THE UNIFIED LEARNING PARADIGM: A FOUNDATION FOR AI

Lev Goldfarb Sandeep Nigam

Faculty of Computer Science
University of New Brunswick
Box 4400,
Fredericton, N.B., E3B 5A3
Canada

goldfarb@unb.ca

Induction has been called "the scandal of philosophy". Unless more consideration is given to the question of which form of knowledge representation [which model] is appropriate for mechanized inductive inferences, I am afraid that induction may become a scandal of AI as well.

Peter Gardenfors, Induction, conceptual spaces and AI, *Philosophy of Science* 57, p.93 (1990)

1. Introduction

As one of us has already repeatedly stressed ([10], [12], [13], [15]), we believe, together with Hermann von Helmholtz [23], that the central and the most pressing issue confronting cognitive science and artificial intelligence is the development of a satisfactory unified inductive learning model (see also [5], [34], [43]).

Unfortunately, this issue was not perceived to be the central issue by the three leading (and founding) schools of AI, which had a very negative effect on the development of AI up to now. In particular, due *only* to the difference between the formal models used originally in some areas of AI and pattern recognition, AI had severed practically all ties with pattern recognition, which was very counter-productive to the development of both areas and particularly to AI.

With the recent rise of connectionism, this situation has begun to change, which is reflected in the content of the recent AI textbooks ([39], [49]). As a result, another issue was brought to the fore: the existing learning models were found grossly inadequate. The present collection of papers is, basically, supposed to address the latter situation.

In this chapter we shall briefly outline the answer that was originally proposed in [15], [11], [12] (see also [5], [34], [42], [43], [19]). In accordance with the topic of the present book, we shall also discuss how the symbolic/numeric dichotomy is resolved within the new (inductive) learning model, and how this model captures the concept of inductive generalization. If, as we mentioned above, at the initial stage of development of AI this dichotomy was overemphasized, the recent connectionist movement, which was revitalised by psychologists, has completely underestimated the formal implications of the dichotomy.

2. On the proper formulation of the inductive learning problem

In view of the space limitation and the fact that we plan to devote a separate paper to this topic, this section is relatively brief.

"Induction is a process of producing a general proposition on the ground of a limited number of particular propositions such that these become special instances of the former. It is a process of going from particular to universal, from concrete to abstract" [48, p. 450].

"Whenever reasoning occurs, one can ask: does the conclusion contain more semantic information than the premises? More precisely, does it rule out some additional state of affairs over and above those ruled out by the premises? If not, then the inference is a valid deduction: its conclusion is true in any situation in which the premises are true. But, if the conclusion does rule out additional state of affairs, it is not valid. *Induction* can be defined as a systematic way of reasoning that increases the given information" [24, p. 219].

"Induction raises perhaps the most difficult problem in the whole theory of knowledge. Every scientific law is established by its means, and yet it is difficult to see why we should believe it to be a valid logical process. … When mankind took to science, they tried to formulate logical principles justifying this kind of inference [induction]. … I will only say that they [these attempts] seem to me very unsuccessful. I am convinced that induction must have validity of some kind in some degree, but the problem of showing how or why it can be valid remains unsolved" [40, p. 11].

In spite of its critical role in cognitive science, neuroscience, AI, and pattern recognition, the problem of inductive learning and its relation to other problems is still not properly understood. For example, in one of the best introductions to cognitive science [24], the chapter on "Induction, concepts and probability" is placed outside part III, "Learning, memory and action", in which we find the chapter on "Learning and learnability." Moreover, it now appears that the entire vector space setting for the pattern recognition problem, including its latest, connectionist, form is absolutely inappropriate (see section 2.2 below).

What is the correct formulation of the inductive learning problem? The *basic* problem can be stated as follows:

Given a finite set C^+ of positive training objects that belong to possibly infinite set C (concept) to be learned and a finite set C^- of negative training objects that do not belong to the concept C, find an analytical model that would allow one to recognize if a new element belongs to the concept C. In other words, on the basis of the finite training set

$$C^+ \cup C^-, \qquad C^+ \cap C^- = \varnothing \qquad (1)$$

the model must be able to form the "idea" of *inductive* generalization corresponding to the concept C.

One should immediately stress that it is meaningless to speak of recognition without the concept of (inductive) generalization, where generalization is simply the state of the agent after the inductive learning process. The concept, or "pattern", formation stage can then be followed by the pattern recognition stage.

