JOURNAL OF ENGINEERING annals OF FACULTY OF ENGINEERING HUNEDOARA

EPISTEMOLOGY TO INVESTIGATE THE SCIENCES - EXPLANATION OF THE COMPUTER SCIENCE: CASE OF INTELLIGENT DATA MINING

Gregorio PERICHINSKY

University of Buenos Aires (UBACYT), Faculty of Computer Science
National University of La Plata, La Plata - ARGENTINA

ABSTRACT
This work is not conventional in strict sense, because it tries to the Scientific Investigation on bases epistemologics, because in methodical form it has allowed the man to increase its knowledge in exponential form. The application of the scientific method is based on the observation, the experimentation and the verification of the empiric base, as much in its information as in its working hypotheses, as formula with the one that has been achieved a more and more clear, more and more precise and more and more wide understanding progressively. This empiric base of the knowledge is what allows being, objective speculation, verifiable and refutable that when chaining terms and enunciated mixed it forms theories, with the scientific method, when completing its stages or phases of the investigation and the crucial fact that it is to contrast the hypotheses. The problem of the explanation [12], it is the main motivation for the formulation of scientific theories, able to explain events that intrigue the scientists who want to understand. The center of gravity of the epistemology and the methodology is the contrast operation and of prediction.

Keywords: Computer Science, Epistemology, Scientific Investigation, Intelligent Data Mining

INTRODUCTION

The knowledge is not infallible, consequently it can be discussed, it can be ratified or it can be rectified, but following the original methodological rules or specifying the reasons of its modification appropriately and when responding with the conclusion. "Science is the rational search of the knowledge, when managing variables in form differentiate in the laboratory investigation and the field investigation. It is a long trip of the intellect that it was born with the man and it is evolutionary alive material."

It is a Body of Doctrine methodologically acquired, objective in the facts, interpretive in the laws, deductive in the hypotheses, on their epistemologics bases and speculative in the theories. It is to create laws that include the problems, in and of itself they group sets disjuntives, according to the grade of deepening of the topic, solving the dilemma or pair <knowledge, practice> [25] [26], when basing the prediction and the explanation of the phenomena, essential operations of the science [12].

Science is the Group of obtained knowledge by means of the observation and the reasoning, systematically structured to deduce principles, theory and general laws with regularities [25] [26] [11] [14]. The tasks of Investigation and the explanation of the problems, the theoretical mark and the role of the hypotheses and the reality, those enunciated should be true, laws and data are used (to explain false facts doesn't make sense).

The scientists discusses the facts that are presented them with a paradigm that constitutes them and it articulates [15] [8]. When a scientific revolution takes place the "reality" becomes something new, because the facts are articulated with the new paradigm. The history of the science and the technology, are a long and clear description of how the technical means and the procedures progress with the increase of effectiveness and operability, in spite of that with the time, the paradigms are substituted each one to other [14]. It is an accommodation process it is not an assimilation process, the evolutionary processes should reach a new equilibrium state, in which the organism recovers their abilities. The functionalism [17] [24], homeostatic behavior of a functional system where the alteration of the characteristic attributes of its operation, a process would allow to the system to take place a recovery.

It is necessary to explain by causes and by reasons (premises and laws) and it is seen that the advance of the paradigms is not continuous neither spasmodic [25] [26] [11] [14], when treating the three scientific ERAS, Previous (Aristotelian), Scientific (Newtonian) and modern (Einstein), converging in the first 30 years of the XX century, with the Cybernetics, the Information Theory and the Natural Languages (Man-Machine relationship) and the metaphor established as theory of a set of equations in the Computer Sciences until what is trying to Explain in this paper, with empiric terms and theoretical terms, the Intelligent Data Mining.

1. INVESTIGATION PROCEDURES

The Investigation is conformed by Procedures that contain Actions and Effects that finally have to extend the scientific knowledge, so much in the "laboratory" like in the "field", when carrying out activities intellectual and experimental in a systematic way with that purpose on a certain matter, in laboratory, field and experimentation ex-post-facto [11] [14].

The laboratory experimentation allows a major control of the variables; invent situations and cases artificially to analyze. The formal paradigm establishes the relationship among elements, as control of an only element or variable that it acts in function of other and the rest of the elements remain constant. The inconvenience is that it is difficult to create artificial conditions without one leave from the reality.

To determine properties and relationships among several components it is necessary to carry out a Factorial Experimentation, that as they are field experiments they take components being of the reality and studies in scale. The researchers enter in a group or organization or institution, they take direct contact with the processes and interactions of the groups, in and of itself they are more controlled.

Experiments ex-post-facto is those in which the variables are not manipulated, due to impediments sometimes of ethical nature and sometimes of technical type. The involved phenomena can be social, economic, historical and astronomical, with qualitative and quantitative variables. The first ones respond to classificatory criterion, the second ones allow some correspondence of numeric order, and in turn they can be divided in ordinal variables and metric variables, subject to measurement scales. The clinical method [11] it is characterized to make deep studies of singular cases, without being a considerable number, comparative with the experimental method, existing differences and rivalries among them, when being related with a diagnostic and therapeutic objective (i.e. developed in psychology and medicine) [9] [27].

The properties to consider in each case are multiple and complex, requiring of a researcher with ability in the analysis of the interactions, and they are always nearby individuals or different and singular people that interact with the researcher.

They are complex properties those that interact and not quantifiable variables.

As long as the experimental method, requires great quantity of cases to evaluate a reduced number of properties, to assure the validity, the objectivity and the experimental dependability, by means of an appropriate design technically; they can be used as much as in exploratory methods as much as for contrast the hypothesis.

An analogy, in an research or in a scientific community, conforms it "Empiric Base", in one an Object, Entity or Situation is a "Datum", or if it is captured it is an "Observation" [14]. Object "Direct" or empiric in the Empiric Area, and "Theoretical Indirect" in the Theoretical Area.

1.1. Dynamics of the knowledge

Of the analysis in the working mode in the sciences, in the search of the knowledge, the classification is the dynamics of the knowledge, being for the observer an action, or procedure or set of procedures, where it is carried out the action.

This analysis is descriptive and normative, an activity of methodological rules, process of the scientists' investigation, interpretation that verifies the hypotheses, is a model so much formal as factual, or theory on part of the reality of a certain system.

It begins with a disconcerting fact or state of disorientation animus and perplexity, it misses of measure in the actions or words of the aspect that it presents a problem for the observer, issue to clarify, because they are a set of facts, circumstances or proposition that hinder the attainment of some objective. Proposed by Aristotle, the Episteme of the Greeks, it based the knowledge by means of methodological rules, until the English empiricist of the XVII-XVIII centuries, changes the concepts of the knowledge like fruit of the experience, contributing the inductive experimental method.

1.2. Inductive method and deductive hypothetic method

The moments of the Novum Organum [11] [14], (1) the collect or direct observation or experimental observation of the data, (2) the inductive method planning the generalization of the behavior models, of the collected data, that when being connected systematically conform sets part or levels, laws of maximum level that can be denominated theories and (3), to deduce or to predict facts or phenomena when applying the laws and theories, in a dynamic precedence. Because of the distortion of the reality, they give way all the prejudices and preconceived attitudes, called idols, they were already the common property of the species, due to ways common thought ("tribe"), biological, mental and cultural ("cavern"); either for an excessive dependence of the language and of the communication ("market") or of the tradition and philosophical inheritance ("theater"), dogmatism. For the neopositivism or logical positivism of the XX century [29] [30] [31] [32], the inductive method is a reduction of the investigation, it is an erroneous precedence, since the data are collected in function of the problem to solve and its generalization, generating theories without leaving of working hypothesis or a heuristic of superior level. One arrives this way, to the Hypothetical-Deductive model of investigation expressed in the introduction [25] [26] [11] [14] [29] [30] [31] [32].

