TAXONOMIC EVIDENCE APPLYING INTELLIGENT INFORMATION ALGORITHM AND THE PRINCIPLE OF MAXIMUM ENTROPY: THE CASE STUDY OF ASTEROIDS FAMILIES

Gregorio Perichinsky[1] , Elizabeth Miriam Jiménez Rey[1] , María Delia Grossi[1] , Félix Anibal Vallejos[1,2] , Arturo Carlos Servetto[1] , Rosa Beatriz Orellana[3] , Angel Luis Plastino[4]

[1]{gperi,ejimenez,mdgrossi,aserve}@mara.fi.uba.ar

[2]felix_vallejos@yahoo.com.ar

[1,2]Computer Science Department - 4 th Floor South Wing - Databases and Operating System Laboratory - Faculty of Engineering - University 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]rorellan@fcaglp.fcaglp.unlp.edu.ar

[3]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

[4]Plastino@venus.fisica.unlp.edu.ar

[4] PROTEM Laboratory - Department of Physical Sciences - Faculty of Sciences - University of La Plata - C.C. 727 or (115 # 48/49) streets - (1900) La Plata - Buenos Aires – Argentina - Phone: (54 221) {483-9061 - 425-0791 (ext. 247)}

ABSTRACT

The Numeric Taxonomy aims to group operational taxonomic units in clusters (OTUs or taxons or taxa), using the denominated structure analysis by means of numeric methods. These clusters that constitute families are the purpose of this series of projects and they emerge of the structural analysis, of their phenotypical characteristic, exhibiting the relationships in terms of grades of similarity of the OTUs, employing tools such as 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 and in this way the significant concept of spectrum of the OTUs is introduced, being based the same one on the state of their characters. A new taxonomic criterion is thereby formulated and a new approach to Computational Taxonomy is presented, that has been already employed with reference to Data Mining, when apply of Machine Learning techniques, in particular to the C4.5 algorithms, created by Quinlan, the degree of efficiency achieved by the TDIDT family's algorithms when are generating valid models of the data in classification problems with the Gain of Entropy through Maximum Entropy Principle.

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

Introduction

Dynamics of Complex Systems: Tools of Statistical Mechanic and Computer Science

The study of the complex systems in an unified outline has been recognized as a new scientific discipline, inside the context of the multidisciplinary fields. This type of systems includes areas so diverse as ecosystems, computers, the human society and its economy, the climate or the physical-chemical systems. The tools for the study of these systems are varied and in the present project are used in such way that combines techniques of statistical mechanic and computer science. The areas to cover are three: description of physical-chemical systems and human systems, developments of the foundation of the used theories, developing and application of new computation tools.

For the human systems [2],[49] it suits to define what type of modelling of the system is carried out. Although there are several ways of defining that it is a model, we will try to give the simplest that is to say that a model is to give a formal frame to the hypothesis set that they arise of a certain observations set. Often these hypotheses look for only to identify the functions that allow reproduce the observed data. We will say then that we have an empiric model. In others it is to identify the mechanisms that it is supposed they generate certain data, and to generate "predictions" about the behaviour of a certain system. In all the cases the model is a representation of the system. A system is a set of entities that have been linked among them by means of relationships. The development of a model is a characteristic process of test and error, and it is developed above the base of the real world. By means of simulations and/or analysis of the model, it is seeking to obtain results that reproduce the data. Axial for example, for the case of the political-economic-social systems, it is possible to define a general strategy for the construction of the quantitative model, or at least quantitative semi dynamics macro evolutions in the society, using diverse concepts coming from statistical mechanic. The qualitative definitions on the individual and collective human behaviour can associate to micro and macro variable of the physical systems, any quantitative model cannot be done without the qualitative definitions that characterize to the social or political behaviour whose modelling is pretended to make.

Objectives

Study of temporal evolution of complex systems in a unified scheme with tools of statistical mechanics and computational. Dynamics and taxonomics description of physical-chemical system. Dynamics and taxonomics description of economic-political-social systems. Relative aspects to the foundation of the statistical mechanics and the quantum mechanics. Development and application of computation algorithms.

