TAXONOMIC EVIDENCE AND ROBUSTNESS OF THE CLASSIFICATION APPLYING INTELLIGENT DATA MINING.

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

(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 - Universy of Buenos Aires
Paseo Colón Nº 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

(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

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

ABSTRACT
Numerical Taxonomy aims to group in families, using so-called structure analysis of operational taxonomic units (OTUs or taxons or taxa).
Clusters that constitute families with a new criterion, is the purpose of this series of papers.
Structural analysis, based on phenotypic characteristics, exhibits the relationships, in terms of degrees of similarity, through the computation of the Matrix of Similarity, applying the technique of integration dynamic of independent domains, of the semantics of the Dynamic Relational Database Model.
The main contribution 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 principles of superposition and interference, and the Invariants (centroid, variance and radius) determined by the maximum of the Bienaymé-Tchebycheff relation, are applied to the spectra of the OTUs.
Using in successive form an updated database through the increase of the cardinal of the tuples, and as the resulting families are the same, we ascertain the robustness of the method.
Through Intelligent Data Mining, we focused our interest on the Quinlan algorithms, applied in classification problems with the Gain of Entropy, we contrast the Computational Taxonomy, obtaining a new criterion of the robustness of the method.

1. Introduction

Classification is an abstraction technique used to collect objects with common properties.
The following hypothesis: 1) each object belongs to one (and only one) class and 2) for each class at least one object belongs to the classification, allow us to delimit the domain of objects.
The association of concepts in systematic way by recourse to numerical variables has been the source of a great variety of numerical classification techniques, that have their origin in Numerical Taxonomy.
The search of classification concepts that facilitate a robust classification structure (not modifiable by the addition of new information and not altered by the incorporation of new entities) constitutes an important endeavor. In such a line, this work develops tools based on Information Theory and, as a result, a new classification technique is found.

We discuss first principles and methods and then, as an application, we investigate the application of our ideas to the classification of celestial bodies (asteroids), and the classification of plants in biological taxonomy, and the classification of objects of other feasibleness and/or formal disciplines.
A correspondence with ideas pertaining to the field of the Dynamic Databases is also established [5] [[6].to. [10]] [[13].to. [21]].
Taxonomic objects are here represented by the application of the semantics of the Dynamic Relational Database Model: Classification of objects to form clusters or families [47].to.[58].
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[12][26][61].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) [57] [58].
Applying the integrated, independent domain technique dynamically to compute the Matrix of Similarity, and, by recourse to an iterative algorithm [60], families or clusters are obtained.
A new taxonomic criterion was thereby formulated.
The considerable discrepancies among the incongruities and existing classifications whose studies results in several disciplines have motivated an interdisciplinary program of research that notices a clustering of objects in stabilized families [64] [65] [66] [67]
In our case, is worked in an interdisciplinary way in Celestial Mechanics[66] [67], Theory of the Information[1][27], Neural Networks[25] and Dynamic Databases [51] and the Algorithmic of the Numerical Taxonomy [12] [61], to achieve the discovery of the depths of the structure formation of the Solar System, an astronomic application is worked out. The result is a new criterion for the classification of Celestial Bodies in the hyperspace of orbital proper elements, Biological Sciences for linguistic and live beings to avoid confusions, uncertainties and ambiguities, in such a way that the classes include the attributes and the relationships[12].
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 works [66] [67] have clarified subtle points concerning the dynamic evolution in the long-term orbits of the asteroids, whose modeling is an essential prerequisite for the proper elements derived (for the classification in families); (ii) the availability of physical data on sizes, shapes, numerical taxonomy, many hundreds of OTUs has provoked new families analyses [1]; (iii) while the most populous families appear in both criteria in quite homogeneous form, the criterion that take into account the composition of the objects and their precedents, is a criterion with more or less difficulty and the criterion which with less difficulty has identified families is that uses data of scientific attributes; (iv) we do not consider in the transformation isotropic and homogeneous sets, changing the values of the attributes to recompute the values of the zones of inter-gap of the objects (if they exist) in the real space with average values and; (v) elimination of groups with few objects, all of which we consider are outside of a Computational criterion.

2. Requirements engineering.

Software engineers have many questions to answer, following Fenton & Pfleeger [22]. We know how to assess our current situation scientifically and to determine the magnitude of change when we manipulate our environment.
There are key components of empirical investigation in software engineering. Other methods, including feature analysis, are not addressed in this paper; their descriptions are available in social science research textbooks, and in a series of columns by Kitchenham[33].