1.3. Phases or stages of the deductive hypothetical model

The stages of the dynamic and holistic system conform a precedence graph that could be denominated phases of the investigation [25] [26] [49] [4], to be dynamic and not mechanical stages, with Moments, Sub-stages, Tasks or Actions, consisting until twelve phases or stages in different authors.

2. MAIN STAGES OF THE SCIENTIFIC INVESTIGATION

They are considered basic stages [11]: Problem, Hypothesis, Theoretical Mark, deductive Procedures, contrastable Consequences, Procedures of Contrast, Evaluation of the results and according to the result the hypothesis is verified or a new hypothesis is generated refuting the previous one. What gives origin to the tasks of Investigation is the problem, the theoretical mark and the role assigned to the hypotheses and the reality and not, the collection of data.

2.1. Problems

Of the problems questions are formulated and it is necessary to try to respond them or to explain to them, transcending the context of the knowledge of the state of a discipline, regarding the relative reality. The implication and the generalization of the problems and questions can be ordered in grades: gradation [25] [26].

2.2. Theoretical mark

In the researches the theoretical mark, with its components of one or more theories, homogeneous or heterogeneous, is present through its hypotheses, conditioning to the queries or questions that are formulated, and to be interested in something or to advance in certain address.

2.3. Hypothesis

By means of a hypothesis or conjecture once formulated with clarity the problem that will investigate and you will proceed to look for their solution.

The set of connected ideas requires Theories and working Hypothesis formed by relationships of compatibility and implication that seek to understand and to explain a certain domain of the reality. The working hypotheses [14] they are stratified in three levels: (1) Level 1: Individuals' description or objects (artifacts) of low-level that they describe, analyze, register, enumerate and they attribute properties. The objects take relationships of the Empiric Base formed by groups of entities, phenomenologies, properties and scheduled relationships; (2) Level 2: Intermediate level of observations that generalize, correlates, submit and they classify; it is a level pre-theoretical and, (3) Level 3: Hypothesis of maximum level and observations that explain, predict, they understand, they systematize, they invent solutions and methodologies. This task result to have many times heuristic value that is to say, it contributes to stimulate the scientist's creativity.

The mechanisms of production of ideas and of resolution of problems, "logic of the discovery", they appeal to procedures of the formal logic, the calculation of predicates or the theory of groups and logical-mathematical systems are usually used, chained premises, as long as theoretical mark.

2.4. Observational consequences

They are observational consequences or contrastable consequences those enunciated inferred deductively of the hypotheses, susceptible of confrontation with the experience. The language is observational, not theoretical. If they were already enunciated about estates, facts or relationships well-known, the hypotheses explain to them.

2.5. Contrastation

The procedures to contrast the observational consequences are a crucial stage of the scientific investigation. The forms that it can acquire this stage have very different modalities, characteristic of the different investigation techniques, the systematic observation, the experimentation, the administration of tests (proofs), and the realization of surveys and recording the interviews and collecting the data statistically processed.

Experimental procedures are applied in those that the modifications are analyzed in the values of the independent variable related with the dependent variable, employees in experimental sciences. The application moments are, (1) the design of the experiment, (2) the realization of the experiment and environment (laboratory or field), and (3) the registration and evaluation of the obtained results. Instead of the experimentation they are making systematic observations or an experimentation ex-post-facto. The enunciated of a law it is one empirically confirmed general hypothesis, inside a theory, that in the hypothetical-deductive system represents an objective regularity [4]. [5] [6]. Most of the enunciated theoretical, alone accepted by an agreement of the scientific community [16]. The propositions, are premises (data) or consequences (theorems) that in whole are a very organized theory, by which must be axiomatic, they are systematically united by the deductive relationship (syntax) and some common topic (semantics).

A model is a medium-range theory because it doesn't has neither the class reference, nor an enunciated that it is a law. [4]. The theories that contain probabilist laws and those that require clauses ceteris-paribus are not refutable, these last ones don't require more factors because the auxiliary hypotheses will complete them, complementarily, in the phenomenon in study, anyway always you can sustain the theory (factual) [7] [33] [34].

2.5.1. Falsificationism

In the evolution of the constrasting, regarding the model hypothetical-deductive epistemologyc, emerge three models planning as more rational [29] [30] [31] [32] [16] denominated falsificationism or refutationism [14]: dogmatic falsificationism, naive falsificationism is the hypothetical-deductive simple and sophisticated falsificationism is the hypothetical-deductive complex, models proposed for the scientific investigation.

The falsificationism and the neopositivism [29] [30] [31] [32] [16] when considering the rational character of the scientific investigation, they reject the justification of the knowledge, where the statements are always demonstrated empirically, contrasting condition or verification.

2.5.1.1. Dogmatic falsificationism

In the Dogmatic Falsificationism there are new hypotheses and they are contrasted rigorously, for scheme of Programming of Scientific Investigation (PSI).

The falsificationism [16] it is metacientífico, because it proves the thesis with the hypothetical-deductive model, concluding with, (1) a net demarcation, among enunciated theoretical and enunciated observational, (2) those enunciated observational are demonstrated by experience, and (3) a theory is scientific if it has an empiric base, expert as the group of a potential falsification of the theory, verifiable for experience.

2.5.1.2. Naive falsificationism

Scientifically the falsificationism should evolve methodologically toward a model of Falsificationism Naive, one conservator of the scientific community [28] and another revolutionary, with changes in the refuted theories [29] [30] [31] [32] [7]; where (1) one for contrast through a confrontation between the theory and the experimentation, (2) an interesting result is to falsify (conclusive), refutations of scientific hypotheses, (3) but the science has shown that they are not simple criterion; in and of itself it is necessary to replace them for refined versions of the same principles, (4) the contrast is a triple confrontation between rivals theories and experimentation, (5) some experiments are of the confirmation more than of the refutation.

The scientific game is given if theories rivals compete in the explanation or the prediction.

Scientific development exists, if there are a competition of a sequence of theories that they share a hard nucleus (hard core), formed by hypotheses. A program of scientific investigation (PSI) it is a succession of theories related T1, T2, T3,..., Tn that ones have been going generating itself departing from the other ones. The scientific community defines a group of hypotheses commons that they form its hard nucleus and she declares it "irrefutable." Any experiment or observation will be able to falsify the hypotheses that compose this nucleus that constitutes the element of continuity of the program. The hard nucleus of all investigation program it is preserved by a body of auxiliary hypotheses that they form a "protector belt" around the nucleus.

2.5.1.3. Sophisticated falsificationism

The falsificationism naive evolves toward the sophisticated falsificationism, when replacing a theory by another in the same investigation program and on the other hand the evolution is a criterion that establishes that a theory T is false if and only if it has proposed another theory T', in those which, (1) T' it has more empiric contained than T, that is to say predicts new facts, unlikely in T or prohibited by T, (2) T' it explains the previous successes of T, that is to say the whole not refuted content of T is included in T', and (3) the excess contained of T', regarding T, it is corroborated. There is continuity in the knowledge and that communication exists among scientists that work in different investigation programs, if there are differences among theoretical and observational terms, it implies the inconmensurability of the scientific theories that compete [15] [8].

The model of the sophisticated falsificationism uses the stratified scheme levels [14] seen before, the auxiliary hypotheses and ad-hoc (post-hoc), it expects the auxiliary hypotheses are approved along the time, for its fundamental role in those "crucial experiments". A program (PSI) refuted it can be recovered with a positive creative development of its heuristic one.

An heuristic is positive if hypotheses are generated so that they protect to the theory and a heuristic is negative when a belt of auxiliary hypotheses is generated so that protect to the hard nucleus, of which hypothesis of the central theory of the Program of Scientific Investigation is replaced [14].

The hypothetical-deductive, dynamic and holistic model, [6], it rejects the whole falsificationism, because those enunciated basic they cannot be verified by observation or experimentation, it rejects the inductive logic, because any universal law will have probability zero of being demonstrated, because for infinite induction the scientific investigation doesn't achieve its truth, finally, it doesn't demarcate the Science of the Non Science, among theoretical and observational enunciated.