Methods of the quantum mechanics and statistic. Principle of maximum entropy.

The principle of maximum entropy (PME) it is broadly not only applied in physics, but also in meteorology, genetics, and in general in processes of any nature where it is wanted to obtain information starting from an incomplete of data set, or using the smallest quantity in previous suppositions. For the case of the physical systems, the PME provides an alternative formulation of the statistical, elegant and compact Mechanics, formulated by [26]. In its proposal, Jaynes introduces to the PME as a canonical method to build main density, in terms of the variables whose means values are known "a priori". However this constructive way doesn't allow assuring that the main opposing density can reproduce means values of other magnitudes of the system in study that they are not part of the information "a priori". This misstatement went the biggest it criticizes to the re-formulation of its ME in terms of the PME. Later on [3], it removes this limitation (with a quantum formulation, applicable also to the classic case) giving a method specifies to find all the operators "outstanding" of the system, this is all the necessary operators to describe the dynamics of the system completely. The operator density in this way built, is valid for temperature different from zero and outside of the equilibrium. The works that the Laboratory of Dynamic Systems has carried out during the last years follow this methodology.

The formulation of Levine is based in that the outstanding operators are those that satisfy the relationship of closure as a property of the algebra, with the Hamiltonian of the system. The main one criticizes to this methodology it was that the Hamiltonian that complete it is generally simple.

Fortunately, we could remove that limitation solving, with the PME, the dynamics of non trivial Hamiltonian, as that of Jaynes-Cummings [21] as well as two levels coupled to a finite and discreet

thermal bathroom [4]. In these cases the associate algebras of operators are infinite. The PME allows finding the temporal evolution of operators' means values, without using the wave function, through the generalization of the theorem of Ehrenfest [40] that drives to a system of differential equations. The Hamiltonian can also be dependent of the time [40]. For the form in that they are built, these systems of equations, don't allow initials conditions arbitrary use, it is necessary to find the partition function through obtaining the diagonal of the density matrix, [4] to be able to build a coherent set of initials conditions. Obtaining this diagonal allows to describe the dynamics of the system in the space of the multipliers of Lagrange associated to each operator, space denominated by us as "dual". This space can be thought as a space of the phases for quantum systems, being conserved in the (even when the multipliers are real numbers) commutation relationships [4]. All the formalism derived for quantum systems is applied directly to classic systems so that the commuters are replacing by brackets of Poisson. The Theory of the Information had been developed by Shannon to be applied to the field of the communications. It starts with the existence of a set of numerable events and of a space of probability, in which each event has a defined probability $p = \{p_1, p_2, \ldots, p_n\}$ that is normalized

$$\sum_{i=1}^{n} p_i = 1. \quad \textbf{(1)}$$

It is possible then, to define the information (I) associated with this distribution of probability, or the ignorance related with this last one, before knowing it (S)

$$I \equiv S = -\sum_{i=1}^{n} p_i \ln p_i. \quad \textbf{(2)}$$

and it is denominated Entropy of Shannon. If it is considered the case of a physical system, it is possible to express the entropy of the system as

$$\boldsymbol{S = -k_B\, Tr(\rho \ln \rho) = -k_B \ln \rho}, \quad \textbf{(3)}$$

where k_B is a constant that one adds to the definition expressed by the Ec. (3) to the effects of giving physical units to the entropy. Von Neumann was the first one in associating S with the entropy of the state described by the operator ρ when taking k_B similar to the Constant of Boltzmann (k_B = 1.38 x 10^{-16} erg/ºK).