The reason why a formal model of inductive learning must be a fundamentally new mathematical model can be intuited from the following fact. To construct such a model, one must find a (new) way of "encoding", or capturing, a possibly infinite set C of positive objects based on the finite training set. Mathematicians, so far, have not addressed this question, and so no scheme to answer it exists within the present mathematics. The potential model should be able to extract the structure of the class C from its finite subset C^+. By *the structure of the class* we mean: 1) the *symbolic* features that make the objects of the class similar to each other and/or different from other objects, and 2) the *emergent*[1] interrelationship among these features. The inductive learning process would then involve the *discovery* (and encoding) of the structure of the class.

Within a typical mathematical model, the structure of the entire set of objects is postulated by a set of axioms, and then one proceeds to study the analytical consequences of the postulates. Besides, so far, it has been the case in mathematics that the only "legitimate" substructures of the "basic" structure are those that satisfy the same set of axioms as the "basic" set, e.g. subgroup, subspace, etc. (See, for example, [9], [27]). On the other hand, in the case of inductive learning model, the axiomatic structure of the basic set is not postulated, and, in addition, the structure of various concepts, though they satisfy some constraints, can be quite different. In connection with this, it is not difficult to understand why Chomsky [37, p. 103] was objecting to the suggestion "that somehow the kind of investigation conducted by Bourbaki [a leading trend-setting group of mathematicians] leads to a theory of structures that relates to the question of [inductive learning model] …''. Of course, this is not to say, as Chomsky does, that such a *new* mathematical model cannot exist. In fact, as was mentioned above, such a model was proposed in [15], [11] and will be outlined below in section 5.

3. Fundamental inadequacy of the present symbolic and numeric learning paradigms

First of all, in order to approach the task of constructing a satisfactory formal learning model, one has to try to understand the fundamental weaknesses of the two basic formal paradigms - vector space based learning models and purely symbolic, or non-numeric, learning models. These weaknesses become particularly visible when the formal paradigms that were developed for entirely different purposes are being adopted for modelling of learning processes, whose main feature; irreversibility; was not captured by those paradigms.

3.1 Inadequacy of the symbolic paradigm

By the symbolic paradigm[2] we mean any analytical learning model that is a fundamentally discrete model, i.e., the topology of the basic set of objects involved in the model is discrete topology: every one-element set is both an open and a closed subset. Any finite or infinite language over a finite alphabet is a typical example of the set involved in the symbolic paradigm. By a

[1] The meaning of the term will be clarified in section 5.1, in which some formal definitions are presented.

fundamentally discrete model, we mean that the underlying discrete set is not a natural subset of some larger non-discrete set, as in the case of any set of integer numbers, which is also a subset of real numbers. The main source of the discrete models is the classical work on the foundation of computability, which, in turn, goes back to the foundations of mathematical logic.

Within AI, a high point for the symbolic paradigm was 1975 Turing Award Lecture "Computer Science As Empirical Inquiry: Symbols and Search" by Newell and Simon, in which "The Physical Symbol System Hypothesis" was stated:

> A physical symbol system has the necessary and sufficient means for general intelligent action. [33, p. 116]

They also state that their

> understanding of symbol systems grew through a sequence of stages. Formal logic familiarized us with symbols, treated syntactically, as the raw material of thought, and with the idea of manipulating them according to carefully defined formal processes. The Turing machine made the syntactic processing of symbols truly machine-like, and affirmed the potential universality of strictly defined symbol systems. The stored-program concept for computers reaffirmed the interpretability of symbols, already implicit in the Turing machine. List processing brought to the forefront the denotational capacities of symbols, and defined symbol processing in ways that allowed independence from the fixed structure of the underlying physical machine. By 1956 all of these concepts were available, together with hardware for implementing them. The study of the intelligence of symbol systems, the subject of artificial intelligence, could begin. [33, p. 126]

The main point of this section is that, in a definite sense, the above hypothesis is incomplete: from analytical point of view the discrete model is not adequate to capture the most basic and necessary part of an intelligent process - inductive learning.

In [32, p. 478] Newell admits that there

> is a difference in world view between myself and many others, including Goldfarb. *Deduction*, *induction*, and *abduction* have always seemed to me a particularly superficial way to classify types of reasoning. *Induction* plays no specific role in my thinking …

Indeed, the inadequacy of the above purely symbolic paradigm may not be very visible in the tasks that do not immediately involve inductive learning.