An Interior History of a discipline or scientific theory is distinguished when it includes variables that can change the theory, if the methodological questions indicate the elements of the External History, EBCP [6] {Economic, Biological, Cultural and Political (Vertehen und geisteswissenschaften ≡ Intuitive, Humanistic and Social understanding)}.

2.6. Hypothesis acceptance

If the results obtained in the empiric contrastation are favorable the hypothesis is corroborated and appropriate to the problem and the acceptance of the hypothesis takes place, but that acceptance will always have a temporary character.

2.7. Hypothesis rejects

If the obtained results were refutatories, it is necessary to determine which premise (or premises) they are responsible of that adverse result, and the rejection of the hypothesis takes place. It is the hypothesis in question, or some of the auxiliary ones that they intervene that they will be evaluated in independent way. If they were not refuted, it is false the initial hypothesis and it will be necessary to abandon it and to propose in their place a new one or one will be able to try to correct modifying it in some aspect, in their scope or in some of their terms.

3. REALISM AND SCIENTIFIC INSTRUMENTALISM (THEORIES)

To build the enunciated it is necessary to build a scientific language [1] that transforms all their terms, even the logical ones, in specific or technical terms, because the terms of the daily language are inadequate and; to build sentences that can be useful to express knowledge like that pointed out by the "instrumentalism", in the problem of the theoretical terms, forming complex expressions that allow to describe a state of things, observable or not.

The instumentalism considers that the theoretical terms although they are specific, they are not empiric neither logical. The constructivism uses the meaning "construct", to insinuate that a theoretical term is in fact a construction based on purely empiric aspects. In epistemology, theoretical and empiric terms are used as "enunciated bridge", linking to the theoretical environment with the observational or practical, they are "rules of correspondence" with the empiric base.

The scientists, to formulate hypothesis or conjectures, use the surprising method and until the deception that an artist uses, when when they imagine a work of art, the imagination power and of creation. They imagine what it can have "behind" of an appearance fenomenologic, they explain the behavior of a phenomenon [29] [30] [31] [32], they invent hypothesis and then they control them.

The activity of the epistemology like investigation of the science, are a dialectical situation of mutual learning, for the scientists. For Albert Einstein, the operational procedure, don't hide the meaning of the theoretical terms, related to the theory notion not to the operational definition, but the operational techniques are applied to introduce concepts.

3.1. The deductive method in the realism and the instrumentalism

The positivist ideas introduce theoretical terms if they increase the predictability of the theories, while a semantic divergence is abysmal. The instrumentalism affirms that the theoretical terms are verbal instruments without reference, without meaning as to build logical deductions, their enunciated are not genuinely hypotheses, and they are not true. They are pseudo-hypothesis or principles of an axiomatic system, rules of correspondence, alone of deductive character. The observational consequences are instruments to mediatize, enzymes or catalysts that allow building a "deductive reaction."

3.2. The realism, the truth, the reality and the metaphor

The realism admits that the theoretical terms make a purely instrumental sense, but there are cases in which entities are non-observable, real whereas clause to the related theoretical term, it allows to obtain a knowledge that transcends the empiric base, to know how the world is in their ontologic foundations beyond the accessible thing to our senses and instruments [5].

A metaphor of a theory can be imagined as a set of equations in which are similar the empiric terms or observable, because its meaning is known. The theoretical terms, would be the incognito ones that have to satisfy certain conditions. Discuss the facts just as they are presented through a paradigm that constitutes them and it articulates, but don't deny the reality [15] [8]. When a scientific revolution takes place it abandon the "reality" like was for to become in something new, because the facts, articulated by the old paradigm, disappear and they are replaced for "new facts." The theses "strong" [15] sustain that the concept of "truth", understood in absolute and Aristotelian sense, is completely useless in science that which is very serious, the philosophical position of the theory of the knowledge and what offers, more than ontology, is a thesis of gnoseology (coherentism). Each scientific community builds paradigms and decides its ontology [15] [13].

3.3. Scientific evolution, realism and instrumentalism for accommodation and equilibrium

"One cannot conceive to the scientific progress as successive approaches to the reality."

"The history of the science and, in particular, that of the technology, is a long and clear description of how the technical means and the procedures of the science to improve it show a progress, increase of effectiveness and operability, in spite of the fact that along the time the paradigms each other are substituted." [14]

It is an accommodation process and it is not an assimilation, it is characteristic of the stages of change in the evolutionary processes that it concludes when a new equilibrium state is reached.

Epistemologically, the psychological thought and the gnoseology [27] they are not different, it is observed: a) the cultural structure incorporates the facts to its system of to conceptualize and to surmise to the universe, b) a dysfunction when appearing a new system and a change of vision and description of the world; c) the equilibrium state is reached in the historical moment in that the new vision is unanimously accepted (major holistic reality and of objects), moments of the historical evolution.

As the science is developed and theories are formulated that some are replaced to others by the processes of assimilation, accommodation and equilibrium, the objects of those that it speaks each theory resemble each other more and more and they approach to what would configure the " real object", to reach.

4. THE PROBLEM OF THE EXPLANATION

The problem of the explanation [12], it is the main motivation for the formulation of scientific theories, able to explain events that intrigue the scientists who want to understand. The center of gravity of the epistemology and the methodology is the the operation of contrastation and of prediction. The inductivism doesn't explain neither it predicts, it proposes an inference type that allows to obtain generalizations with data and samples.

They are three essential operations of those that it is in charge of the science: foundation, prediction and explanation.

The foundation of an enunciated is to indicate the reasons for which it can be considered verified, and for the deductive hypothetical method, it is "sufficiently corroborated". The prediction, refers to observational consequences, it is not known if the enunciated is true, but the prediction offers elements that try to anticipate if in the future, it will occur in the described way. The prediction is weaker than the foundation, it doesn't prove the truth and it is not even equal to corroboration, it is necessary to verify and to establish that what was predicted has been completed and be admitted as knowledge; the observation has an essential performance.

The word "explanation" it is used in different senses and it is necessary to discriminate against them [14], this way, (1) to give action rules or pragmatic sense, (2) to give the meaning of a word, saying which their application is, (3) the scientific explanation is the one that attempts, of an enunciated that it should be true, to agree with, (4) the explanation of the nomology deductive is a model that explains for laws, it presupposes: a) deduction; b) what is deduced expressed the fact; and c) among the premises used in the deduction they figure laws, (5) the explanation of laws is when there is some regularity in the facts of a theory like derivative hypothesis, being the explanation the motor to build the scientific theories and they are used to deduce and to understand, the theory exists and if the theory doesn't exist and it is necessary to invent it to explain, (6) not to admit any speculation like explanatory theory and if it doesn't exist it is a potential explanation, as Motor of the history of the science that originates new theories (phenomenon of "change of theories"), (7) the explanation of facts complicates the explanation, not of a law, but of a singular fact that describes the intriguing fact to explain.

The alone laws don't allow to deduce factual aspects and to the inverse one, with data, without laws, it won't be possible to carry out the deduction.

Conclusion: explanations are not requested of what is false and the logical structure of the explanation scheme is the "symmetry principle between explanation and prediction". If one makes a prediction and this it is completed, then, automatically, it becomes explanation.

The Prediction is an event that will occur, in epistemologic felt it is the deductive connection among knowledge that are already possessed. In the scientific practice it is necessary to carry out predictions by means of theories and laws and to give up the prophecies.

The statistical model of explanation is a model where the laws are enunciated statistical or probabilistic that establishes regularity in their terms. In the genetic explanation, laws are not used, the pertinent facts are chained by precedence and if they are explanations teleologic, they explain a present fact, with something that will happen in the future.

In the functionalism [17] [24], it is a scheme with homeostatic behavior, functional system, where the alteration of variables or factors that characterize their operation, are a process that allows him to recover their structure and it is when it explains "for causes and for reasons" [48], by means of premise-laws and causal laws.

