DENIS ENĂCHESCU

University of Bucharest

ELEMENTS OF STATISTICAL LEARNING. APPLICATIONS IN DATA MINING

Lecture Notes

CONTENTS

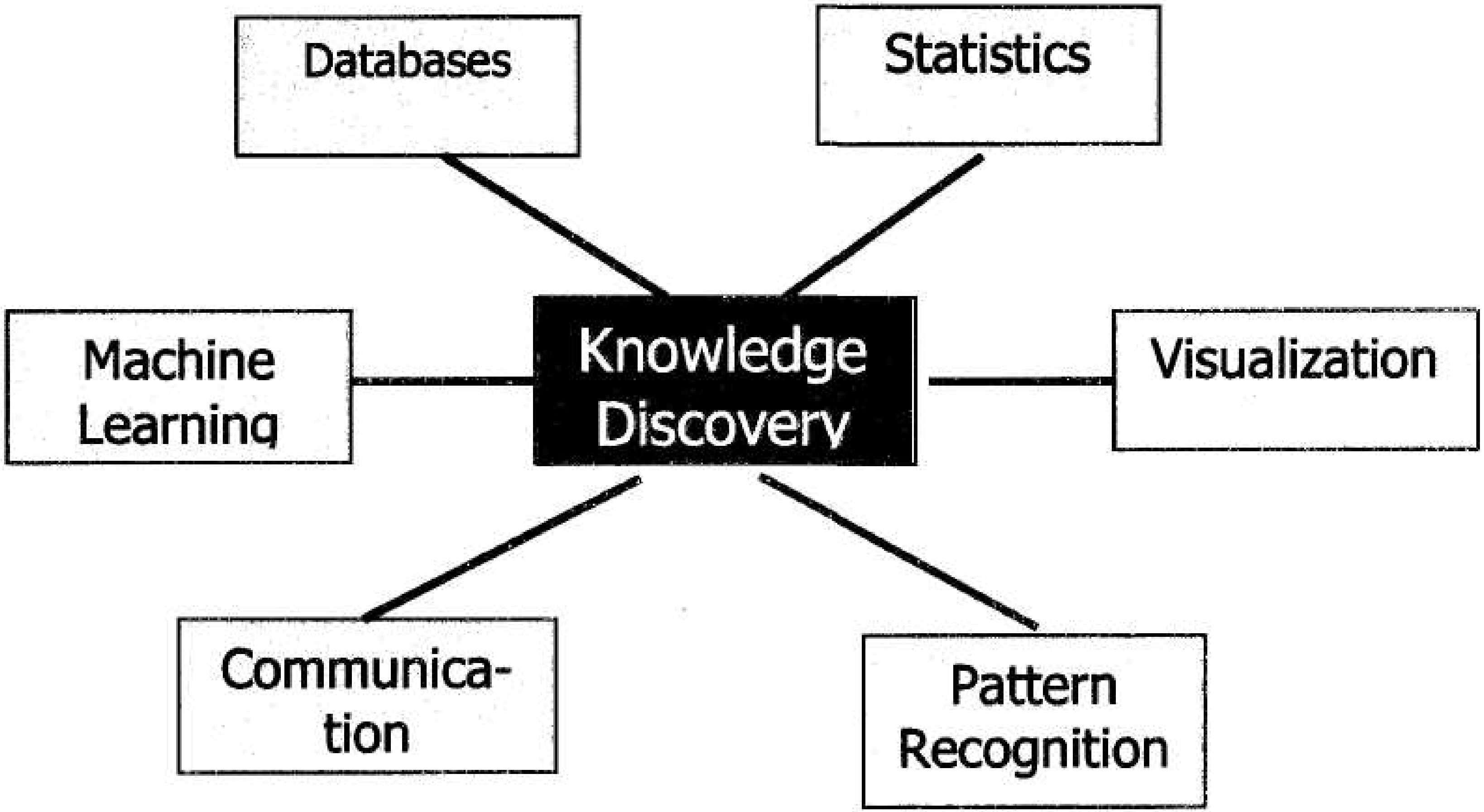


Figure 1-1 Relationships among Knowledge Discovery and other disciplines. (From Han[1]).

[1] http://www.es. sfu,ca/~han

1 The Nature of Machine Learning

1.1 Basic Definitions and Key Concepts

Learning **(**understood: **artificial, automatic)** *(Machine Learning)* This concept includes any method making it possible to build a model of reality starting from data, either by improving a partial or less general model, or by creating the model completely. There are two principal tendencies in learning, that resulting from the artificial intelligence and qualified *symbolic system*, and that resulting from the statistics and qualified *numerical*.