2.1. Application in Celestial Bodies. Asteroids Families.

We will have the classification of celestial bodies in mind, asteroids in particular, that constitute a fundamental topic in Celestial Mechanics because of a variety of reasons. First of all, a fundamental grouping of asteroids was established, the so-called "families of Hirayama" [[29][30][31]]. Basically, they

constitute a natural laboratory for the study of CHAOS, on the one hand, and remnants of the early times of the formation of our Solar System, on the other one to achieve the discovery of the depths of the structure formation of the Solar System from this natural Laboratory that show the asteroids, their orbits, shape, constitution and physical properties.

In Astronomy, taxonomic classification began 50 years ago, used by several authors, for grouping different celestial bodies, among other: Galaxies, Variable Stars, Asteroids, Comets and Cumulous.

The considerable discrepancies among the incongruities and existing classifications of astrophysical study results have motivated an interdisciplinary program of research that noticies a clustering of asteroids in stabilized families [66][67].

With different alternative and pauses the topic of asteroid families maintains a particular interest motivating new researches.

Following Arnold the orbital elements distribution in asteroid belts is not at random showing the families existence, such that the groups of asteroids whose Celestial Mechanics attributes are approximated to a cluster for certain special values [2] (1969).

It has been verified the agglomeration in families (clustering) correcting the perturbation periodic produced by secular variations caused by the major planets, taking the proper elements.

On this base a cluster must have a greater density, since this is taken in a region whose center will be an asteroid. It is clearer and convenient to take an ellipsoidal region instead of a rectangular one since this result is inconsistent.

According to Arnold and following the Poisson Law the number of elements of a set must be less than a certain waited number, with which this work is not agreed because the events do not follow this law by contradicting all that is developed until now: it is 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.

For Carusi and Valsechi and about 1982 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 [3] [4].

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.

Researchers have arrived to the conclusion that the problem of the classification of the asteroids in families is clearly defined and practically solved; simplistic vision that is not agreed with by this work.

It can be observed that the growth in observations among 1969 [2], 1978 [3], 1979 [63] and 1982 [4] 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 [63] 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 (σ_j) 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.

The rejection criteria of a member of a family are not clear, they are arbitrary or directly are not exposed in the projects and by logical consequence are not automatic, like in Natural Sciences applications, plants [12], in which the specialist define the regions of the clusters.

In the works of Knezevic and Milani the proper asteroid elements of an analytical theory of second order [34][35][38], of asteroids identified in the principal belt (main-belt), are much more exact than those of mechanical attributes in a region of a family.
The algorithm that [63] permits to calculate a code of quality (QC) that indicates how much iteration one must accomplish so that converge.
All this development appears less clear and arbitrary, there is not a formal basis in the relationship convergence quantity of iterations (why QC?) and the number of objects (?).
Zappala,Cellino,Farinella and Knezevic [66] found an important criterion since an improved classification was noted in dynamic families, analyzing a numbered OTUs database. The families are identified then by comparison with similar dendrograms, derivatives from a distribution "quasirandom" of elements that compare the structure to gross scale of the real distribution.
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 (if they exist).
It is arrived thus to constitute families [66][67] with an actually important method and totally automated methods.
The criterion of Chapman et al. (1989) [66][67] is different, the families of Williams with 12 or more members are taxonomically different from the precedent ones, of those with less than five members will be definitely not different (something which does not imply that they will be necessarily and generically "unreal").

2.2.Spectral analysis classification criterion

With these motivations, we have decided to accomplish with our spectral analysis criterion, the classifications extended to the proper elements database of asteroids in families [22] [23] [24] [27] [28] [29] [30] [31].We recognize that the works of Zappala [66][67] 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 transformed space not clearly univocal.
We do not consider in the transformation of isotropic and homogeneous sets, changing the values of the attributes to recompute the values of the regions that emerges from the clustering eliminating groups of few 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.

2.3 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 [37] [38] [[41].to.[46]]. 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[44b][45]. 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.

2.3.1. Constructing the decision trees

2.3.1.1. 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.

2.3.1.2. 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 $A_i<=N$ 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[44b][45]. 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.

2.3.1.3. 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 [32].

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 $\{C_1, C_2, \ldots, C_k\}$. 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 $\{O_1, O_2, \ldots, O_n\}$. T is partition of the subsets $T_1, T_2, \ldots, T_n$ 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.