As it was indicated in the previous paragraph, the Operator of Density ρ determines the physical state and the entropy of them. This state is partially characterized by means of the knowledge of a certain number of outstanding observables for the physical problem of interest. Only in case the operators make a Complete Set of Observables that they commute, the determination of the state it is univocal and the entropy value is zero. In the case of partial information, the knowledge of the values means of a limited number of operators will imply the existence of different Operators of Density that they satisfy the conditions imposed by the Ecs. (1)-(2). It arises, consequently, the problem of the election of one of these Operators of Density as representation of the physical state. It is in this point that [26] is introduces in the theory, the Principle of Maximum Entropy: given an observables set $\{\hat{O}_1, \hat{O}_2, \ldots, \hat{O}_n\}$ whose value means,

$$(\hat{O}_i) = Tr(\rho\, \hat{O}_i),\ i=1,\ldots,n, \quad \textbf{(4)}$$

they are the only information that one has of the physical system and that they will be denominated Outstanding Operators, the Operator of Density of the system is that which maximizes the entropy, defined through the Ec. (3). The Operator of Density that satisfies this condition is obtained by the Method of the Multipliers of Lagrange.

$$\rho = \exp\left(-\sum_{i=0}^{n} \lambda_i\, \hat{O}i\right) \quad \textbf{(5)}$$

where $\hat{O}_i$ is the Operator of Identity that is added to the initial set, to the effects of satisfying the condition

$$\textbf{Tr } \rho = 1. \quad \textbf{(6)}$$

Using the Ecs. (5) and (6) it is possible then, to relate the entropy of the system with the means values of the operators

$$S = k_B \sum_{i=0}^{n} \lambda_i\, (\hat{O}i) \quad \textbf{(7)}$$

Of here in more it will be considered $k_B = 1$. The means values and the Multipliers of Lagrange are related by the Equation

$$\lambda_0 = \ln Tr \left[\exp\left(-\sum_{i=0}^{n} \lambda_i\, \hat{O}i\right)\right] \quad \textbf{(8)}$$

Being obtained

$$(\hat{O}_i) = \frac{\partial \lambda_0}{\partial \lambda_i}, \quad i=1,\ldots,n, \quad \textbf{(9)}$$

of Lagrange.

The results exposed precedently was presented to be applied to a variables set of the system whose value mean are of interest [26]. These values mean were averages of classic observables related with the system. In the previous section it has denominated them to him "operators" because these results can extend without difficulty to quantum operators. The set of operators used is formed with the variables that, a priori, they seem outstanding.

If a posteriors of the study of the system it is observed that it is necessary to incorporate some operator to this set to allow a more guessed right

description, the initial set is redefined. This method makes impossible the deduction of results, since doesn't allow to distinguish when a not-prospective result is product of the lack of some operator or it constitutes a new result of the model in study. These limitations of the theory were overcome [3], since they extended it to sets of quantum operators that they can or not to commute to each other and they also elaborated not a constructive method that allows, not only to determine which the set of interest associated with a given physical system is, but also to endow to the dynamics, of a set structure of Lie.
To introduce these new concepts it is convenient to work with the logarithm of the Density Matrix,

$$\ln \rho = -\sum_{i=0}^{n} \lambda_i \hat{O}i \quad \textbf{(10)}$$

that it also fulfils an equation of the type,

$$i\frac{d \ln \rho}{dt} = \left[H(t), \ln \rho\right] \quad \textbf{(11)}$$

Replacing the Ec. (10) in the Ec. (11) it is proven that this will be valid for all time, if the commute of the observables ***{Ô₁, Ô₂,..., Ôₙ}*** with the Hamiltonian H(t) it satisfies

$$\left[H(t), \hat{O}i\right] = i\sum_{i=0}^{n} \hat{O}i\, g_{li} \qquad \textbf{i=1,...,n (12)}$$

where g_{li} are complex numbers that are interpreted as the constants of structure of a semi-algebra of Lie. If the initial set doesn't fulfil the condition (12), they will incorporate to him all the necessary operators to satisfy it. Those (n + 1) x (n + 1) elements g_{li} the matrix G conforms, and they establish the dynamics of the physical system, since as it will be seen, they determine the equations of evolution of the Multipliers of Lagrange and of the means values of the Outstanding Operators. Adding the closure condition from the semi-algebra to the maximization of the entropy has an important effect since it allows to obtain, for a Hamiltonian of a physical system of interest, a complete set of Outstanding Operators by means of the application of a canonical procedure. The Ecs. (11) and (12) they form a coupled set of differential equations for the Multipliers of Lagrange,

