbital elements, in particular their principal movement, the inclination and the eccentricity, are observed condensations in different places that seem at random, but there are some cases in which taking into account only the quantities of the probability is not so evident [1].

The asteroids are also grouped by having nearby inclinations or the plans of the orbital have practically the same pole (that of the orbit of Jupiter), other groupings do not have the same center but the drawing of the graph taking the eccentricity and the length of the perihelion instead of the inclination and the length of the node distribution has the shape of a circumference. Continuing the development of the mentioned theory do not exist doubts of the fact that there are physical relationships that connect the asteroids. Because of this it is that we can venture that there exist associated asteroid families. The theory remains verified and thus the families training such as KORONIS (fhn-158), EOS (fhn-221), THEMIS (fhn-24), FLORA (fhn-244), MARIA (fhn-170) and PHOCAEA (fhn-25) (where fhn is family head number).

The orbital elements distribution in asteroid belts is not at random showing the families existence, such that the groups of asteroids whose semimajor-axis, their eccentricity and their inclination (or the sine of the same) are approximated to a cluster for certain special values following to Arnold (about 1969 there was less than 1735 objects) [1]. It has been verified the agglomeration in families (clustering) correcting the perturbation periodic produced by secular variations caused by the major planets, like Jupiter, taking the proper elements. Other groupings have been identified by proper resonance characteristics or current of impelled asteroids (JET STREAMS) through the FLORA family and objects that cross MARS in orbits of superior order eccentricity.

Taking into account that Celestial Bodies are based on physical attributes, on phenotypic characteristic of characters or attributes of the asteroids and finally on their genotypic or common origin. Nearby vicinity

condition should be taken account and the high density families are the most stable and less random. Families of Hirayama are confirmed and the small families are of low density and the probability to belong to the families is high and therefore their coupling by the pair-group method is possible.

About 1982, Carusi and Valsechi there is a record of 2125 smaller planets, asteroid type, grouping which produce discrepancies in the results of the classification computational methods based on physical and dynamical parameters [1].

This discrepancy among the statistic methods is disconcerting since the relationship among the members of a family with respect to the dynamical parameters and any physical study that is accomplished on the same should be concurrent. It can be observed that the growth in observations does not solve the discrepancies. Of the methods of families identification the discrepancies emerge by their probabilist criteria and the future new asteroids discovery seem that exists a contradiction between them, but in spite of all this, if there is congruity, the suspected families appear in the reality (scientific method of contrast) but if the methods are arbitrary they are always debatable in addition to the methodological doubt [the authors].

For **Williams** the problem of Arnold was already discussed in function of their criterion of distribution density uniform Poissonian and the proper elements. In the 1980s the analysis techniques by similarity and a generalized distance but with the use of personal judgements or manual managing is what is usual and not an automatic classification. Because of this appears the consideration of the variance ( $\sigma$ ) of the domains and families for the process of elements identification within the family or the subsequent. The accepted classes have been split into two types: 1), if the class has been identified in two intervals, without noticeable differences and 2), if the class was found mixed coupling with other less important classes in overlap intervals, being able to exist masked families or less reliable contours, these aspects should emerge of the proper statistic method.

These projects of the Jet Propulsion Laboratory, California Institute of Technology, gave as a result crossing orbits of major planets and that are split into families, by the characteristic of the method. A characteristic is that the strong resonance does not appear in asteroid and the weak one is taken as noise.

The distances are taken from a right line SUN-PLANET (Mars MXR, Jupiter JXR, Saturn SXR, etc.) and the proper values are more exact within belt than outside it (something which endorses the theory of the authors).

For **Knezevic and Milani** the proper asteroid elements of an analytical theory of second order, of asteroids identified in the principal belt (main-belt), are much more exact than those of eccentricity and small inclination in the region of the family Themis. This is because the short periodical perturbations are eliminated and are taken into account the principal second dependent order effects, according to the

results of the consistent algorithm with the modern dynamic theories of Kolmogorov-Arnold-Moser, they are about 3495 asteroids of the edition of the Leningrad Ephemerides of the Minor Planets. Hildas, Trojans and the nearby to the Earth ( $q < 1.1$  u.a.) were discarded.

All this development appears less clear and arbitrary, there is not a formal basis in the relationship convergence quantity of iterations (code of quality QC) and the number of asteroids.