2.3.1.4. 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.

2.3.1.5. 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. Numerical Taxonomy.

We infer an **analogy** of the **taxonomic representation** [47].to.[58] **in dynamic relational database [62].**
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[12] [61].
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 [12] [61].

3.1. Calculation of the Average and of the Standard Deviation for the Normalization

In normalizing characters we compute the average value and the standard deviation of each string (the states of each character) and express each state as a deviation of the average in standard deviation units. The normalization of the states of the character makes the average of all characters to vanish. Likewise, variances adopt the value unity. We have

$$\overline{Xj} = \left(\sum_{i}^{n} \text{Xij} \right) / \text{n}$$

$$\sigma\text{j} = \left(\left(\sum_{i}^{n} \left(\text{Xij} - \overline{Xj} \right)^2 \right) / \left(\text{n} - 1 \right) \right)^{1/2}$$

$$\overline{X'ij} = \left(\text{Xij} - \overline{Xj} \right) / \sigma\text{j}$$

For the normalized domains we calculate both the average difference among characters (its absolute value) and the concomitant taxonomic distances. For the latter we consider two metrics: that of Minkowski and the so-called Manhattan one [12] [44]. The quantity

$$\overline{D} = \left(\sum_{i}^{n} \left| \text{Xij} - \text{Xik} \right| \right) / \text{n}$$

is the mean difference among characters,

$$\Delta\text{jk} = \left[\sum_{i}^{n} \left(\text{Xij} - \text{Xik} \right)^2 \right]^{1/2}$$

is the distance Δjk among OTUs, and we consider further the average value

djk = (($\sum_{i}^{n}$ (Xij - Xik)^2 / n))$^{1/2}$

due to the fact that Δjk grows with the number of characters.
The expectation value (d) of the djk for a normal distribution of zero and variance unity is:
E(d) = ((n - 1)! (π / n)$^{1/2}$) / (2 $^{n-2}$ [((n/2) - 1)!]^2)
After using the Stirling approximation we get
E(d) $\approx$ $\sqrt{2}$ (1 - 1/n)^{½} ((1 + (1 / (n - 2)))^{½} (1 / e))
and the expectation value of the variance for (d) turns out to be
$E(\sigma^2_d)$ = 2 - [E(d)]^2 $\approx$ 1/n.

3.2. Dispersion
The variance is a moment of second order and represents to the moment of inertia of the distribution of objects (masses) with respect to their center of gravity (the so-called centroid) [11].

$\overline{X'ij}$ = (Xij - $\overline{Xj}$) / σj is a normalized variable that represents the deviation of the Xij with respect to their mean (in units of σj).
As usual, we take the dispersion to be given by the variance σ^2_d .The mean-squares method is now to be applied.
Let g (Xij) be a not negative function of the variable Xij. For all k > 0 will have the probability function:
P [g (Xij) $\geq$ K] $\leq$ (E (g (Xij)) / K.
Theorem of Tchevicheff
Let S be the set of all the Xij that satisfy the inequality g (Xij) $\geq$ K. The truth of the theorem stems from the relationship (valid in any number of dimensions):