$$\frac{d\lambda_i}{dt} = \sum_{j=0}^{n} g_{ij}\, \lambda_j \quad \textbf{, i = 1,2,...,q ,} \quad \textbf{(13)}$$

to those that are added the initial conditions $\lambda_j(t_0)$, compatible with the Ecs. (4)-(9). For the case of independent Hamiltonian of the time, the coefficients g_{ij} are also independent of the time and the Ecs. (13) they become a system of differential equations at constant coefficients. In this case the solutions are of the type

$$\lambda_j(t) = \sum_{i=1}^{K} \exp(r_i t) \sum_{m=0}^{\gamma} a_{im}^{j}\, t^m \quad \textbf{, (14)}$$

where ***K*** is the number of roots (r_i) different from the corresponding secular equation, a_{im}^{j} are constant to determine starting from the initial conditions and ***γ + 1*** are the r_i multiplicity. This same discussion can be applied to the means values of the operators using the Theorem of Ehrenfest [Ec. (5)]. If the Hamiltonian is independent of the time, when using the Ec. (12) it is obtained

$$\frac{d < \hat{O}_i >}{dt} = -Tr\left(\rho \sum_{j=0}^{N} \hat{O}_j\, g_{ji}\right) = -\sum_{j=0}^{N} < \hat{O}_j > g_{ji} \quad \textbf{(15)}$$

that is to say, the Theorem of Ehrenfest in function of the constants of structure of the algebra g_{li}.

Computational Algorithms

The brain processes the information through the neurons, cells, able to make decisions that communicate dynamically (synapsis). In 1948, Mc Culloch, among other, began to develop mathematical models to imitate the operation of the human brain, task that continuous in uninterrupted form until the present, originating the theory of neural networks [7],[9], where they converge the mathematics, the neuron-anatomy, the theory of the information, the psychology, computer sciences, and the theoretical physics [10]. Thus it results that neural networks (NN) is a model neuron-physiologic of the cognitive activities, a dynamic not-linear system, a computational structure, etc., depending who is the specialist that is defining one [31]. It is as much a multidisciplinary field of high incidence in the promotion of the knowledge as in the applications to several areas. The association of concepts in systematic form, with numeric variables, of classification conforms the numeric taxonomy [8], [48] it disciplines defined as the numeric evaluation of the resemblance and similarity between taxonomics units (taxon) and the cluster of those units into taxa (plural of taxon), based on the state of their characters. The search of classificatory concepts that they allow a classification structure that doesn't modify it neither because of the adding of new information (stability of the classification and the taxonomics evidence) [32] [11] [33] [34] [35] [37] [38] [39], nor it is altered for the incorporation of new entities, it motivates us to look for new analytic tools. Thus, we try to develop techniques based on the Theory of the Information and a classification

technique was investigated whose foundation is the numeric taxonomy. The Taxonomy in celestial bodies, asteroids in particular, it is a fundamental topic in the Celestial Mechanics per a variety of reasons, among those that it is enough to mention that they constitute a Natural Laboratory for the study of the CHAOS, on one hand, and that they are remainders of the time of formation of the solar system, per other part [36] [38] [39]. The knowledge applied in the process of generation of navigation plans in autonomous intelligent systems (exploration robots [5]) that move in an unknown territory can take a wingspread, that is criticized the consumed time, taken the decision in the next action to be executed [14] [15] [16]. In this context the neural networks and the genetic algorithms arise as a solution alternative to some of the navigation problems [17] [18] [19], in particular those of handling of obstacles and detection of local steps, for those that the robot could exhibit a behaviour reagent [27]. The genetic Algorithms also arise like an alternative to the classic models of treatment of images, we obtained encouraging results to the date [11].