5. REDUCTIONISM

When it faced with uncertainty an important problem, it reduce, it is a posture philosophical denominated reductionism that tries objects or environments of certain nature that can be characterized in terms that are of another environment or different nature. In definitive they are of philosophical importance, ethics and metaphysical, it implies the nature of the reality.

The connection is advised the connection between reduction and explanation, if a procedure exists to reduce a discipline to other and, a theory to another of a previous discipline, where the laws of the discipline that it has been reduced, become derivative hypothesis of the theories of major range. The laws of a discipline will be explained by the laws or the theories from the basic discipline to which reduced the first one that will be besides reduced, explained, based on the successful theories of the fundamental discipline.

The thesis of the "ontologic reductionism" it is that when connecting the laws of a discipline with those of other, the objects that it treats a discipline will be reduced to entities and simplifying the complex thing, (see The Discourse of the Rene-Descartes´s Method), the "semantic reductionism" it doesn't reduce entities, it translates to the language of a basic discipline as a semantic and syntactic problem. The connection with the explanation is the "methodological reductionism" when reducing a basic theory to another reduced one that implies a semantic reduction of the language of a basic theory and it discovers that a theory is derived of another. For the deductive dependence of the "machine of deducing."

Hypothesis is formulated on correlations with the basic theory, the reduced theory and the group of rules of correspondence they link expressions of the vocabulary [23].

6. CASES OF EXPLANATION: INTELLIGENT DATA MINING

Most of the applications of the Artificial intelligence to tasks of practical importance build a model of usable knowledge for a human expert. In some cases, the task that the expert accomplish is a classification, that is to say, it assigns objects to categories or certain classes according to his properties [43]. In a classification model, the connection between classes and properties can be defined using diagrams of flow until complex procedures and desestructurates. If we restrict our discussion to executable models, represented with computational methods, they exist different ways of how to build a model. A model can be obtained commencing from outstanding interviews with experts. If classifications stored previously exist, they are examined to build an inductive model, by means of a generalization of specific examples. The systems ID3 and C4.5 belong to this second group [2] [53].

6.1. Theoretical Mark

When scheming the problem of learning of a model of data starting from examples from a theoretical mark, we find the following scheme [2]: Given:

- A set C of classes,
- A set E of preclassified examples

Find: One hipothesis H (set of sentences) such that:

$$\forall e \in E:\ H \cap e = c \wedge H \cap e \neq c' \qquad 6.1.1.$$

Where c is the class of the example e and c' $\in$ C-{c}. The obtained results are presented as:

- a decision tree,
- a set of decision rules.

The system will generate a decision tree, fruit of the nature of the algorithms of the TDIDT family. The tree of induction that has resulted will be built from the root toward the leaves (top-down). The generated model is very useful for the user since allows an easy visualization of the results. Also, it is transformed the tree into decision rules that can be used by other classification programs or being transformed in SQL sentences to classify new data quickly.

6.1.2. Input data

Before analyzing the TDIDT family it should be kept in mind that not all the classification tasks are appropriate for this inductive focus, the next requirements that are listed should be completed [20] [21] [22], [35] [41]:

- **Attribute-value description:** *Plane files.*
- **Predefined classes:** type {attribute-value$_1$, attribute-value$_2$,..., attribute-value$_n$, *class$_k$*}.
- **Discrete classes and disjoints,** given the nature of the decision trees, the classes should be discrete or become discrete in the case of being continuous.
- **Sufficient data,** model generated by inductive generalization are not valid if they cannot be distinguished of the casual ones.
- **The data of training can contain errors**: according to Mitchell, the learning methods using decision trees are robust in front of the errors, so much in the values of the classes like in the values of the attributes of the data of training [20].
- **The data of training can contain values that fault in the attributes** in the methods of the family TDIDT. The treatment of those fault values varies from an algorithm to another ID3 and C4.5.
- **Generated logical models:** the programs only build classifiers that can be expressed as decision trees or as a group of production rules.

6.1.3. Generated results. Characteristic of the decision trees

The decision trees represent a structure of data that organizes efficiently to the describers. A tree in a such way is built that it is achieved a test about the value of the describers in each node and in accordance with the answer leaves descending in the branches, until arriving at the end of the path where it is the value of the classifier. One can analyze a decision tree like a black box in function of whose parameters (describers) a certain value of the classifier is obtained. A decision tree can be analyzed like a disjunction of conjunctions. Each path from the root until the leaves represents a conjunction, and the entire path is alternative, that is to say, they are disjunctions.

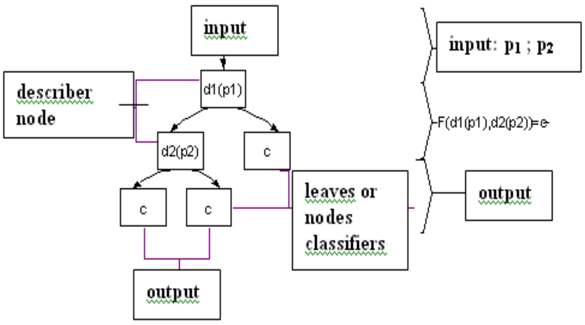

Figure 6.1.3.1.: Structure of a decision tree

6.1.4. Characteristic of the decision rules

The rules of decision or of production they are an alternative to the decision trees, and all decision trees can be taken to rules of this type [52] [3].

Antecedent ⇒ Consequent

Where the antecedent is a conjunction among different tests of value on the values of the attributes; and the consequent one is a class for all the cases that satisfy to the antecedent.

For example: **If attribute$_1$=”value a” and attribute$_2$= “value y”, then Clase$_K$**

The rules of decision are presented in order, and they should be interpreted in that way. The order determines which rule they should be executed first. When classifying a new case you advances in the list until arriving to an antecedent that is satisfied by the case, then the class of the case it is the corresponding to the consequent of this rule. The C4.5 in particular, to add a last rule to the list, this it doesn't has antecedent, it is the rule with the default class, that is to say, if the case didn't satisfy none of the previous rules, then it is of the class indicated by the last rule that it doesn't has antecedent.

In the case of the rules of decision, to add a new rule simply implies to add it to the list of rules without necessity of making structure changes, while to add a new rule in a tree would imply to redo the structure of the same one.

6.1.5. Presentation of the results

As much ID3 as C4.5 generate a classifier in the way of a decision tree whose structure is [43]:

- A leaf, indicating a class, or
- A node of decision that specifies some test to be carried out on an only attribute, with a branch and subtree for each value possible of the test.

The decision tree generated by the C4.5 has several particular characteristics: each leaf has associates two numbers that indicate the number of cases of covered trainings for each leaf and the quantity of them classified erroneously by the leaf. It is in certain way, an estimator of the success of the tree on the cases of training. The ID3, on the other hand, doesn't classify erroneously to the data of training, with that which they are not necessary this type of indicators. It is in and of itself that this algorithm, contrary to the C4.5, the risk runs of falling in overfitting, precise indicators.

Michie criticizes the ID3 when sustaining that the results demonstrate that the algorithms like the ID3 are considered "súper-programs" but incomprehensible for people [18] [19].

It should be simplified to the decision trees and the trees generated by C4.5 and by ID3 they become a group of production rules or of decision, more comprehensible than the trees, when these last ones are too extensive or leafy (bushy).

6.2. General description of the algorithms

The main algorithm of the systems of the family TDIDT, to which the ID3 and the C4.5 its descendant belong, is the process of generation of an initial decision tree starting from a data set of training. The original idea is based on a work of Hoveland and Hunt 50´s years, culminated in the book Experiments in Induction [50] [51] that describes several experiments with several implementations of systems of learning of concepts (concept learning systems - CLS).