$$Eg(Xij) = \int_{-\infty}^{\infty} g(Xij)dF \geq K\int_{S} dF = KP(S)$$

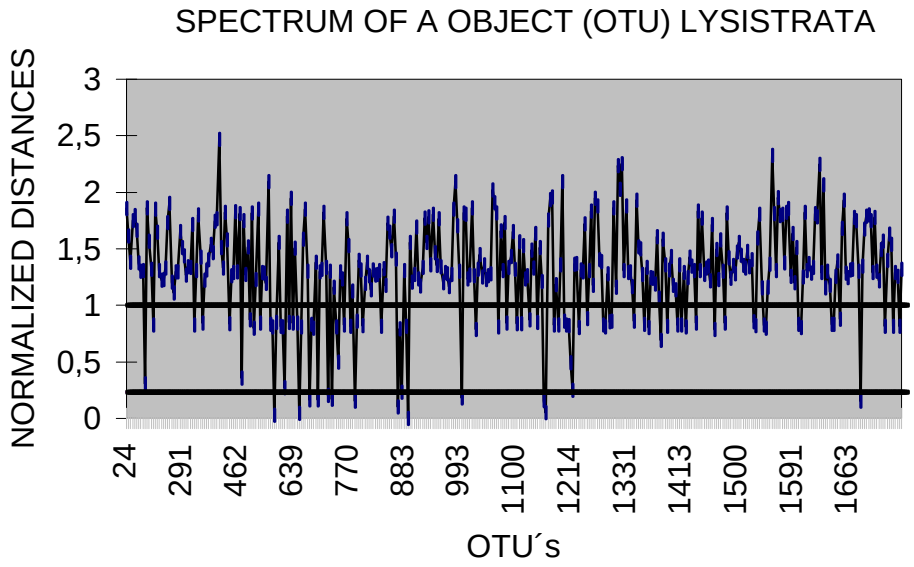
If g (Xij) = (Xij - $\overline{Xj}$)^2 , K = k^2 σj^2, which leads, for all k > 0, to the inequality of Bienaymé-Tchevicheff:
P (| Xij - $\overline{Xj}$ | $\geq$ k. σj) $\leq$ 1 / k^2
This inequality shows that
$\overline{Xj}$ - k. σj < Xij < $\overline{Xj}$ + k. σj
(maximal value equal to 1 / k^2).
In particular, for an average value $\overline{Xj}$ and deviation σj with a mass 1 / 2.k^2 located at each the points Xij = $\overline{Xj}$ $\pm$ k. σj one has
P (| Xij - $\overline{Xj}$ | $\geq$ k. σj) = 1 / k^2
a maximal limit value that can not be improved upon.
This inequality shows that the quantity of mass of the distribution is to be found in the interval
$\overline{Xj}$ - k. σj < Xij < $\overline{Xj}$ + k. σj.
The inequality permits one to fix both distribution levels and the radius of a cluster.
If we take k=2 then we obtain **k . σj** = $\sqrt$ **2 . σj** as the maximal value.

3.3. Clusters and Spectra.
In discussing Sequential, Agglomerative, Hierarchic and Nonoverlapping (SAHN) [61] 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 $\mathbf{t_{j,k} = t_j + t_k}$ 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 [23][24][28][59], 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.

3.3.1. Variation Range Normalization.

There exist sound reasons for considering that the weight of a character should be inversely proportional to its variability. For normally distributed quantitative characters their information content (in the information theory sense) is proportional to the variance. If the variances are made equal, then each character contributes an equal informational amount. Such an uniform probability yields, of course, the maximum possible entropy.
In a more general sense we may argue that the variation contributes most of the information, and that the gross character size and range of variation should contribute little toward phenetic resemblance, in terms of that information relevant for taxonomic purposes.



One observes in the graph, for the line of equal Invariant (ordinate unity), a region that clearly shows the objects that constitute it. Objects belonging to other regions are to be found above such a line. Below the line

at ordinate 0.2343 one sees objects of a family. Above these two lines we encounter other objects. A more detailed analysis is required in order to ascertain to which family these objects belong.

I. Iteration around the center (centroid). `Invariants:`

- **Average Distance: 0.1321**
- **Density: 13**
- **Dispersion: 0.059**
- **Range: 0.2343**

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: 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:

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

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

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

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:

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

In all the calculations related to the entropy, we define *0log0 equal to 0.*
If the attribute *at* divide the set *S* in the subsets S_i, i = 1,2, ... , n, then, the total entropy of the system of subsets will be:

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

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:

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

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

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

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

{ [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[39][40],. 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]Abramson,N., "Information Theory and Coding". McGraw Hill. Paraninfo. Madrid. 1966.
[2]Arnold,J.R., "Asteroids Families and Jet Streams". The Astronomical Journal. 74: pp 1235-1242. 1969.