The criterion of **Zappala, Cellino, Farinella and Knezevic** (1992 and subsequent) is important since an improved asteroids classification was noted in dynamic families, analyzing a numbered asteroids database, whose proper elements have been computed in a new second-order, fourth-degree secular perturbation theory by, and verified their stability in the long term. The multivariate criterion uses the technique of hierarchic clustering data analysis. It was applied to build for each zone of the asteroids belt a "dendrogram", graph, in the proper elements space, with a distance in function related to the necessary incremental velocity of the orbital change after the ejection from the fractional parent body.

The parameters of importance associated with each family, measured as random concentrations results, (as to transform the zones anisotropy and inhomogeneous into homogeneous zones and isotropy of the inter-gaps zones in the asteroids belt modifying mechanical attributes as the semimajor-axis and the inclination) and the hardness parameters (stability), were obtained repeating the classification procedure after varying the velocity elements in small quantities to recompute the real zones from the calculations with the artificial changing of the coefficients of the distance function.

The most important and healthy families are as usual Themis, Eos, and Koronis, that jointly include 14% of the known principal belt of the population; but 12 more reliable and healthy families that were found throughout the belt, the majority departed partially of previous classifications.

It is the case of FLORA in the region of the interior belt, giving rise for a very difficult reliable families identification, mainly when have a high density and the accuracy of the inclinations and proper eccentricities is poor mainly on account of the proximity of a strong secular resonance.

It is arrived thus to constitute 21 families with an actually important method and totally automated methods.

### **3.2. Spectral analysis classification criterion**

We have decided to accomplish with our **spectral analysis criterion**, the classifications extended to the proper elements database of asteroids in families[1]. We recognize that the works of Zappala are very important (automatic classification and hierarchic method), and a point of inflection in the early 90's but is different the approach because we work in computational taxonomy, in a taxonomic hyperspace, and not in a criterion of the composition and physical precedents and cosmochemical.

Zappala use a confusing methodology, with only one variable of velocity, and that transforms a homogeneous space into inhomogeneous one and conversely not clearly univocal.

Incorporating thus an updated and larger set of osculating elements that were derived from the secular perturbation theory, whose accuracy (specifically, the stability in the time) has been extensively verified by numerical integration in the long-term; in automatic form, and to prejudice the technique of data analysis in not-random groups is not used in the proper elements space as in the criterion of Zappala and quantitatively the statistical importance of these groups; with robustness of the statistics for the important families with respect to the small random variations of proper elements, all based on an analysis on Computational Taxonomy.

We do not consider in the transformation of isotropic and homogeneous sets, changing the values of the eccentricity and the semiaxis to recompute the values of the zones of inter-gap of the asteroids belt into the velocities in average, or eliminating groups from 5 or fewer objects, all of which we consider are outside a Computational criterion.

Thus, a new approach to Computational Taxonomy is presented, that has been already employed with reference to Data Mining.

### 3.3. Numerical Taxonomy.

We infer an **analogy** of the **taxonomic representation** [1] in **dynamic relational database**.

We explain the theoretical development of a domain's structured Database and how they can be represented in a Dynamic Database.

Immediately we apply our model to the structural aspects of the taxonomy, applying Scaling Methods for domains[2] [4].

We define numerical methods used for establishing and defining clusters by their taxonomic distances.

We shall let  $C_{jk}$  stand for a general dissimilarity coefficient of which taxonomic distance,  $d_{jk}$ , is a special example. Euclidean distances will be used in the explanation of clustering techniques.

In discussing clustering procedures we make a useful distinction between three types of measure.

We use clustering strategy of space-conserving or the space-distorting strategies that appears as though the space in the immediate vicinity of a cluster has been contracted or dilated and if we return to the criterion of admission for a candidate joining an extant cluster, this is constant in all **pair-group** method.

Thus we can represent the **data matrix** and to compute the **resemblance of normalized domains**. The steps of clustering are the **recomputation** of the coefficient of similarity for future admission followed by the **admission criterion** for new members to an established cluster.

The strategies of both **space-conserving** and **space-distorting** that appear in the immediate vicinity of a cluster either contract or dilate the space, and this is constant in all **pair-group** methods [1].

#### 3.3.1. Dispersion

Once a typical value it is known of the variable of the states of the characters, it is necessary to have a parameter that give an idea of how scattered, or concentrated, are their values respect to the mean value[19].