6.2.1. Division of the data

The method "divide and reign" is carries out in each step a partition of the data of the node according to a realized test on the "better" attribute. Any test that not divides to T in a way trivial, such that at least two different subsets {Ti} they are not empty, possibly it will be in a partition of subsets of an only class, still when most of the subsets contain a single example. However, the process of construction of the tree doesn't point merely to find any partition of this type, but to find a tree that it reveals a structure of the domain and, therefore, have to be able to predict. For it, it is necessary an important number of cases in each leaf or, said otherwise, the partition should have the smallest quantity in possible classes. In the ideal case, it suits to choose in each step the test that generates the smallest tree. Then, a compact decision tree is looking for that is consistent with the data of training. All the possible trees could be explored and to choose the simplest. Unfortunately, an exponential number of trees should be analyzed. The problem of finding the consistent smaller decision tree with a group of training is of NP-complete complexity.

Most of the methods of construction of decision trees, including the C4.5 and the ID3, they don't allow to return to previous states, that is to say, they are sweet-toothed algorithms without towards back return. Once a test has been chosen for partition the current group, typically being based on the maximization of some local measure of progress, the partition is summed up and the consequences of an alternative election are not explored. For this reason, the election should be much realized.

6.2.2. Election of the division criterion

To carry out the division of the data in each step, the methods of the Information Theory are used. In a beginning, the ID3 used the gain like division criterion. However, the tests showed that this criterion was not effective in all the cases and better results were obtained if the ceriterion was normalized in each step. Therefore, the proportion of gain of information is used, with more success. The C4.5 also uses the criterion of proportion of gain that is more robust and it is more consistent than the gain criterion [37].

The proposed solution allows the use of both criterions. They should be studied and to compare the obtained results with the ID3 and with the C4.5 using the gain and the proportion of gain.

6.2.3. Criterion of gain

The information of gain definition can be calculated as the decrease in entropy. That is to say:

$$I(S,at) = H(S) - H(S,at) \tag{6.2.3.1.}$$

Let us suppose that we have a possible test with n that partition the T combined of the training in the subsets T_1, T_2,. . ., T_n. If the test is carried out without exploring the subsequent divisions of the subsets Ti, the only available information to evaluate the partition is the distribution of classes in T and its subsets.

$$H(T,X) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times H(T_i) \tag{6.2.3.2.}$$

Let us consider a similar measure after T has been partitioned according to the n of the test X. The prospective information (entropy) it can be determined as the pondered sum of the subsets, in the following way

$$I(T,X) = H(T) - H(T,X) \tag{6.2.3.3.}$$

The quantity measures the information gained when partition T according to the test X. The gain criterion, then, it selects the test that maximizes the gain of information. That is to say, before partition the data in each node, the gain is calculated that would be of partition the data set according to each one of the possible attributes. It is carried out the partition that is in the major gain.

6.2.4. Criterion of proportion of gain

The gain criterion has a very serious defect: it presents a very strong tendency to favor the tests with many results. Let us analyze a test on an attribute that is the primary key of a data set, in which, we will obtain an only subset for each case, and for each subset we will have $I(T,X) = 0$, then the gain of information will be maximum. From the point of view of the prediction, this division type is not useful. This inherent tendency to the gain criterion can be corrected by means of a normalization, in which the apparent gain is adjusted, attributable to tests with many results. Let us consider the content of information of a message corresponding to the results of the tests. For analogy to the definition of the $I(S)$ we have:

$$I\_division(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right) \qquad 6.2.4.1.$$

This represents the potential information generated when dividing T in n subsets, while the gain of information measures the outstanding information to a classification that it is born of the same division. Then,

$$proportion\_of\_gain(X) = \frac{I(T,X)}{I\_division(X)} \qquad 6.2.4.2.$$

expressed the useful proportion of information generated in the partition. If the partition is almost trivial, the information of the division will be small and this proportion will become unstable. To avoid this phenomenon, the criterion of proportion of gain selects a test that maximizes the previous expression, subject to the restriction that the information of the division is great, at least as great as the gain average on all the realized tests.

6.3. ID3

The algorithm ID3 was designed for J. Ross Quinlan [40] [41]. The ID3 takes objects of a well-known class and it describes them in terms of a fixed collection of property or of attributes, and it produces a decision tree on these attributes that it classifies all the objects correctly [41], there are certain qualities that differ to this algorithm of other general systems of inference. The first one is based on the form in that the required effort to carry out an induction task grows with the difficulty of the task. The ID3 was designed specifically to work with masses of objects, and the required time to process the data it only grows lineally with the difficulty, as product of:

- the quantity of objects presented as examples,
- the quantity of attributes given to describe these objects, and
- the complexity of the concept to be developed (measured by the quantity of nodes in the decision tree)

This lineal function is gotten at cost of the descriptive power: the concepts developed by the ID3 only take the form of decision trees based on the given attributes, and this "language" it is much more restrictive that the logic of first order or the logical multivalued, in which other systems express their concepts [41].

The ID3 is a much simpler mechanism for the discovery of a collection of objects belonging to two or more classes. Each object should be described in terms of a fixed group of attributes, each one of which it has its group of possible values of attributes. For example, the attribute humidity can have the values {high, low}, and the attribute climate, {sunny, cloudy, rainy}.

A classification rule in the form of a decision tree can be built for any combined C of attributes in that way [41]. If C is empty, then it associates it arbitrarily to anyone of the classes. If not, C contains the representatives of several classes; an attribute is selected and is partition C in combined disjoints C_1, C_2,..., C_n, where C_i it contains those members of C that have the value i for the selected attribute. Each one of these subsets is managed with the same strategy. The result is a tree in which each leaf contains a class name and each interior node specifies an attribute to be tested with a branch corresponding to the value of the attribute.

6.3.1. Description of the ID3

The objective of the ID3 is to create an efficient description of a data set by means of the use of a decision tree. Given consistent data, that is to say, without contradiction among them, the resulting tree will describe the input set to the perfection. Also, the tree can be used to predict the values of new data, assuming whenever the data set on which one works is representative of the whole data.

Given:

- A data set.
- A describers set of each date.
- A classifier/classifiers set for each object.

It is wanted to obtain a simple decision tree being based on the entropy, where the nodes can be:

1. Intermediate nodes: where are the chosen describers according to the entropy criterion that it determines which branch it is the one that should take.
2. Leaves: these nodes determine the value of the classifier.

This procedure of formation of rules will always work since two objects belonging to different classes don't exist but with identical value for each one of their attributes; if that case was presented, the attributes are inadequate for the classification process.

There are two important concepts to keep in mind in the ID3 algorithm [3]: the entropy and the decision tree. The entropy is used to find the most significant parameter in the characterization of a classifier. The decision tree is an efficient mean and intuitive one to organize the describers that can be used with functions predictives.

6.3.2. ID3 Algorithm

Next the algorithm of the method ID3 is presented for the construction of decision trees in function of a previously classified data set.

```
Function ID3
(R: set of attributes not classifiers,
C: attribute classifier,
S: traning set) it returns a decision tree;
Begin
If S is empty,
to return an only node with Value Fails;
If all the records of S have the same value for the attribute classifier,
To return an only node with this value;
If R is empty, then
to return an only node with the most frequent value in the attribute classifier in the records of S [Note: there will be errors, that is to say, records that won't be very classified in this case];
If R is not empty, then
D ← attribute with major Gain(D,S) among the attributes of R;
Be {dj| j=1,2,..., m} the values of the attribute D;
Be {Sj| j=1,2, .., m} the subsets of S corresponding to the values of
dj respectively;
To return a tree with the named root as D and with the named arcs
d1, d2, .., dm that they go respectively to the trees
ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), .., ID3(R-{D}, C, Sm);
End
```

6.3.3. Prunnig of the decision trees

The pruning of the decision trees is carried out with the objective that these are more comprehensible. That which implies that they have less level and/or be less leafy. The pruning applied in the ID3 is carried out once the tree has been generated and it is a quite simple mechanism: if of a node many branches are born, which finish all in the same class, then this node is replaced by a leaf with the common class. Otherwise, all the nodes children are analyzed.