Precision vs**. Generalization,** the great dilemma of the learning. Precision is defined by a difference between a measured or predicted value and an actual value. To learn with too much precision leads to an "over-fitting", like the learning by heart, for which unimportant details (or induced by the noise) are learned. To learn with not enough precision leads to an "over-generalization" and the model applies even when the user does not wish it.

Intelligibility (should be *Comprehensibility* but tends to become *Understandability*). For a few years, mainly under the push of the industrialists, the researchers have started to try also to control the intelligibility of the model obtained by the mining of data. Until now, the methods of measurement of intelligibility are reduced to check that the results are expressed in the language of the user and that the size of the models is not excessive. Specific methods of visualization are also used.

The criterion of success. The criterion of success is what is measured in the performance evaluation. It thus acts as a *criterion relative to an external observer*. For example, the performance will be measured according to the error count made by the learner in the course of learning, or according to its error rate after learning. More generally, the measurement of performance can include factors independent of the adequacy to the data of learning and of very diverse natures. For example, *simplicity* of the learning result produced by the learning machine (LM), its *comprehensibility*, its *intelligibility* by an expert, its facility to the integration in a current theory, the low computational cost necessary to its obtaining, etc.

Here, it should be made an important remark. The criterion of success, measured by an external observer, is not necessarily identical to the *performance index* or to the *loss function* that is *intern to the LM* and used in the internal evaluation of the learning model. For example, an algorithm of learning of a connectionist network generally seeks to minimize a standard deviation between what it predicts on each example of learning and the desired exit.

The protocol of learning. The learning and its evaluation depend on the *protocol* that establishes the interactions between the LM and his environment, including the supervisor (the oracle). It is thus necessary to distinguish between the *batch learning*, in which all the data of learning are provided all at the start, and the *on-line learning* in which the data arrive in sequences and where the learner must deliberate and provide an answer after each entry or groups entries.

The protocol also stipulates the type of entries provided to be learned and the type of awaited exits. For example, a scenario can specify that at every moment the LM receive an observation $\mathbf{x}_i$, that it must provide an answer y_i and only then, the supervisor produces the correct answer

u_i. One speaks then naturally about a *prediction task*. More, the tasks known as prediction are interested to envisage correctly a response in a precise point.

In contrast, in the *identification tasks* the goal is to find a total explanation among all those possible, which once known will make possible to make predictions whatever the question.

The scenario will be then different. By example, the learning system must yet provide after each new entry $(\mathbf{x}_i, u_i)$ an assumption on the "hidden function" of the supervisor by which this one determines u_i as function of $\mathbf{x}_i$. It is conceived that the criterion of success is not the same in the case of a prediction task as in that of an identification task. In this last case, indeed, one asks much more from the LM since one awaits from him an explicit assumption, therefore a kind of explanation from his predictions.

In addition, the LM can be more or less active. In the protocols described up to now, the LM receives passively the data without having influence on their selection. It is possible to consider scenarios in which the LM has a certain initiative in the search for information. In certain cases, this initiative is limited, for example when the LM, without having the total control of the choice of the learning sample, is simply able to direct its probability distribution; the *boosting* methods are an illustration of this case. In other cases, the LM can put questions about the class of membership of an observation, one speaks then *of learning by membership querie*s, or to even organize experiments on the world, and one speaks then *of active learning*. The play of Mastermind, which consists in guessing a configuration of colors hidden pawns by raising questions according to certain rules, is a simple example of active learning in which the learner has the initiative of the questions.

The task of learning. It is possible to approach the objective of the process of learning following several points of view.

- The knowledge point of view.

The goal of the learning can be *to modify the contents* of knowledge[2]. One speaks then *of knowledge acquisition*, of *revision*, and, why not, of *lapse of memory*.

The goal of the learning can also be, without necessarily modifying the "contents" of knowledge, *to make it more effective* compared to a certain goal, by reorganization, optimization or compilation for example. It could be the case of a player of chess or a mental calculator who learns how to go more and more quickly without knowing new rules of play or of calculation. One speaks in this case about *optimization of performance* (*speed-up learning*).

[2] Measured, for example, by its deductive closure, i.e., in a logical representation, all that can be deduced correctly starting from the current base of knowledge.

- The environment point of view.