In contrast, one of the leading mathematicians of this century John von Neumann insisted on the inadequacy of the symbolic paradigm as far ago as the late 1940s. In view of the importance of his remarks, which he repeatedly made during the late 40s, they are reproduced here from "The General and Logical Theory of Automata", read at the Hixon Symposium in September 1948 and published in 1951. It is important to note that the same work reviews the relevant developments, including "The McCulloch-Pitts Theory of Formal Neural Networks", which von Neumann brought to the attention of others, as well as "Turing's Theory of Computing Automata". Von Neumann himself made outstanding contributions to logic and had excellent knowledge of the contemporary logical developments. Since no fundamentally new relevant developments in symbolic logic occurred since the late 40s, his remarks below are as relevant today as they were then.

> We are very far from possessing a theory of automata which deserves that name, that is, a properly mathematical-logical [continuous-discrete, or numeric-symbolic] theory. There exists today a very elaborate system of formal logic and specifically, of logic as applied to mathematics. This is a discipline with many good sides but also with certain serious weaknesses. … About the inadequacies, however, this may be said: Everybody

> who has worked in formal logic will confirm that it is one of the technically most refractory parts of mathematics. The reason for this is that it deals with rigid, all-or-none concepts and has very little contact with the continuous concept of the real or of the complex number, that is, with mathematical analysis. Yet analysis is the technically most successful and best elaborated part of mathematics. Thus formal logic is, by the nature of its approach, cut off form the best cultivated portions of mathematics and forced onto the most difficult part of the mathematical terrain, into combinatorics.
>
> The theory of automata, of the digital, all-or-none type … is certainly a chapter in formal logic. It would, therefore, seem that it will have to share this unattractive property of formal logic. It will have to be, from the mathematical point of view, combinatorial rather than analytical. [30, pp. 99 - 100]
>
> All of this will lead to theories [future theories] which are much less rigidly of an all-or-none nature than past and present formal logic. They will be of a much less combinatorial, and much more analytical, character. In fact, there are numerous indications to make us believe that this new system of formal logic will move closer to another discipline which has been little linked in the past with logic. This is thermodynamics … and is that part of theoretical physics which comes nearest in some of its aspects to manipulating and measuring information. Its techniques are indeed much more analytical than combinatorial, which again illustrates the point that I have been trying to make above. [30, p. 101]

Essentially, what von Neumann is saying is that a purely discrete model is not powerful enough to fit the bill. Note, however, that he does not suggest to eliminate the discrete part of the model, as was done by the connectionists, but rather to enlarge the classical logical model by introducing in it in some, possibly new, form a continuous structure. This can be more clearly seen from another passage taken from his 1949 University of Illinois lectures [31, p. 62]:

> I have been trying to justify the suspicion that a theory of information is needed and that very little of what is needed exists yet. Such small traces of it which do exist, and such information as one has about adjacent fields indicate that, if found, it is likely to be similar to two of our existing theories: formal logics and thermodynamics. It is not surprising that this new theory of information should be like formal logics, but it is surprising that it is likely to have a lot in common with thermodynamics.
>
> Though this new theory of information will be similar to formal logics in many respects, it will probably be closer to ordinary [continuous] mathematics than formal logics is.

Another leading mathematician, A. N. Kolmogorov, also stressed the importance of continuous structures in the model of intelligent automata:

> Our discussion leads us to the vital conclusion that information and control processes in living organisms are undoubtedly based on a complex interaction between discrete (digital) and continuous mechanisms which perform both deterministically and probabilistically. The discrete mechanisms, however, play a leading part in the information and control processes in living organisms. [26, p. 30]

The inadequacy of the discrete models were also pointed out by a number of cognitive scientists (see, for example, [4], [35]).

The analytical poverty of the symbolic models, which von Neumann stressed, has resulted in the quite prominent lack of efficient inductive learning algorithms in AI. Moreover, in the light of the above remarks, the entire attempt to develop them within the purely symbolic setting should be considered as absolutely futile.