6.3.4. Passage to rules of decision

To pass to rules of decision, the ID3 traverse the tree from the root until the leaves and it generates a rule for each traversed path. The antecedent of each rule will be compound for the conjunction of the tests of value of each visited node, and the class will be the corresponding to the leaf. The traverse of the tree is based on the traversed of having preorder (of root to leaves, of left to right). When being working with n-ary trees, this path is only.

6.3.5. Unknown attributes

It is necessary that all the cases presented to the ID3 are characterised for the same attributes. This limits the application of the algorithm, since not always it is had the whole necessary information. Let us imagine a historical database in which were adding attributes as it considered necessary, for the first cases of the same one they won't be known the values of the new attributes. Can the ID3 work with unknown attributes, does it consider them as if was a new value, in and of itself, you arrives to the convention that the unknown values, should they be expressed with a "?" in the data. The "?" it constitutes a new possible value for the attribute in question [38].

Once calculated the gain and proportions of gain for all the available attributes, the attribute should be chosen according to which is divided to this data set. So much in the case of the gain like in that of proportion gain, the best attribute for the division is that which maximizes it. In this example, the division according to the attribute STATE is the one that bigger gain and proportion of gain offers. This means that the node root of the tree will be a node that evaluates the attribute STATE [39].

6.3.6. Transformation to decision rules

As it was explained in the section 6.3.4 to pass a decision tree to rules of decision, the ID3 traversal preorder and every time that it arrives to a leaf, it writes the rule that has as consequent the value of the same one, and as antecedent, the conjunction of the tests of value specified in all the nodes traversed from the root to arrive to this leaf. Let us analyze the passage from the tree to rules of decision.

A limitation of the ID3 is that it can be applied to any data set, provided the attributes are discrete. This system doesn't have the easiness of working since with continuous attributes it analyzes the entropy on each one of the values of an attribute, therefore, it would take each value of a continuous attribute individually in the calculation of the entropy, that which is not useful in many of the domains. When one works with continuous attributes it is generally thought of ranges of values and not in particular values.

Several ways exist of solving this problem of the ID3, as the grouping of values presented in [10] or make discrete of the same ones explained in [3], [43]. The C4.5 solved the problem of the continuous attributes by means of make discrete.

The process described for the construction of decision trees assumes that the calculation operations, especially, those of evaluation of the relative frequencies (in those that should be elements) of the set C, they can be carried out efficiently, that which means, in the practice that so that the process is quick, C should reside in memory. The solution applied by ID3, for instances in the memory, is a solution iterative to create successive decision trees of precision more and more bigger, until arriving to the optimum decision tree [42]. The method can be summarized as [41]:

```
To choose a random group of instances (called window).
{repeat:
form a rule to explain the current window
find the exceptions to the rule in the rest of the instances
create a new window starting from the current window and the exceptions to the rules generated starting from it.
until they are not exceptions to the rule.}
The process finishes when it is formed a rule that doesn't have exceptions and be correct for all C.
```

6.4. C4.5

The C4.5 is based on the ID3, therefore, the main structure of both methods is the same one, it builds a decision tree by means of the algorithm "divide and reign" and it evaluates the information in each case using the entropy approaches and gain or gain proportion, as it is the case.

6.4.1. C4.5 algorithm

El algorithm C4.5 is similar to ID3, vary the way of realize the test on the attributes.

```
Function C4.5
(R: set of attributes not classifiers,
 C: attribute classifier,
 S: traning set) it returns a decision tree;
Begin
If S is empty,
   to return an only node with Value Fails;
If all the records of S have the same value for the attribute classifier,
   To return an only node with this value;
If R is empty, then
to return an only node with the most frequent value in the attribute classifier in the records of S [Note: there will be errors, that is to say, records that won't be very classified in this case];
If R is not empty, then
D ← attribute with major Proportion of Gain (D,S) among the attributes of R;
Be {dj| j=1,2,..., m} the values of the attribute D;
Be {Sj| j=1,2, .., m} the subsets of S corresponding to the values of
dj respectively;
To return a tree with the named root as D and with the named arcs
d1, d2, .., dm that they go respectively to the trees
C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2),..., C4.5(R-{D}, C, Sm);
End
```

6.4.2. Characteristic peculiar of the C4.5 test used

In each node, the system should decide which test it chooses to divide the data. The three types of tests possible proposals for the C4.5 are [43]:

1. the test "standard" for the discrete attributes, with a result and a branch for each value possible of the attribute.
2. a more complex test, based on a discrete attribute where the possible values are assigned to a variable number of groups with a possible result for each group, instead of for each value.
3. if an attribute A it has continuous numeric values, it is carried out a binary test with results A ≤ Z and A > Z, for that which is the limiting value of Z, should be determined.

All these tests are evaluated in the same way, looking at the result of the gain proportion, or alternatingly, that of the gain, resultant of the division that they take place. It has been useful to add an additional restriction: for any division, at least two of the subsets Ti should contain a reasonable number of cases. This restriction that avoids the almost trivial subdivisions is only kept in mind when the set T is small.

6.4.3. Test on continuous attributes

The tests for continuous values work with an arbitrary limiting value. The utilized method for the C4.5 is very simple [43] [46]. First, the cases of training T are ordered according to the values of the attribute A continuous that is being considered. A finite number of these values exists.

Be {v$_1$, v_2,. . ., v_m} the values that takes the attribute A. Any limiting value among vi saw and v_i *and* v_{i+1} will have the same effect when dividing the cases among those whose value stops in A it belongs to the subset {v$_1$, v_2,. . ., v_i} and those whose value belongs to {v$_{i+1}$, v_{i+2},. . ., v_m}. Then, they only exist m - 1 divisions possible of according to the value of A and all are examined. When being orderly, the successive tests for all the values, they can be carried out in an only passing.

Typically it is chosen the half point of the interval like representative limiting value, then the i-th limiting value would be:

$$\frac{v_i + v_{i+1}}{2} \qquad 6.4.2.1.$$

C4.5 differs of other algorithms in that it chooses the biggest value of A in the whole group of cases of training that doesn't exceed the half-point presented, instead of the half-point in itself, as limiting value; this way makes sure that all the limiting values that appear in the tree and/or the rules occur at least once in the data.

The method of binarization of attributes utilized has a great disadvantage. While all the operations of construction of a decision tree grow lineally with the number of cases of training, the classification of d continuous values grows in proportion of ***d x log(d)***. Then, the required time to build a tree starting from a great data set of training, it can be dominated by the classification of data with continuous values.

6.4.4. Unknown attributes

C4.5 assumes that all the ignored results of tests are distributed in form probabilistic according to the relative frequency of the well-known values. A case (possibly fractional) with an unknown value it is divided in fragments whose weight are proportional to this relative frequencies, giving for result that a case can follow multiple path in the tree. This is applied so much when the cases of training are divided during the construction of the tree, as when the tree is used to classify cases [38] [39].

6.4.5. Evaluation of the test

The modification of the gain criterion is quite direct. The gain of a test measures the information about the ownership to a class that can be expected as a result of partition a data set of training, calculated when subtracting the information that is expected that it is necessary to identify the class of an object after the partition to the same quantity before the partition. It is evident that a test cannot provide information of ownership to a class if one doesn't know the value of an attribute.

Be T the data set of training and X a test based on an attribute A, let us suppose that the value of A it is only known in a fraction F of cases in T.

Be I(T) and IX(T) calculated according to, if the attribute ***at*** divides the set S in the subsets S_i, *i* = 1, 2, ..., n, then, the total entropy of the system of subsets will be:

$$H(S,at) = \sum_{i=1}^{n} P(S_i)\cdot H(S_i) \qquad 6.4.5.1.$$

except that are only had in bill the cases for those which the value of A it is known. The gain definition can be corrected to:

$$\begin{aligned} &Gain\,(X) = probabilit\,y\_A\_be\_known \times (I(T) - I_X(T)) \\ &+ probabilit\,y\_A\_be\_unkown \times 0 = \\ &F \times (I(T) - I_X(T)) \end{aligned} \qquad 6.4.5.2.$$

or, in other words, the apparent gain of looking to the cases with well-known values, multiplied by the fraction of this cases in the group of training.