The task of the learning can also be defined compared to what the learning agent must carry out "to survive" in its environment. That can include:

- *To learn how to recognize patterns* (for example: handwritten characters, birds, the predatory ones, an ascending trend of a title to the bourse, appendicitis, etc.). When the learning is done with a professor, or supervisor, who provides the wished answers, on have a *supervised learning*. If not, one speaks *of unsupervised learning.* In this last case, the task of learning at the same time consists in discovering categories and finding rules of categorization.
- *To learn how to predict.* There is then a concept of temporal dependence or causality.
- *To learn how to be more effective.* It is the case in particular of the situations of resolution of problem, or search for action plans in the world.

- The abstract classes of problems point of view.

Independently from the learning algorithm, it is possible to characterize the learning process by a general and abstract class of problems and processes of resolution. Thus a certain number of disciplines, in particular resulting from mathematics or information theory, were discovered an interest for the problems of learning.

  - *The theories of compression of information.* In a certain direction, the learning can be approached like a problem of extraction and compression of information. It is a question of extracting essential information or the initial message from an ideal transmitter, cleared of all its redundancies. In a sense, the nature sciences, such astronomy or ornithology, proceed by elimination of the superfluous or redundant details and by the description of hidden regularities.

– *The cryptography*. From the similar point of view, near to the goals of the information theory, the learning can be regarded as one *try decoding* or even of decoding of a message coded by the ideal transmitter and intercepted in whole or part by the learner agent. After all, it is sometimes like the scientist studying nature. It is then logical to see under which conditions a message can "be broken", i.e. under which conditions learning is possible.

– *The mathematical / numerical analysis*. The learning can also be examined like one *problem of approximation*. The task of learning is to find an approximation as good as possible of a hidden function known only by the intermediary of a sample of data. The problem of learning becomes often that of the study of the conditions of approximation and convergence.

– *The induction*. In the Seventies and at the beginning of the Eighties, under the influence from the cognitive point of view, a broad community of researchers, particularly active in France, is leaning on the learning as *a problem of generalization.* This approach starts from two essential hypotheses. First, the cognitive learning agent must learn something that another cognitive agent equivalently knows. It is thus normally able to reach the target knowledge perfectly. Second, knowledge and data can be described by a language. One seeks then the operators in this language who can correspond to operations of generalization or specialization useful for induction, and one builds algorithms using them, making it possible to summarize the data while avoiding the over-fitting and the drawing of illegitimate consequences.

– *The applied mathematics.* Finally, the engineer can be tempted to see in the learning a particular case of the *resolution of an inverse problem*. Let us take two examples:

  ▪ one can say that the theory of probability is a theory sticking to a direct problem (being given a parameterized model, which are the probabilities associated with such event?), while the theory of the statistics attacks an inverse problem (being given a sample of data, which model does make it possible to explain it, i.e. can have produced it?).

- being given two numbers, it is easy to find the product of it (direct problem). It is on the other hand generally impossible to find starting from a number those of which it is the product (inverse problem).

The inverse problems are thus often problems that one known as *ill-posed,* i.e. not having a single solution. According to this point of view, the study of the learning can be seen like that of the conditions making possible to solve an ill-posed problem, i.e. constraints which it will have to be added so that the procedure of resolution can find a particular solution.

- The structures of data or types of concerned assumptions

It frequently arrives that one imposes the type of structure (or the *language of expression of the assumptions)* that must be sought by the learning system. That makes it possible to guide at the same time the determination of the learning algorithm to be used, but also the data that will be necessary so that the learning is possible. Without to seek to be exhaustive, we quote among the principal structures of studied data:

- *the Boolean Expressions*, who are often adapted to learn concepts definite on a language of attribute-values(for example the rules of an expert system).
- *the grammars and the Markovian Processes* allowing representing sequences of events.
- *the linear/nonlinear functions* making possible to discriminate objects belonging to a subspace or its complementary.
- *the decision trees* who allow the classifications by hierarchies of questions. The corresponding decision tree is often at the same time concise and comprehensible.

- *the logical programs* who allow learning from the relational concepts.

- *the Bayesian Networks* allowing at the same time to represent universes structured by relations of causality, to take into account, and to express measurements of certainty or confidence.

Sometimes the learning can consist in changing the *structure of data* in order to find an equivalent but most computational effective structure. It is once again, under another angle, the problem of performance optimization.

To simplify, we will suppose that the LM seek an approximation of the target function inside a family $\mathcal{H}$ of *hypothesis functions*. It is the case, for example, of the learning using a neurons network of which architecture constrained the type of realizable functions to a certain space of functions.