It is interesting to note that the evidence from biological central nervous systems is unequivocal about the critical role of non-discrete (non-symbolic) mechanisms:

> The great majority of synapses in the mammalian (and human) brain are chemical. … The difference between electrical and chemical synapses is an important one. A chemical synapse is by nature plastic, it can be altered in many ways and at many different steps to increase or decrease its activity. Electrical synapses are rigid and unchanging. … Learning and memory could not develop in a nervous system that had only electrical synapses. [46, p. 99]

> Contrary to the dogma of the 1940s, the brain is almost completely an analog machine, not a digital one. … Digital coding in the nervous system is used only to transmit information over long distances. Information about sensory events in the foot, for instance, travels to the spinal cord as a series of digital spikes called action potentials; it is broken down in the spinal cord to an analog code for determining what to do about the events. [2, p. 8]

Finally, it appears that within a purely symbolic model the resolution of the semantic paradox, which was brought to the attention of AI and cognitive scientists by Searle ([44], [45]), is not possible:

> The failure of AI can also be understood as the impossibility of *grounding* symbol meanings within a pure symbol system. … Without the projected interpretation, the system, be it ever so systematically interpretable, is hanging from a sky hook; it is ungrounded. Hence, on pain of infinite regress, the system in our heads that is capable of projecting meanings onto symbol systems cannot itself be merely a symbol system. Our symbol meanings must be grounded non-symbolically. [22, p. 78]

3.2 Inadequacy of the vector space learning paradigm[3]

In spite of the fact that we rejected the classical symbolic learning paradigm, the presently popular alternative, connectionist paradigm, is also fundamentally inadequate as a learning model.

The main reason for the inadequacy is the choice of a vector space as the object representation space. By its very definition, a vector space, which emerged as a convenient model for representing some physical processes, is not suited for the inductive learning process that is supposed to provide solution to the problem formulated in section 2. Why?

From the axiomatic definition of a vector space it follows that the only “legitimate” classes/concepts in such a space are its vector subspaces of various dimensions and some (polyhedral) subsets associated with them[4] . The representation of the concept/generalization by a polyhedral region in an n-dimensional Euclidean space implies the trivialization of the learning process. This is because, first, such concept representation allows the agent “to learn” the *exact* class “generalization” only in the case when the training set contains a set of vectors that span the boundary (which is a piecewise linear set) of the corresponding region. The reason why we stress *exact* class generalization is that when analyzing the *abstract (mathematical)* model of inductive learning, one should demand of it, by analogy with our own abilities, the capability to capture *the exact structure of the class* based on the finite training set.

[3] Of necessity, the treatment of the topic is very brief.

[4] See also the last paragraph in section 2.

Second, even if such a polyhedral region or, more realistically, its approximation is constructed, it has no relation to the *structure* of the concept that is being learned: the constructed boundary hyperplanes reveal no information about the structure of the original class [34]. In fact, the constructed region *cannot* reveal any structural information, simply because the vector space is not an appropriate mathematical model for capturing structural class properties. The latter cannot be accomplished without participation of some conventional symbolic language.

The third fundamental deficiency of the connectionist models, related to the above, is a principal absence of a natural formal framework for the process of new feature construction during the learning process. It should be clear that to encode the concept structure the agent must be able to discover appropriate new features (see also section 2).

Fourth, on a more formal level, one can show that even for a very small training set of structural pattern, with a natural distance function, there is *no* Euclidean space in which one can isometrically (preserving the interdistances) represent the training set. Only expanding the class of Euclidean spaces to the class of pseudo-Euclidean, or Minkowski, spaces can such a representation be constructed. But, even then, such representation can be carried out only after the interdistance matrix for the training set is computed, i.e., only after the corresponding structural (non-vector) distance measure is obtained [16, chapters 1-4]. As we shall argue below, we believe that the inductive learning is the process of synthesis of the corresponding distance measure.

For a related discussion of the deficiencies of the vector space learning models see [15, section 4.1].

To summarise, the weight adjustment process, so characteristic of the connectionist models, captures only one side of the learning process. The other and perhaps more fundamental side of the learning process is construction of new features, which together with the corresponding weights (plus the training set) can capture the inductive concept structure (see section 5.2). The vector space is not the most appropriate mathematical model to accommodate such a learning process. The needs of such a process require, as was mentioned above, an introduction of the new mathematical model - evolving transformation systems. As we shall see, the new model is not discrete and is not continuous but a new symbiosis of both.