It is considered to the variance as a moment of second order and represents the moment of inertia of the distribution of objects ( mass ) with respect to their gravity center: centroid.

When  $\overline{X'ij} = (X_{ij} - \overline{Xj}) / \sigma_j$  [2] is a normalized variable the one which represents the deviation of  $X_{ij}$  with respect to their mean in units of  $\sigma_j$ .

The normalization of the states of the character causes that the average of all character will be of value zero and variance of unitary value.

If we take as value of the dispersion to the variance  $\sigma_d^2$ , we express the principle of minimal square.

It will be  $g(X_{ij})$  a not negative function of the variable  $X_{ij}$ , for all  $k > 0$  will have to be the probability function:

If  $g(X_{ij}) = (X_{ij} - \overline{Xj})^2$ ,  $K = k^2 \sigma_j^2$ , obtaining for all  $k > 0$  the inequality from Bienaymé-Tchevicheff:

$$P(|X_{ij} - \overline{Xj}| \geq k \cdot \sigma_j) \leq 1 / k^2$$

This inequality shows that the quantity of ( OTUs ) mass of the located distribution would be of the interval

$$\overline{Xj} - k \cdot \sigma_j < X_{ij} < \overline{Xj} + k \cdot \sigma_j$$

it is to what is maximal value equal to  $1 / k^2$ , giving a utilization idea of  $\sigma_j$  as measure of the dispersion or concentration.

#### 3.3.2. Clusters and Spectra.

In discussing Sequential, Agglomerative, Hierarchic and Nonoverlapping (SAHN) [4] clustering procedures we make a useful distinction between the three types of measure.

We shall be concerned with clusters **J**, **K** and **L** containing **tj**, **tk** and **tl** OTUs, respectively, where **tj**, **tk** and **tl** all  $\geq 1$ . OTUs **j** and **k** are contained in clusters **J** and **K**, and **l**  $\in$  **L**, respectively. Given two clusters **J** and **K** that are to be joined, the problem is to evaluate the dissimilarity between the resulting joint cluster and additional candidates **L** for further fusion. The fused cluster is denoted (**J,K**), with **t<sub>j,k</sub>** = **t<sub>j</sub>** + **t<sub>k</sub>** OTUs.

The cluster center or centroid represents an average object, which is simply a mathematical construct that permits the characterization of the Density, the Variance, the taxon radius and the range as **INVARIANT** quantities.

The states of the taxonomic characters in a class, defined ordinarily with reference to the set of their properties, allow one to calculate the distances between the members of the class. The distances can be established by the similarity relationship among individuals (obtaining a matrix of similarity that has been computed).

Considering characteristic spectra [1], in addition to the states of the characters or attributes of the OTUs,

we introduce here the new **SPECTRAL** concepts of i)**OBJECTS** and ii)**FAMILY SPECTRA**.

Within the taxonomic space this method of clustering delimits taxonomic groups in such a manner that they can be visualized as characteristic spectra of an OTU and characteristic spectra of the families.

We define an individual spectral metric for the set of distances between an OTU and the other OTUs of the set. Each one provides the states of the characters and, therefore, is constant for each OTU, if the taxonomic conditions do not change (in analogy with the fasors) having an individual taxonomic spectrum (ITS).

The spectrum of taxonomic similarity is the set of distances between the OTUs of the set, that determine the constant characteristics of a cluster or family, for a given type of taxonomic conditions.

Invariants are found that characterize each cluster. Among them we mention the variance, the radius, the density and the centroid.

These invariants are associated with the spectra of taxonomic similarity that identify each family.

### 3.4. Tests of Intelligent Data Mining

A software system was constructed to evaluate the C4.5 algorithm. This system takes the training data as an input and allows the user to choose whether he wants to construct a decision tree according to the C4.5. If the user chooses the C4.5, the decision tree is generated, then it is pruned and the decision rules are built.

The decision tree and the ruleset generated by the C4.5 are evaluated separate from each other.

We use the system to test the algorithms in different domains, mainly Elita: a base of asteroids.

#### 3.4.1. Compute of the Information Gain

In the cases, in those which the set T contains examples belonging to different classes, is accomplished a test on the different attributes and is accomplished a partition according to the "better" attribute. To find the "better" attribute, is used the theory of the information, that supports that the information is maximized when the entropy is minimized. The entropy determines the randomness or disorder of a set.