We defined the task of learning like that of a problem of estimating a function starting from the observation of a sample of data. We turn now to the principles allowing carrying out this estimate.

The exploration of the hypothesis space. Let $\mathcal{H}$ be a hypothesis space, $\mathcal{X}$ a data space and $\mathcal{S}$ a training sample. The task of learning is to find a hypothesis h approximating as well as possible, within the meaning of a certain measurement of performance, a target function f based on the sample $\mathcal{S}=\left\{\left(\mathbf{x}_i,u_i\right)\right\}_{i=\overline{1,m}}$ in which one supposes that each label u_i, was calculated by the function f applied to the data $\mathbf{x}_i$.

How to find such a hypothesis $h \in \mathcal{H}$? Two questions arise:

1. How to know that a satisfactory hypothesis (even optimal) was found, and more generally how to evaluate the quality of a hypothesis?

2. How to organize the research in $\mathcal{H}$?

Whatever the process guiding exploration of $\mathcal{H}$, it is necessary that the LM can *evaluate* the hypothesis h that it considers at each moment t of its research. We will see that this evaluation utilizes an intern performance index (for example a standard deviation between the exits calculated from h and desired targets u provided in the training sample). It is this performance index, more possibly, the other information provided by the environment (including the user for example), which allows the LM to measure its performance on the training sample and to decide if it must continue his research in $\mathcal{H}$ or it can stop.

By supposing that at the moment t, the LM judge unsatisfactory his current assumption h_t, how can it change it? It is there that the effectiveness of the learning is decided and in this context, the structure of the space $\mathcal{H}$ plays an important role. More this one will be rich and fine, more it will be possible to organize the effectively exploration of $\mathcal{H}$. Quickly let us examine three possibilities in an ascending order of structuring:

- the space $\mathcal{H}$ of hypothesis does not present *any structure.* In this case, only a random exploration is possible. Nothing makes it possible to guide research, nor even to benefit from the information already gained on $\mathcal{H}$. It is the case where nothing is known *a priori* on $\mathcal{H}$.

- *a concept of neighborhood is definable on* $\mathcal{H}$. It is then possible to operate an exploration by techniques of optimization like the gradient method. The advantage of these techniques, and what makes them so popular, it is that they are of a very general use since it is often possible to define a concept of neighborhood on a space. A fundamental problem is that of the relevance of this concept. A bad neighboring relation can indeed move away the LM from the promising areas of the space! In addition, it is still a low structure, which, except in particular cases (differentiability, convexity, etc. of the function to be optimized), does not allow a fast exploration.

- It is sometimes possible to have a stronger structure making it possible to organize the exploration of $\mathcal{H}$. In this case, for example, it becomes possible to modify an erroneous hypothesis by specializing it just enough so that it does not cover any more the new negative example, or on the contrary by generalizing it just enough so that it covers the new provided positive example. This type of exploration, possible in particular when the space of hypothesis is structured by a language, is generally better guided and more effective than a blind exploration.

By what precedes, it is obvious that more the structuring of the space of the hypothesis is strong and is adapted to the problem of learning, more the learning will be facilitated. On the other hand, of course, that will require a preliminary deliberation.

1.2 *Short History*

The artificial learning is a young discipline at the common frontier of the artificial intelligence and the computer science, but it has already a history. We brush it here rapidly, believing that it is always interesting to know the past of a discipline because it can reveal, by the updated tensions, its major problems and its major options.

The theoretical preliminary principles of the learning are posed with the first results in statistics in the years 1920 and 1930. These results seek to determine how to inhere a model starting from data, but especially how to validate an assumption based on a sample of data. Fisher in particular studies the properties of the linear models and how they can be derived starting from a sample of data. At the same period, the computer science born with the work of Gödel, Church and especially Turing in 1936, and the first simulated data become possible after the Second World War. Besides the theoretical reflections and the conceptual debates on the cybernetics and the cognitivism, the pioneers of the domain try to program machines to carry out intelligent tasks, often integrating learning. It is particularly the case of the first simulations of tortoises or cybernetic mice, which one places in labyrinths while hoping to see how they learn to leave it more and more quickly. On his side, Samuel at IBM, in the years 1959-1962, develops a program to play the American Backgammon, which includes an evaluation function of the positions enabling him to become quickly a very good player.

In the years 1960, the learning is marked by two currents. On the one hand, a first connectionism, which under the crook of Rosenblatt father of the perceptron, sees developing small artificial neurons networks tested in class recognition using supervised learning. On the other hand, the conceptual tools on pattern-recognition are developed.