Methodology

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**[36]**.**
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[8][20][48].***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) [36] with the Maximum Entropy Principle (MEP).
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 noticies a clustering of asteroids in stabilized families [50].
In our case, is worked in an interdisciplinary way in Celestial Mechanics [50], Theory of the Information [1][22], Neural Networks[13] and Dynamic Databases [36] and the Algorithmic of the Numerical Taxonomy [8] [48], 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 [36] 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 [36].
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.

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 [28] [[41]..[47]]. 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[45][46]. 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 generates 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.

Constructing the decision trees

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 these 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.

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[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.

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 [25].
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 branches 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 over fitting of the data.

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.

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.

Requirements engineering.

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 [36].
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) [36]. 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 [36].
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 probabilistic 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 judgments 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.

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 hardiness 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.

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 [36]. 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.

Numerical Taxonomy.

We infer an **a n a l o g y** of the **taxonomic representation** [36] **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[8] [48].
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 [36].

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 [29].
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 $\overline{X'ij}$ = (Xij - $\overline{Xj}$) / σj **(16)** is a normalized variable the one which represents the deviation of Xij with respect to their mean in units of σj [8].
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 σ^2_d , we express the principle of minimal square.
It will be g (Xij) a not negative function of the variable Xij, for all k > 0 will have to be the probability function:
If g (Xij) = (Xij - $\overline{Xj}$)^2 , K = k ^{2} σj^2 **(17)** , obtaining for all k > 0 the inequality from Bienaymé-Tchevicheff:
P (| Xij - $\overline{Xj}$ | ≥ k. σj) ≤ 1 / k^2 **(18)**
This inequality shows that the quantity of (OTUs) mass of the located distribution would be of the interval
$\overline{Xj}$ - k. σj < Xij < $\overline{Xj}$ + k. σj **(19)**
it is to what is maximal value equal to 1 / k^2, giving a utilization idea of σj as measure of the dispersion or concentration.

Clusters and Spectra.

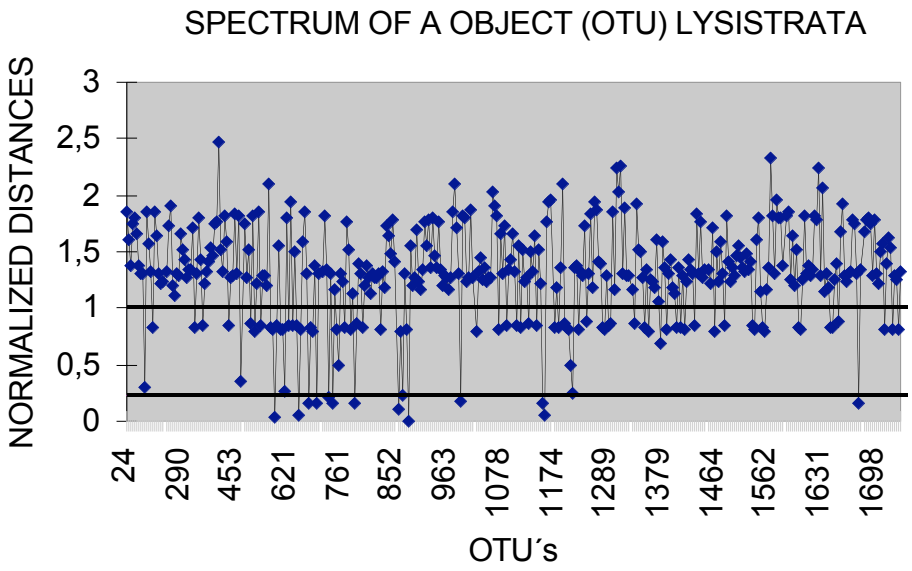
In discussing Sequential, Agglomerative, Hierarchic and No-overlapping (SAHN) [48] clustering procedures we make a useful distinction between the three types of measure.
We shall be concerned with clusters **J**,**K** and **L** containing **tj**, **tk** and **tl** OTUs, respectively, where **tj**, **tk** and **tl** all **≥ 1**. OTUs **j** and **k** are contained in clusters **J** and **K**, and **l ∈ 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 [36], 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.