[3]Carusi,A., Massaro,E. “On Asteroids Classifications in Families”. Astronomy and Astrophysics. Supplements 34, p 81. 1978.
[4]Carusi,A., Valsecchi,G.B. “On Asteroids Classifications in Families”. Astronomy and Astrophysics. pp 327-335. 1982.
[5]Batini, C., Ceri, S., Navathe, S.B. “Conceptual Databases Design” Addison Wesley. 1998.
[6]Codd E. F. “Relational Completeness of Data Base Sublanguages”. Database Systems, Courant Computer Science Symposia Series 6, Englewood Cliffs, New Jersey, Prentice-Hall. 1972.
[7]Codd E. F. “Extending the Data Base Relational Model to Capture More Meaning” ACM TODS 4, 4 pp 397-434. 1979.
[8]Codd E. F. “Relational Data Base. A Practical Foundation for Productivity” CACM 25, 2. 1982.
[9]Codd E. F. “How Relational is your Database Management System ?”. Computer World. 1985.
[10]Codd E. F. “The Relational Model for Database Management: Version 2”. Addison Wesley. 1990.
[11]Cramer, Harald. “Mathematics Methods in Statistics”. Aguilar Edition. Madrid. Spanish. 1958.
[12]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.
[13]Date,C.J. “An Introduction to Datase Systems Vol. I”. 6th Ed. Addison Wesley. 1995.
[14]Date,C.J. “Relational Database: Selected Writings”. Addison Wesley. 1986.
[15]Date,C.J. “Relational Database: Further Misconceptions #1”. Info DB, spring, 1986.
[16]Date,C.J. “A SQL Standard”. Addison Wesley. 1987.
[17]Date,C.J. “Where SQL Falls Short”. Datamation pp 84-86. 1987.
[18]Date,C.J. “An Introduction to Datase Systems". Addison Wesley. 1998.
[19]Date,C.J. “Date on Databases” On proceeding of the Codd & Date Relational Database Symposium”. Madrid. 1992.
[20]de Miguel, A., Piatttini, M. "Concepts and Design of Databases." Addison Wesley.1994. Spanish.
[21]Elmasri,R., Navathe,S. “Fundamentals of Database Systems”.The Benjamin/Cummings Publishing Company and Addison Wesley. 1997.
[22]Fenton, N.E., Pfleeger, Sh.L. “Software Metrics”. PWS Publishing Company. 1997.
[23]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.
[24]Frank,N.H. “Introduction to Mechanics and Heat”. Science Service. Washington. Editorial Atlante. Spanish. 1949.
[25]Freeman,J.A., Skapura,D.M. “Neural Networks. Algorithms, applications and techniques of programming”.Addison Wesley. Iberoamericana. Spanish. 1991.
[26]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.
[27]Hamming, R.W. “Coding and information theory”. Englewood Clifs, NJ: Prentice Hall. 1980.
[28]Hetcht,E. and Zajac,A., “Optic”. Fondo Educativo Interamericano. pp. 5-11-206-207-293-297-459-534. Spanish 1977.
[29]Hirayama,K. “Groups of Asteroids Probably Common Origin”. Proceeding of Physics-Mathematics Society. Japan II:9. pp 354-351. 1918.
[30]Hirayama,K. “Groups of Asteroids Probably Common Origin”. The Astronomical Journal: 31, pp 185-188. 1918.
[31]Hirayama,K. “Present State of the Families of Asteroids”. Proceeding of Physics-Mathematics Society. Japan II:9. pp 482-485. 1933.
[32]Hunt, E.B., Marin, J., Stone, P.J. 1966 (1995-AI). Experiments in Induction. New York: Academic Press, USA.
[33]Kitchenham,B., Pickard,L., Pfleeger, S.L. “Case studies for method and tool evaluation”. IEEE Software, 12(4) pp 52-62. 1995.
[34]Knêzevîc,Z., Milani,A. “Asteroids Proper Elements from an Analytical Second Order Theory”. Astronomy and Astrophysics. pp 1073. 1990.
[35]Knezevic,Z., Milani,A. Farinella,P., Froechle,Ch., Froechle, Cl, “Asteroids Family Identifications and Proper Elements”. Icarus, 93, 316. 1991.
[36]Kohonen, T. “Self - Organization and Associative Memory”. Berlin: Springer - Verlag. 1989.
[37]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-Kauffman, USA.
[38]Milani,A. “Asteroids Family Identifications and Proper Elements”.Celestial Mechanics. 57,59. 1993.
[39]Mitchell, T. 1997. Machine Learning. MCB/McGraw-Hill, Carnegie Mellon University, USA.
[40]Mitchell, T. 2000 Decision Trees. Cornell University, www.cs.cornell.edu/courses/c5478/2000SP, USA.
[41]Quinlan, J.R. 1986. Induction of Decision Trees. In Machine Learning, Ch. 1, p.81-106. Morgan Kaufmann.
[42]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.
[43]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.