At the end of 1960, publication of the book of Minsky and Papert (1969) which states the limits of the perceptron causes the stop for about fifteen years of almost all researches in this field. In a concomitant manner, the accent put in artificial intelligence in the years 1970, on knowledge, their representation and the use of sophisticated inference rules (period of the expert systems) encourages work on the learning systems based on structured knowledge representations bringing on the stage complex rules of inference like the generalization, the analogy, etc.

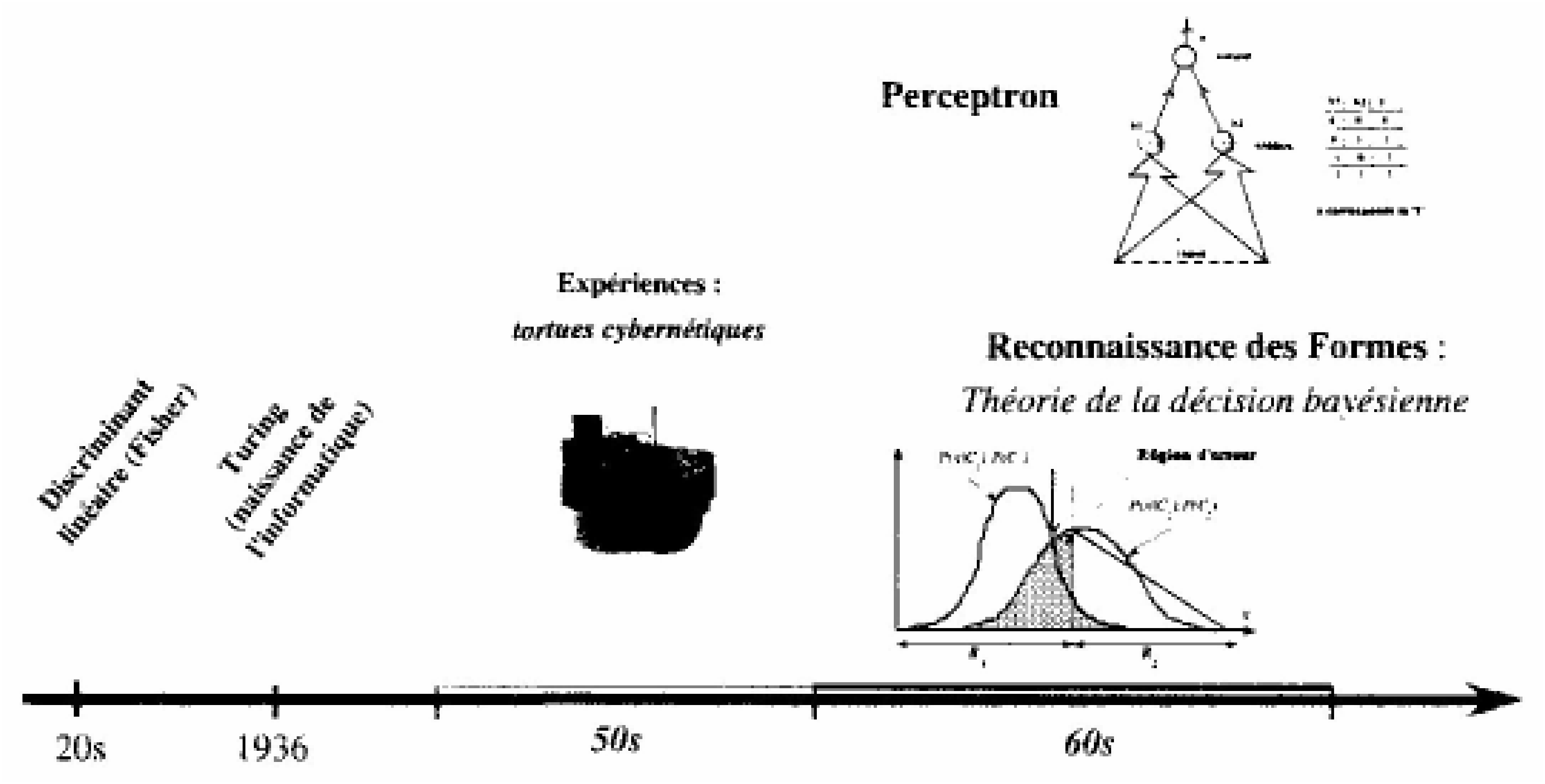

Perceptron
Expériences :
tortues cybernétiques
Reconnaissance des Formes :
Théorie de la décision bayésienne
Discriminant linéaire (Fisher)
Turing (naissance de l'informatique)
20s
1936
50s
60s