The calculation of the gain proportion is carried out in the same way. The definition of information of the division can modify in a similar way, considering the cases with unknown values as another group, then, if a test has n results, its information of the division is calculated as the test divided n+1subset.

$$I\_division\;(X) = -\sum_{i=1}^{n+1} \frac{|T_i|}{|T|+1} \times \log_2\left(\frac{|T_i|}{|T|+1}\right) \qquad 6.4.5.3.$$

6.4.6. Partition of the group of training

A test can select of the group of possible tests, as before, but using the modified versions of gain and information of the division. If the test X with results O_1, O_2, ..., O_N is chosen and has some values ignored for some of the data of training, the partition concept should be generalized, according to an approach probabilistic.

When a case T with a well-known result Oi it is assigned to the subset T_i, this means that the probability that the case belongs to T_i it is 1 and that it belongs to all the other subsets it is 0. When the result is ignored, it can only be carried out a weaker statistical affirmation. Then, associates with each case of the subset T_i a weight representing the probability that the case belongs to each subset. If the result for the case is known, then the weight is 1; if the case has an ignored result, then the weight is the probability of the result simply is Oi in this point. Each subset T_i are a collection of cases you fraction them possible, such that $|T_i|$ it should be reinterpreted as the sum of the weight that fraction them of the cases belonging to the subset.

The cases of training in T can have non unitary weight, since T can be the result of a previous partition. Then, in general, a case of T with weight p whose result is unknown, it is assigned to each subset T_i with weight:

$$P \text{ x probability\_of\_result\_}O_i \qquad 6.4.6.1.$$

The probability_of_result_O_i is considered as the sum of the weight of the cases in T with well-known values that have result Oi, on the sum of the pesos of the cases in T with well-known result for the test.

6.4.7. Classification of a new case

It takes a similar focus when the decision tree is used to classify a case. If in a node of decision the outstanding attribute is unknown, in such way that the result of the test cannot be determined, the system explores all the possible results and it combines the resulting classifications arithmetically. As for each attribute multiple path can exist from the root of the tree until the leaves, a "classification" it is a distribution of classes more than an only class. When the total distribution of classes for a new case has been established this way, the class with the highest probability, it is assigned as "the" class anticipated.

The information of the division is still determined starting from the group of complete training and it is bigger, since an extra category exists for the unknown values.

Each leaf in the resulting decision tree has associates two values: (N/E). N is the sum of the cases you fraction them that they arrive to the leaf; and E is the number of covered cases for the leaf that don't belong to the class of the same one.

6.4.8. Pruning of the decision trees

The recursive method of partition to build the decision trees described previously, will subdivide the group of training until the partition contains cases of an only class, or until the test doesn't offer improvement some. This gives as a result, generally, a very complex tree that overfitting the data when inferring a bigger structure that the required one for the cases of training [20] [44] [45]. Also, the initial tree is generally extremely complex and has a superior proportion of errors to that of a simpler tree. While the increase in complexity is understood at first sight, the biggest proportion of errors can be more difficult of visualizing.

To understand this problem, let us suppose that we have a data set two classes, where p 〈 0.5 a proportion of the cases belong to the majority class. If a classifier assigns all the cases with uncertain values to the majority class, the prospective proportion of error is clearly 1 - p. If, on the other hand, the classifier assigns a case to the majority class with probability p and to the other class with probability 1 - p, its prospective proportion of error is the sum of:

- the probability that a case belonging to the majority class is assigned to the other class, p x (1 - p),
- the probability that a case belonging to the other class is assigned to the majority class, (1 - p) x p that it gives 2 x as a result p (1 - p). As p it is at least 0.5, this is generally superior at 1 - p, then the second classifier will have a bigger proportion of errors. A complex decision tree has a great similarity with this second classifier type. The cases are not related to a class, then, the tree sends each case at random to some of the leaves.

A decision tree is not simplified erasing the whole tree in favor of a branch, but rather the parts of the tree are eliminated that don't contribute to the accuracy of the classification for the new cases, producing a less complex tree, and therefore, more comprehensible.

They exist, basically, two ways to modify the method of recursive partition to produce simpler trees: to decide not to divide more a group of cases of training or to remove some part of the structure built by the recursive partition retrospectively.

The first focus, well-known as pre-pruning, has the advantage that it doesn't get lost time in building a structure that then will be simplified in the final tree. The systems that apply it, generally look for the best way to leave the subset and they evaluate the partition from the statistical point of view by means of the theory of the gain of information, reduction of errors, etc. If this evaluation is smaller than a predetermined limit, the division is discarded and the tree for the subset is simply the most appropriate leaf. However, this type of method has the disadvantage that it is not easy to stop a partition in the appropriate moment, a very high limit can finish with the partition before the benefits of subsequent partitions seem evident, while a too low limit is in a too light simplification.

The C4.5 uses the second focus, the method "divides and reign" it processes the data of training freely, and the tree produced overfitting is pruned later. The processes computational extras invested in the construction of parts of the tree that then will be pruned they can be substantial, but the cost doesn't overcome the benefits of exploring a bigger quantity of possible partitions. The growth and the pruning of the trees are more slow, but more reliable.

The pruning of the decision trees will take, without a doubt, to classify a bigger quantity of the cases of training erroneously. Therefore, the leaves of a pruned tree won't necessarily contain an only class but a distribution of classes, like it was explained previously. Associated to each leaf, there will be a distribution of classes specifying, for each class, the probability that a case of training in the leaf belongs to this class.

Generally, the simplification of the decision trees is carried out discarding one or more subtrees and replacing them for leaves. The same as in the construction of trees, the classes associated with each leaf are when examining the covered cases of training for the leaf and choosing the most frequent case. Besides this method, the C4.5 allows to replace a subtree for some of its branches.

Let us suppose that it was possible to predict the proportion of errors of a tree and their subtrees. This immediately would take to the following pruning method: "To begin with the leaves and to examine each subtree. If a substitution of the subtree for a leaf or for their more frequently utilized branch, it takes to a proportion of errors anticipated (predicted error rate) smaller, then to prune the tree according to it, remembering that the proportions of errors anticipated for all the subtrees that contain it will be affected". As the proportion of errors anticipated for a tree diminishes if they diminish the proportions of errors anicipated in each one of their branches, this process would generate a tree with a proportion of errors minimum preventively.

Examples of this family are:

“It prunes according to the complexity of the cost (Cost-complexity pruning).”

“Pruning of reduction of errors (Reduced-error pruning) [36].”

6.4.9. Estimate the proportion of errors for the decision trees

Once pruned, the leaves of the decision trees generated by the C4.5 will have two associate numbers: N and E. N is the quantity of covered cases of training for the leaf, and E it is the quantity of errors anticipated if a group of N new cases was classified by the tree.

The sum of the errors anticipated in the leaves, divided the number of cases of training, is an immediate estimador of the error of a tree pruned on new cases.

7. INTEGRATION OF THE SYSTEM - CONCLUSIONS

To study the proposed algorithms a system it was developed that integrates the ID3 and the C4.5. The system receives the data of training like entrance and it allows the user to choose which algorithm and with what criterion of decision (gain or gain proportion) he wants to apply. Once generated the tree and the rules of decision, the user can evaluate the results on the supporting data. In the case of the ID3, this evaluation is carried out starting from the rules of decision whose performance, is identical to that of the trees. The evaluation of the results of the C4.5, on the other hand, is carried out for separate and they are obtained, therefore, two different evaluations, one for the tree and another for the rules. This is due to that, the classification model generated with the C4.5 like decision tree is different to the one generated as rules of decision.

REFERENCES

[1] Althusser, Louis. Pour Marx. Semantic theory. F. Maspero. Paris. France. 1965

[2] Blockeel, H., De Raedt, L. *Top-Down Induction of Logical Decision Trees*. Katholieke Universiteit Leuven, Departament of Computer Science, Celestijnelaan, Belgium. 1997.