We suppose that we have negative and positive examples. In this context the entropy of the subset  $S_i$ ,  $H(S_i)$ , it can be calculated as:

$$H(S_i) = -p_i^+ \log p_i^+ - p_i^- \log p_i^- \quad (3.4.1)$$

Where  $p_i^+$  is the probability of a example is taken in random mode of  $S_i$  will be positive. This probability may be calculated as

$$p_i^+ = \frac{n_i^+}{n_i^+ + n_i^-} \quad (3.4.2)$$

Being  $n_i^+$  the quantity of positives examples of  $S_i$ , and  $n_i^-$  the quantity of negatives examples.

The probability  $p_i^-$  is calculated in analogous form to  $p_i^+$ , replacing the quantity of positives examples by the quantity of negatives examples, and conversely.

Generalizing the expression (3.4.1) for any type of examples, we obtain the general formulation of the entropy:

$$H(S_i) = \sum_{i=1}^n -p_i \log p_i \quad (3.4.3)$$

In all the calculations related to the entropy, we define  $0 \log 0$  equal to 0.

If the attribute  $at$  divide the set  $S$  in the subsets  $S_i$ ,  $i = 1, 2, \dots, n$ , then, the total entropy of the system of subsets will be:

$$H(S, at) = \sum_{i=1}^n P(S_i) \cdot H(S_i) \quad (3.4.4)$$

Where  $H(S_i)$  is the entropy of the subset  $S_i$  and  $P(S_i)$  is the probability of the fact that an example belong to  $S_i$ . It can be calculate, used the relative sizes of the subsets, as:

$$P(S_i) = \frac{|S_i|}{|S|} \quad (3.4.5)$$

The gain of information may be calculate as the decrease in entropy. Thus:

$$I(S, at) = H(S) - H(S, at) \quad (3.4.6)$$

Where  $H(S)$  is the value of the entropy a priori, before accomplishing the subdivision, and  $H(S, at)$  is the value of the entropy of the subsets system generated by the partition according to  $at$ .

The use of the entropy to evaluate the best attribute is not the only one existing method or used in Automatic Learning. However, it is used by Quinlan upon developing the ID3 and his succeeding the C4.5.

#### 3.4.2. Numerical Data

The decision trees can be generated so much as discrete attributes as continous attributes. When it is worked with discrete attributes, the partition of the set according to the value of an attribute is simple.

To solve this problem, it can be appealed to the binary method. This method consists in forming two ranges of agreement values to the value of an attribute, that they can be taken as symbolic.

## 4. Results and Conclusions.

### 4.1. Results of the C4.5.

The C4.5 with post-pruning results in trees smaller and less bushy. If we analyze the trees obtained in the domain, we'll see that the percentages of error obtained with the C4.5 are between a 3% and a 3.7%, since that the C4.5 generate smaller trees and smaller rulesets. Derivative of the fact that each leaf in a tree generated covers a distribution of classes.

## 4.2. Error percentage

{ELITA} { [1]: C4.5-Gain Trees [2]: C4.5-Gain Rulers [3]: C4.5-Proportion of Gain Trees [4]: C4.5-Rulers Proportion of Gain Trees} < 3%

From the analysis of this value we could conclude that no method can generate a clearly superior model for the domain. On the contrary, we could state that the error percentage doesn't appear to depend on the method used, but on the analyzed domain.

## 4.3. Hypothesis space

The hypothesis space for this algorithm is complete according to the available attributes. Because any value test can be represented with a decision tree, this algorithm avoid one of the principal risks of inductive method that works reducing the spaces of the hypothesis.

An important feature of the C4.5 algorithm is that it use all the available data in each step to chose the "best" attribute; this is a decision that is made with statistic method. This fact favors this algorithm over other algorithms because analyze how the input dataset take the representation into decision trees in consistent forms.

Once an attribute has been selected as a decision node, the algorithm does not go back over their choices. This is the reason why this algorithm can converge to a local maximum[20]. The C4.5 algorithm adds a certain degree of reconsideration of its choices in the post-pruning of the decision trees.

Nevertheless, we can state that the results show that the proportion of error depends on the data domain. For future study, we suggest an analysis the input datasets with the numerical method of clustering and choosing for the domain the method that maintains a low percentage error in extended databases as a robustness of the method.