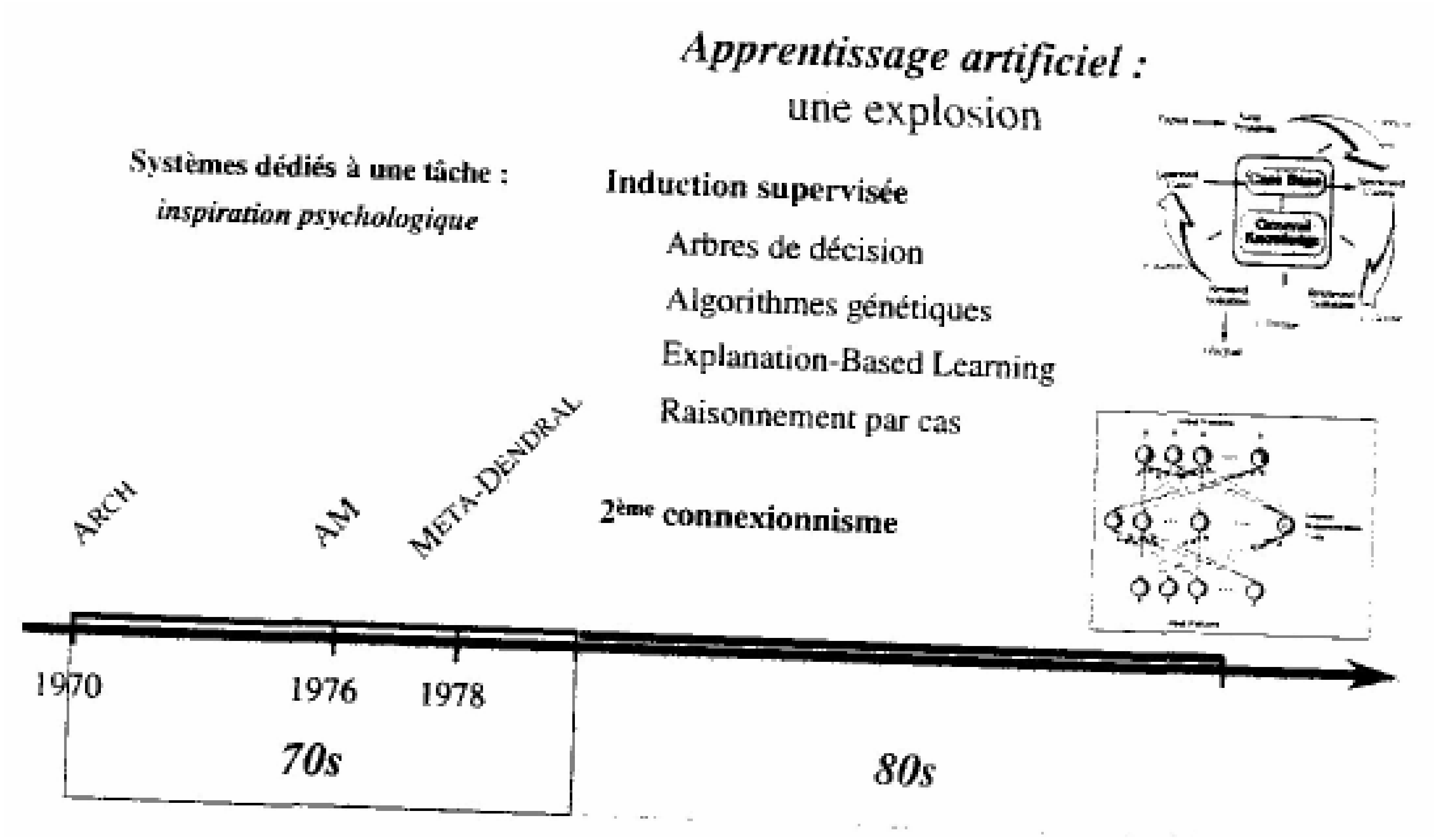
Figure 1-1 The first period of the artificial learning

It is then the triumph of impressive systems realizing the specific tasks of learning by simulating strategies used, more or less, in the human learning. It must be cited, the system ARCH of Winston in 1970, which learns how to recognize arches in a world of blocks starting from examples and counterexamples; the system AM of Lenat in 1976, which discovers conjectures in the field of arithmetic by the use of a set of heuristic rules or even the system META-DENDRAL of Mitchell which learns rules in an expert system dedicated to the identification of chemical molecules.