[44]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.
[44b]Quinlan, J.R. 1993 C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, California, EE.UU.
[45]Quinlan, J.R. 1996. Improved Use of Continuous Attributes in C4.5. Basser Departament of Computer Science, University of Science, Australia.
[46]Quinlan, J.R.1996. Learning First-Order Definitions of Functions. Basser Departament of Computer Science, University of Science, Australia
[47]Perichinsky, G. “Multiple states of multiple state automata to key fast validation”. 11th. International Symposium Computer at University. Catvat. Zagreb. Yugoslavia. 1989.a.
[48]Perichinsky, G., Servetto, A., Crocco, E. “Relational Data Bases Structured on Dynamic Domains of Attributes".18th Sessions. Operative Research and Informatic’s Argentine Society. 1989.b.
[49]Perichinsky, G., Servetto, A. "Dynamically Integrated Independent Domains on Data Bases”.19th Sessions Operations Research and Informatic’s Argentine Society. 1990.
[50]Perichinsky, G. et Al. “Data Base Model Manager Structured on Independent Domains". Faculty of Science. National University of La Plata. Spanish. 1992.
[51]Perichinsky, G., Feldgen, M., Clúa, O. “Conceptual Contrast of Dynamic Data Bases with the Relational Model” in Proceedings International Association of Science and Technology for Development. 14th Applied Informatics Conference. Innsbruck, Austria. 1996.
[52]Perichinsky, G. , Feldgen, M. , Clúa, O. "Dynamic Data Bases and Taxonomy" in Proceedings International Association of Science and Technology for Development. 15th Applied Informatics Conference. Innsbruck. Austria. 1997.
[53]Perichinsky, G., Jimenez Rey, E.; Grossi, M. “Domain Standardization of Operational Taxonomic Units (OTUs) on Dynamic Data Bases” in Proceedings International Association of Science and Technology for Development. 16th Applied Informatics Conference. Garmisch-Partenkirchen. Germany. 1998.
[54]Perichinsky, G., Jimenez Rey, E.; Grossi, M. “Spectra of Objects of Taxonomic Evidence on the Dynamic Data Bases” in Proceedings International Association of Science and Technology for Development. 16th Applied Informatics Conference. Garmisch-Partenkirchen. Germany. 1998.
[55]Perichinsky, G., Jimenez Rey, E. and Grossi, M.D. “Application of Dynamic Data Bases in Astronomic Taxonomy” International Association of Science and Technology for Development. 17th Applied Informatics Conference Proceedings. Innsbruck. Austria. 1999.
[56]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. Innsbruck. Austria. 2000.
[57]Perichinsky, G., Orellana, R. and Plastino, A.L. "Spectra of Taxonomic Evidence in Databases." Proceedings of International Conference on Computer Science, Software Engineering, Information Technology, e-Business and Applications (CESITeA´02). Foz do Iguazu. Brazil. 2002.
[58]Perichinsky, G., Orellana, R., Plastino, A.L., Garcia Martinez, R., Servente, M. and Servetto, A. C. "Taxonomic Evidence Applying Algorithms of Intelligent Data Mining" Proceedings of International Conference on Computer Science, Software Engineering, Information Technology, e-Business and Applications (CESITeA´03). Rio de Janeiro. Brazil. 2003.
[59]Sawyer, R.A. “Experimental Spectroscopy”. Dover Publication. New York. 1963.
[60]SEI "Rationale for SQL Ada module Description language (SAMeDL)" Ver. 2.0 CMU/SEI-92-TR-16, oct 1992.
[61]Sokal, R.R., Sneath, P.H.A. "Numerical Taxonomy". W.H. Freeman and Company. 1973.
[62]Wiederhold,G.**“**Data Base Design”. McGraw-Hill Book Company. 1983/1985.
[63]Williams,J.G. “Asteroids Family Identifications and Proper Elements”. Jet Propulsion Laboratory. Palomar-Leiden minor planets. Asteroids Edition. University of Arizona Press. p 1034. 1989.
[64]Williams,J.G., “Asteroids Family Identifications and Proper Elements”.Icarus. 96,251. 1992a.
[65]Williams,J.G.,“Asteroids Family Identifications and Proper Elements”. Icarus. Sbm. 1992b.
[66]Zappala, V., Cellino,A., Farinella,P., Knêzevîc,Z., "Asteroid Families. I. Identification by Hierarchical Clustering and Reliability Assessment". The Astronomical Journal, 100, 2030. 1990.
[67]Zappala, V., Cellino,A., Farinella,P., Milani,A., "Asteroid Families. II. Extension to Unnumbered Multiopposition Asteroids" The Astronomical Journal, 107, 772. 1994.