## 5. Corollary

From what has been said, the work uses the Sequential, Agglomerative, Hierarchic and Nonoverlapping clustering procedures, spectral analysis criterion and invariants to accomplish classifications in extended databases, of proper asteroid elements, to structure families.

The pre-classified data is an important input to Intelligent Data Mining, and Computational Taxonomy in Databases will have always a low percentage error in extended databases as a robustness of the method; to combine a sure result.

## References

[1]Perichinsky, G., Orellana, R., Plastino, A.L., Jimenez Rey, E. and Grossi, M.D. "Spectra of Taxonomic Evidence in Databases." Proceedings of XVIII International Conference on Applied Informatics. (Paper 307-7-1 ).Innsbruck. Austria. 2000.

[2]Crisci, J.V. , Lopez Armengol, M.F. "Introduction to Theory and Practice of the Numerical Taxonomy" , A.S.O. Regional Program of Science and Technology for Development. Washington D.C. Spanish. 1983.

[3]Gennari,J.H. "A Survey of Clustering Methods" (b). Technical Report 89-38. Department of Computer Science and Informatics. University of California., Irvine, CA 92717. 1989.

[4]Sokal, R.R., Sneath, P.H.A. "Numerical Taxonomy".W.H.Freeman and Company. 1973.

[5]Zappala, V, Cellino,A., Farinella,P., Milani,A., The Astronomical Journal, 107, 772. 1994

[6]Abramson,N., "Information Theory and Coding". McGraw Hill. Paraninfo. Madrid. 1966.

[7]Hamming, R.W. "Coding and information theory". Englewood Clifs, NJ: Prentice Hall. 1980.

[8]Freeman,J.A., Skapura,D.M. "Neural Networks. Algorithms, applications and techniques of programming". Addison Wesley. Iberoamericana. Spanish. 1991.

[9]Michalski, R. S. 1998. *A Theory and Methodology of Inductive Learning*. En Michalski, R. S., Carbonell, J. G., Mitchell, T. M. (1983) Machine Learning: An Artificial Intelligence Approach, Vol. I. Morgan-Kaufman, USA.

[10]Quinlan, J.R. 1986. *Induction of Decision Trees*. In Machine Learning, Ch. 1, p.81-106. Morgan Kaufmann.

[11]Quinlan, J.R. 1987. *Generating Production Rules from Decision trees*. Proceeding of the Tenth International Joint Conference on Artificial Intelligence, p. 304-307. San Mateo, CA., Morgan Kaufmann, USA.

[12]Quinlan, J.R. 1988. *Decision trees and multi-valued attributes*. En J.E. Hayes, D. Michie, and J. Richards (eds.), Machine Intelligence, V. II, p. 305-318.Oxford University Press, Oxford, UK.

[13]Quinlan, J.R. 1993. *Learning Efficient Classification Procedures and Their Application to Chess Games*, In R. S. Michalski, J. G. Carbonell, & T. M. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann, V. II, Ch. 15, p. 463-482, USA.

[13b]Quinlan, J.R. 1993 C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California, EE.UU.

[14]Quinlan, J.R. 1996. *Improved Use of Continuous Attributes in C4.5*. Basser Department of Computer Science, University of Science, Australia.

[15]Quinlan, J.R.1996. *Learning First-Order Definitions of Functions*. Basser Department of Computer Science, University of Science, Australia

[16]Hunt, E.B., Marin, J., Stone, P.J. 1966 (1995-AI). *Experiments in Induction*. New York: Academic Press, USA.

[17]Hirayama,K. "Present State of the Families of Asteroids". Proceeding of Physics-Mathematics Society. Japan II:9. pp 482-485. 1933.

[18]Cramer, Harald. "Mathematics Methods in Statistics".Aguilar Edition.Madrid.Spanish.1958.

[19]Mitchell, T. 1997. *Machine Learning*. MCB/McGraw-Hill, Carnegie Mellon University, USA.

[20]Mitchell, T. 2000 *Decision Trees*. Cornell University, www.cs.cornell.edu/courses/c5478/2000SP, USA.

[21]Feynman, R.P., Leighton, R.B. & Sands, M. "Lectures on physics, Mainly Mechanics, Radiation and Heat". pp. 25-2 ff, 28-6 ff, 29-1 ff, 37-4. 1971.

[22]Hetcht,E. and Zajac,A., "Optic". Inter-American Educational Fund. pp. 5-11-206-207-293-297-459-534. Spanish 1977.