It is also a period during which the dialogue is easy and fertile between the psychologists and the experts of the artificial learning. From where assumptions relating concepts like the short-term and long-term memories, the procedural or declaratory type of knowledge, etc. also the ACT system of Andersen testing general assumptions on the learning of mathematical concepts in education.

However, also spectacular they are, these systems have weaknesses, which come from their complexity. Indeed their realization implies necessarily a great number of choices, small and large, often implicit, and who of this fact do not allow an easy replication of the experiments, and especially throw the doubt about the general and generic range of the proposed principles. It is why years 1980 saw gradually drying up work relating to such simulations with some brilliant exceptions like the systems ACT or SOAR.

Moreover, these years saw a very powerful come back of connectionism in 1985, with in particular the discovery of a new algorithm of learning by the gradient descent method for multi-layer perceptrons. That deeply modified the study of the artificial learning by opening large the door at all the concepts and mathematical techniques relating on optimization and the convergence properties. Parallel to the intrusion of continuous mathematics, other mathematicians engulfed themselves (behind Valiant in 1984) in the breach opened by the concept of space of versions due to Mitchell.



Apprentissage artificiel :
une explosion

Systèmes dédiés à une tâche :
inspiration psychologique

Induction supervisée

Arbres de décision

Algorithmes génétiques

Explanation-Based Learning

Raisonnement par cas

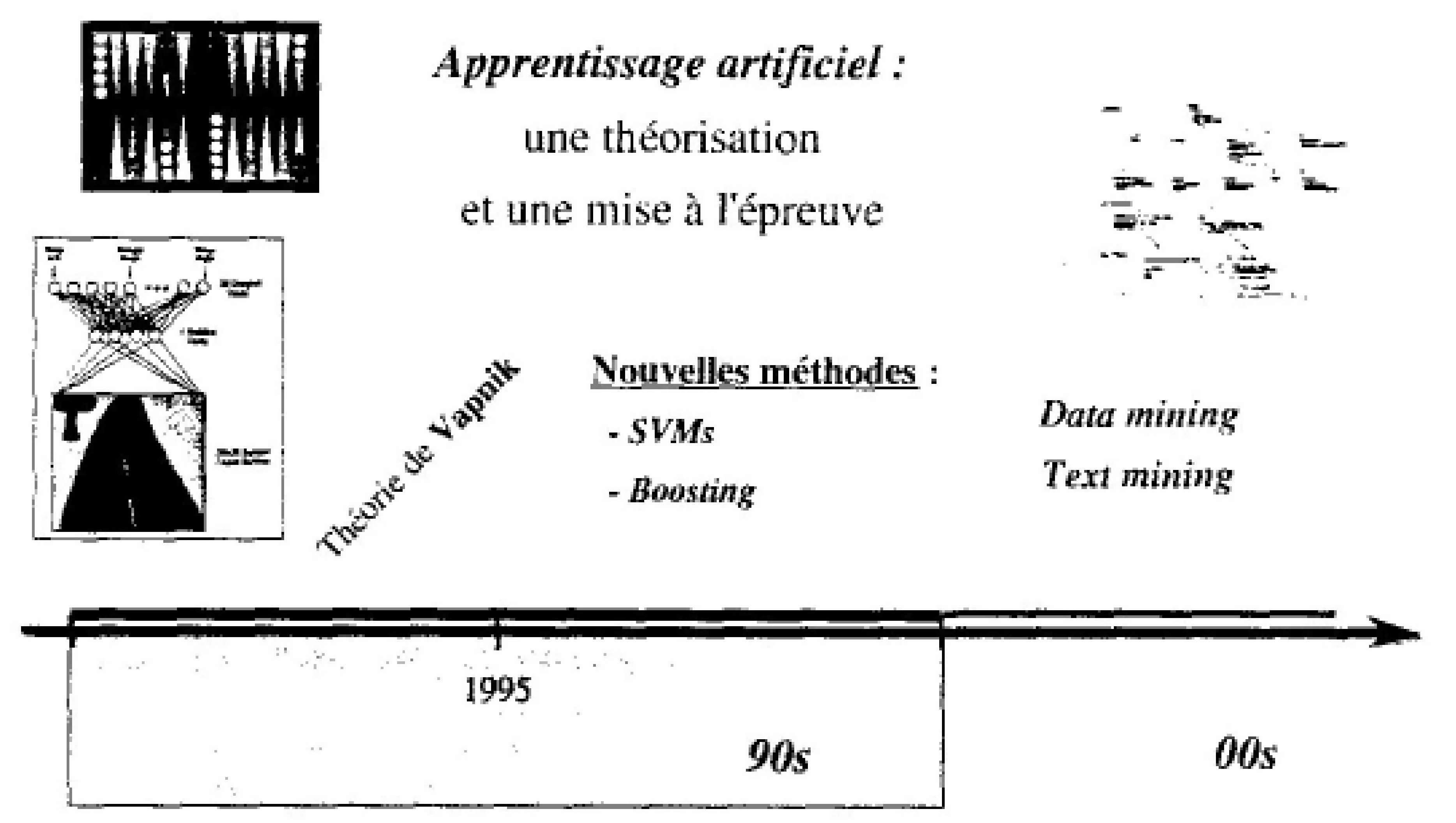
2ème connexionnisme

Figure 1-2 The second period of the artificial learning.

Of only one blow the learning was seen either as the search for algorithms simulating a task of learning, but like a process of elimination of hypothesis not satisfying an optimization criterion. It was then a question within this research framework how a sample of data drawn by chance could make it possible to identify a good hypothesis in a given space of hypotheses. It was extremely misleading, and as the language used in this new research direction was rather distant from that of the experts of the artificial learning, those continued to develop algorithms simpler but more general than those of the previous decade: decision trees, genetic algorithms, induction of logical programs, etc.

It is only in the years 1990, and especially after 1995 and the publication of a small book of Vapnik (1995), that the statistical theory of the learning truly influenced the artificial learning by giving a solid theoretical framework to the interrogations and empirical observations made in the practice of the artificial learning.