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.

Algorithm.

The algorithm entails building up the data matrix, normalizing it, constructing the matrix of similarity, the spectra of objects and families formed by clustering, and, finally, performing an analysis of the pertinent invariants.

I. Data Matrix
II. Characters Selection
III. Construction of attributes domain
IV. Normalization
V. Distances analysis
VI. Matrix of Similarity
VII. Dispersion analysis
VIII. Identification of OTUs in the clusters
IX. Analysis of Invariants
X. Characteristic Spectrum of Objects
XI. Characteristic Spectrum of Families (Clusters)
XII. Iteration around the center (centroid)

Invariants:

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

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.

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:

$$H(S_i) = -p_i^+ \log p_i^+ - p_i^- \log p_i^- \quad \textbf{(20)}$$

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

$$p_i^+ = \frac{n_i^+}{n_i^+ + n_i^-} \quad \textbf{(21)}$$

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 ***(20)*** for any type of examples, we obtain the general formulation of the entropy:

$$H(S_i) = \sum_{i=1}^{n} - p_i \log p_i \quad \textbf{(22)}$$

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, … , j , … ,n , then, the total entropy of the system of subsets will be:

$$H(S, at) = \sum_{i=1}^{n} P(S_i) \cdot H(S_i) \quad \textbf{(23)}$$

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:

$$P(S_i) = \frac{|S_i|}{|S|} \quad \textbf{(24)}$$

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

$$I(S,at) = H(S) - H(S,at) \quad \textbf{(25)}$$

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.

Numerical Data

The decision trees can be generated so much as discrete attributes as continuous 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.

Results and Conclusions

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.

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.

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 avoids 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 [30]. 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.

Corollary

From what has been said, the work uses the Sequential, Agglomerative, Hierarchic and No overlapping 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]Acedo, C.F., Proto, A.N., Proceedings of NeuralP97, Neural Networks and their Applications. Theory and modeling. Marseilles, March 12-14, 1997.

[3]Alhassid, Y. and R.D. Levine, J. Chem. Phys. Quantum Formulation and Classic Case of Principle of Maximum Entropy 67, (1977) 4321.

[4]Aliaga, J, Crespo, G. and Proto, A.N. Aliaga, J. and AN. Proto. Principle of Maximum Entropy and Dynamics of non trivial Hamiltonian. Phys. Letter. A142 (1989) 63. Phys. Rev. A August (1991). Lett.70 (1993) 434.

[5]Bares, J.; Hebert, M.; Kanade, T.; Krotwow, E.; Mitchel,T.; Simmons, R, y Whittaker, R. 1989. Assembler: An autonomous Robot for Planetary Exploration. IEEE Computer Vol 22. Ner. 6, pags 18-26.

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

[7]Caianiello, E. R. (Ed.). 1988. Parallel Architectures and Neural Networks Strongly Connected. World Scientific (Ed.).

[8]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.

[9]Domany, E., Hemmen, 1. L., & Schulten, K. 1991. Model of Neural Networks. Springer-Verlag.

[10]Erickson, G. and Ray Smith, C. (Eds.). Maximum-Entropy and Bayesian Methods. 1989.

[11]Fernandez, V., Garcia Martinez, R, Rodriguez, L. & Gonzalez, R 1996. Genetic Algorithms Applied to Clustering. Proceedings of the International Conference on Signal and Image Processing. Pages 97-99. Orlando. Florida. Noviembre.

[12]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.

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

[14]Fritz, W.; Garcia Martinez, R.; Blanque, I; Rama, A; Adobatti, R Y Sarno, M. 1989. The Automous Intelligent System. Robotics and Autonomous Systems. Vol 5, nro. 2, pags. 109-125. Elsevier.

[15]Garcia Martinez, R Heuristic theory formation as a machine learning method. 1993a. Proceedings VI International Symposium on Artificial Inteligence. Pages 294-298. Editorial LIMUSA Mexico.