[3] Blurock, Edward S. Blurock, *The ID3 Algorithm*, Research Institute for Symbolic Computation 1996.,www.risc.uni-linz.ac.at/people/bulrock/ANALYSIS/ manual/document, Austria.

[4] Bunge, Mario. The scientific investigation. Editorial Ariel. Barcelona. Spain. 1969.

[5] Bunge, Mario. Rationality and Realism. Editorial Alianza. Madrid. Spain. 1983.

[6] Bunge, Mario. Social science under debate. Editorial Sudamericana. Buenos Aires. Argentina. 1999.

[7] Duhem, Pierre, Gilbert Ryle and Albert Einstein thesis of the Empiric Science, Circle of Vienna. Austria. 1934.

[8] Feyerabend, Paul Karl. Treaty against the method. Tecnos. Madrid. Spain. 1981.

[9] Freud, Sigmund. Their complete Works. Editorial New Library. Madrid. Spain. 1967-1980.

[10] Gallion, R., St Clair, D., Sabharwal, C., Bond, W.E. Dynamic ID3: A Symbolic Learning Algorithm for Many-Valued Attribute Domains. Engineering Education Center, University of Missouri-Rolla, St. Luis, EE.UU. 1993.

[11] Gianella, Alicia E. Introduction to the epistemology and the methodology of the science. Editorial of the National University of La Plata. Buenos Aires. Argentina. 2000.

[12] Hempel, Carl Gustav. The scientific explanation. Studies on the philosophy of the Science. Editions Iberian Paidós, CORP. Barcelona. Spain. 1996.

[13] Kant, Immanuel. Critic of the pure reason. Editorial Losada. Buenos Aires. Argentina. 1973.

[14] Klimovsky, Gregorio. The misfortunes of the scientific knowledge. A-Z. Buenos Aires. Argentina. 1994.

[15] Kuhn, Thomas S. The structure of the scientific revolutions. Editorial of the Fund of the Economic Culture. Mexico. 1980.

[16] Lakatos, Imre. Sophisticated falsificationism. Rodolfo Gaeta y Susana Lucero. Editorial Universitaria de Buenos Aires. Argentina. 1999.

[17] Malinowski, Bronislaw. The argonauts of the western Pacific. Editorial Planeta. Barcelona. Spain. 1986.

[18] Michie, D. On Machine Intelligence (2nd ed.), Ellis Horwood, Chichester, Reino Unido. 1986.

[19] Michie, D. Machine Learning in the next five years, EWSL-88, 3rd European Working Session on Leaming, Pitman, Glasgow, Londres, Reino Unido. 1988.

[20] Mitchell, T. *Machine Learning*. MCB/McGraw-Hill, Carnegie Mellon University, EE.UU. 1997.

[21] Mitchell, T. *Decision Trees*. Cornell University, www.cs.cornell.edu/courses/ c5478/ 2000SP, EE.UU. 2000a.

[22] Mitchell, T. Decision Trees 2. Cornell University, www.cs.cornell.edu/ courses/c5478/2000SP, EE.UU. 2000b.

[23] Nagel, Ernest. La estructura de la ciencia. Paidós. Buenos Aires. Argentina. 1968.

[24] Parsons, Talcott. The structure of the social action (1937), The social system and Societies: perspectives evolucionistas and comparative (1966). Harvard University Press. USA.

[25] Perichinsky, Gregorio. Diagnostic Interdisciplinary National Congress and Professional and Scientific Perspectives toward the XXI Century. With Academic auspice of the National University of La Plata. President of the Commission of Science and Technology. Writer of conclusions. Exponent in the plenary. Buenos Aires. Argentina. 1991.

[26] Perichinsky, Gregorio. Investigation, Education and Projection in Computer Science. Proceedings of the First International Congress of Computer Engineering. Pages 306-314. Faculty of Engineering. University of Buenos Aires. Argentina. 1995.

[27] Piaget, Jean y Apostel, L. Construction and validation of the scientific theories.Paidós. Buenos Aires. Argentina. 1978.

[28] Poincaré, Henri. Science and method. Academic Scientist from France. Paris. 1908.

[29] Popper, Karl. The logic of Scientific Discovery. Science Editions. New Cork. USA. 1961.

[30] Popper, Karl. Conjetures and Refutations. Routledge and Kegan. London. U.K. 1963.

[31] Popper, Karl. Conocimiento objetivo. Tecnos. Madrid. España. 1974.

[32] Popper, Karl. Realism and the objective of the science. Tecnos. Madrid. Spain. 1985. Empirism of the Circle of Vienna. Austria, (Rudolf Carnap) - Philosophical Society of Berlin, Germany (Ernest Nagel, Hans Reichenbach and Carl Gustav Hempel). 1934.

[33] Quine, William van Orman. The methods of the logic. Ariel. Barcelona. Spain. 1975.

[34] Quine, William van Orman. Theories and things. Cambridge. Harvard University Press. USA. 1981.

[35] Quinlan, J. “Introduction of Decision Trees”. Machine Learning. Vol.1. N° 1. Pp. 81-106. 1986.

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

[37] Quinlan, J.R. *Decision trees and multi-valued attributes*. En J.E. Hayes, D. Michie, and J. Richards (eds.), Machine Intelligence, Volumen II, páginas. 305-318.Oxford University Press, Oxford, Reino Unido. 1988.

[38] Quinlan, J.R. *Unknown Attribute Values in Induction*. Basser Departament of Computer Science, University of Science, Australia. 1989.

[39] Quinlan, J. R. *Learning Logic Definitions from Relations*. En Machine Leaming, Vol 5, páginas 239-266. Oxford University Press, Oxford, Reino Unido. 1990.

[40] Quinlan, J.R. *The Effect of Noise on Concept Learning*, En R. S. Michalski, J. G. Carbonell, & T. M. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann, Vol. I, Capítulo 6, páginas149-167. San Mateo, CA: Morgan Kaufmann, EE.UU. 1993a.

[41] Quinlan, J.R. *Learning Efficient Classification Procedures and Their Application to Chess Games*, En R. S. Michalski, J. G. Carbonell, & T. M. Mitchells (Eds.) Machine Learning, The Artificial Intelligence Approach. Morgan Kaufmann,Vol. II, Capítulo 15, páginas 463-482, EE.UU. 1993b.

[42] Quinlan, J.R. *Combining Instance-Based and Model-Based Learning*. Basser Departament of Computer Science, University of Science, Australia. 1993c.

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

[44] Quinlan, J.R. *MDL and Categorical Theories*. Basser Departament of Computer Science, University of Science, Australia. 1995a.

[45] Quinlan, J.R., Cameron-Jones, R.M. *Oversearching and Layered Search in Empirical Learning*. Basser Departament of Computer Science, University of Science, Australia. 1995b.

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

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

[48] Ryle, Gilbert. See [7] The Thesis of Empiric Science with Pierre Duhem and Albert Einstein.

[49] Samaja, Juan. Epistemology and Methodology. Editorial Universitary of Buenos Aires. Argentina. 1993.

[50] Hunt, E.B., Marin, J., Stone, P.J. Experiments in Induction. New York: Academic Press, EE.UU. 1966.

[51] Hunt, E.B. Artificial Intelligence. New York: Academic Press, EE.UU. 1975.

[52] Witten, I.H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Diego, EE.UU. 2000.

[53] Perichinsky Gregorio, Jiménez Rey Elizabeth Miriam, Grossi María Delia, Vallejos Félix Anibal, Servetto Arturo Carlos, Orellana Rosa Beatriz, Plastino Angel Luis. Taxonomic Evidence of Classification. Applying Intelligent Data Mining. Galactic and Globular clusters. Annals of the Faculty of Engineering Hunedoara - Journal of Engineering. Tome V. Fascicule 2. ISSN 1584 – 2665. University “Politechnica” Timisoara Faculty of Engineering – Hunedoara. 2007.