The current development of the discipline is dominated at the same time by a vigorous theoretical effort in the directions opened by Vapnik and the theorists of the statistical approach, and by redeployment towards the application of the developed techniques to great applications of economic purpose, as the mining of socio-economic data, or with finality, like the genomic one. It is undeniable that for the moment the learning is felt like necessary in very many fields and that we live a golden age for this discipline. That should not however forget the need for joining again the dialogue with the psychologists, the teachers, and more generally all those which work for the learning in a form or another.



Apprentissage artificiel :

une théorisation

et une mise à l'épreuve

Théorie de Vapnik

Nouvelles méthodes :

- ***SVMs***
- ***Boosting***

Data mining

Text mining

1995

90s

00s

Figure 1-3 The third period of the artificial learning.

A non-exhaustive list of reviews specialized on the artificial learning is:

- *Machine Learning Journal*
- *Journal of Machine Learning Research* (available free on http://www.ai.mit.edu/projects/jmlr/)
- *Journal of Artificial Intelligence Research* (JAIR) accessible free on Internet (http://www.ai.mit.edu/projects/jmlr/)
- *Data Mining and Knowledge Discovery Journal*
- *Transactions on Knowledge and Date Engineering*

Table 1.1 - Core tasks for Machine Learning

Task category	**Specific tasks**
Classification	Classification, Theory revision, Characterization, Knowledge refinement, Prediction, Regression, Concept drift
Heuristics	Learning heuristics, Learning in Planning, Learning in Scheduling, Learning in Design, Learning operators, Strategy learning, Utility problem, Learning in Problem solving, Knowledge compilation
Discovery	Scientific knowledge discovery, Theory formation, Clustering
Grammatical inference	Grammar inference, Automata Learning, Learning programs
Agents	Learning agents, Multiagent system learning, Control, Learning in Robotics, Learning in perception, Skill acquisition, Active learning, Learning models of environment
Theory	Foundations, Theoretical issues, Evaluation issues, Comparisons, Complexity, Hypothesis selection
Features/Languages	Feature selection, Discretization, Missing value handling, Parameter setting, Constructive induction, Abstraction, Bias issues
Cognitive Modeling	Cognitive modeling

The Information Society Technologies Advisory Group (ISTAG) has recently identified a set of “grand research challenges” for the preparation of FP7 (July 2004). Among these challenges are

- The 100% safe car
- A multilingual companion
- A service robot companion
- The self-monitoring and self repairing computer
- The internet police agent
- A disease and treatment simulator
- An augmented personal memory
- A pervasive communication jacket
- A personal everywhere visualiser
- An ultra light aerial transportation agent
- The intelligent retail store

If perceived from an application perspective, a multilingual companion, an internet police agent or a 100% safe car are vastly different things. Consequently, such systems are investigated in largely unconnected scientific disciplines and will be commercialized in various industrial sectors ranging from health care to automotive.