[16]Garcia Martinez, R 1993b. Measures for theory formation in autonomous intelligent systems. Proceedings RPIC'93. Pages 451-455. University National of Tucuman. Argentine.

[17]Garcia Martinez, R 1996. Planning while Learning-by-Interaction Systems: A Theoretical Approach. Proceedings del II Internacional Congress on Informatics. Pages 410-416. Buenos Aires.

[18]Garcia Martinez, R & Borrajo Millan, D. 1996. Unsupervised Machine Learning Embedded in Autonomous Intelligent Systems. Proceedings of the XIV International Conference on Applied Informatics. Pages 71-73. Innsbruck. Austria.

[19]Garcia Martinez, R y Borrajo, D. 1997. Planning, Learning and Executing in Autonomous Systems. Lecture Notes in Artifical Intelligence. Springer-Verlag. 1997.

[20]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.

[21]Gruver, J.L., Aliaga, J., Cerdeira, H.A. and Proto, AN. Principle of Maximum Entropy for Temporal Evolution of Operators. Phys.Rev. A 50 - A 184335 – b A 190 - (1994) 5274 A 190 - 363 c .E. 51 (1995) 6263.

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

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

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

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

[26]Jaynes, E.T., Phys. Rev. A canonical method of Principle of Maximum Entropy 106 (1957) 620; 108 (1957)171.

[27]Mahadevan, S. y Connell, J. 1992 Automatic Programming of Bahavior-Based Robots using Reinforcement Learning. Artificial Intelligence vol 55 pages 311- 365.

[28]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.

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

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

[31]Muller, B. & Reinhart, J. 1991. Neural Networks. Springer-Verlag.

[32]Perichinsky, G. 1989. Multiple states of multiple state automata to key fast validation. 11th. International Symposium Computer at University. Catvat. Zagreb. Yugoslavia.

[33]Perichinsky, G., Jimenez Rey, E. & Grossi, M.D. 1998. Pages 191-195. Domain standardization of operational taxonomic units (OTUs) on dynamic databases. Proceedings of the XVI International Conference on Applied Informatics. Garmisch-Partenkirchen. Germany.

[34]Perichinsky, G., Jimenez Rey, E. & Grossi, M.D. 1998. Pages 165-168. Spectra of objects of taxonomic evidence on the dynamic data bases. Proceedings of the XVI International Conference on Applied Informatics. Garmisch-Partenkirchen. Germany.

[35]Perichinsky, G., Jimenez Rey, E. & Grossi, M.D. 1999. Application of Dynamic Data Bases in Taxonomy Astronomic. Proceedings of the XVII International Conference on Applied Informatics. Pages 120-126 .Innsbruck. Austria.

[36]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.

[37]Perichinsky, G., Jimenez Rey, E., Grossi, M.D., García Martínez, R. & Proto, A., 2000. Knowledge Discovery Based on Computational Taxonomy and Intelligent Data Mining. VI Argentinean Congress of Computer Science, CACIC, CD. National University of San Juan Bosco. Seat of Ushuaia. Argentina.

[38]Perichinsky, G., Orellana, R., Plastino, A.L. 2002. Spectra of Taxonomic Evidence in Databases.III. Application in Celestial Bodies. Asteroids families. Pag. 212-226. International Association for (ACIS) Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications. Institute (SEITI), Central Michigan University. Foz do Iguazú. Brazil.

[39]Perichinsky, G., Servente, M., Servetto, A., García Martínez, R., Orellana, R., Plastino, A.L. 2003. Taxonomic Evidence Applying Algorithms of Intelligent Data Mining. Asteroids families. (pp 308-315). International Association for (ACIS) Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications. Institute (SEITI), Central Michigan University. Río de Janeiro. Brazil.

[40]Proto, AN., Maximun Entropy Principle and Quantum Mechanics. Condensed Matter Theories, Vol. 5 Valdir Aguilera-Casaca Ed. Plenun Press. 1989.

[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.

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

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

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

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

[49]Weidlich, W Phys. Rep. Modeling Formal Frame Prediction 204 (1991)1.

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