As we mentioned at the beginning of this section, the connectionist models are inadequate as inductive learning models. One of the main reasons they were mistaken for such is that *in some cases* they can be used as recognition models either directly or after the learning process has been completed. The learning process can be viewed as the process of the features discovery as well as the process of synthesis of the appropriate distance function defined on the object set (see sections 4, 5). Once the learning process is completed, i.e. the appropriate features as well as the corresponding distance function are constructed, the vector space recognition model can then be used as an efficient *auxiliary* tool for the on-line recognition process (see, for example, [15, section 3.3] or [14]). We are planning a separate paper [14] to clarify the relation between the two stages.

4. Original motivation for the unified learning model

The new learning model ([15], [11]) emerged as the result of an attempt to find the general model for pattern recognition that would unify two basic but analytically incompatible recognition models: the vector space, or geometric, model and the syntactic, or structural, model ([16], [17]). Thus, it is interesting to note that the goal of the unification that prompted the present collection of papers was the driving motivation behind the development of the new learning model, which started 15 years ago [18].

From the very beginning of the above development it became clear that the concept of distance function, or dissimilarity measure, is the most natural tool for unifying the symbolic and the numeric models. This answer has already been provided by the historical development of pattern recognition itself, as well as some other fields of knowledge, which contributed to the emergence of pattern recognition, such as biological taxonomy and psychology:

> Similarity [or, equivalently, dissimilarity] plays a fundamental role in theories of knowledge and behaviour. It serves as an organizing principle by which individuals *classify objects, form concepts, and make generalizations.* [47, p. 257, our emphasis]

In other words, from the very beginning of pattern recognition it was clear that the distinctive feature of the recognition problem is the similarity based construction of the decision rule from the finite training set: if $o_1 \in C^+$ and there is a way of measuring the degree of "closeness" of o_2 to o_1 then one can also judge about the degree of membership of o_2 in the class $C \supset C^+$. Formally, the ability to relate objects by the degree of their closeness is expressed by the introduction of distance function on the space of objects, i.e., functions that for every pair of objects o_1 and o_2 uniquely assign a non-negative number d_{ij}:

$$d_{ij} = \Delta (o_1 , o_2) , \qquad (2)$$

where $\Delta (o_1 , o_2)$ satisfies some simple axioms. Definitions or, more accurately, construction of this function determines to a considerable degree the decision rule. It is the latter part that is a creative, non-trivial stage in the solution of the pattern recognition problem. Depending on the approach to the construction of Δ, one obtains different recognition rules, in spite of the fact that the same training set was the initial input [38, pp. 169 - 170].

The emergence and popularity of the Levenshtein, or edit, distance between two strings [8, p. 248] reinforced the view of the fundamental role of the concept of distance function in pattern recognition.

Having postulated the distance function as the basic underlying and unifying concept, a new analytical scheme was proposed ([16], [17], [18]) to improve the efficiency and reliability of the decision algorithms for the patterns represented as elements of a pseudo-metric, including metric, space. In essence, the scheme shows how to efficiently transform the decision problem from the general distance space - in which no global structure to support efficient decision algorithms exists - to that in a low-dimensional Minkowski, or pseudo-Euclidean, vector space - in which efficient mechanisms similar to the connectionist schemes can be employed.

Thus, during the first stage in the development of the new learning model, a general analytical scheme for organizing the recognition process under very general metric assumptions was proposed. However, the question of the dynamics of the distance function itself - its structure, and its evolution during learning - remained unanswered.

During the second stage in the development, first outlined in [15], the latter question was basically answered, and it became clear that the name inductive learning should most appropriately refer to the newly emerged process, in which the "right" metric in the learning environment is synthesized. It also became clear that none of the existing mathematical models can serve the needs of the latter process, and, therefore, previously, one could not even properly state or address the above question, despite its overwhelming importance to AI and cognitive science. We now turn to a very brief outline of this inductive learning model.

5. Inductive learning as evolving transformation system

Both areas, artificial intelligence and pattern recognition, despite their intention to model the corresponding cognitive processes, have not offered any fundamentally new, irreversible, models of the processes. While the irreversibility of the basic cognitive processes became apparent already in the last century, the most developed of the natural sciences, physics, has not offered even a single irreversible model (see, for example, [7, lecture 5]). Thus, without any useful analytical paradigm to follow, without the realization of the need for such a new paradigm, and also without the realization of the profound connections between themselves, both of the above areas made negligible progress, especially considering the amount of invested resources.

In the following section the concept of a transformation system (TS) is outlined. The most important feature of the TS is that the new "continuous" structure associated with it is introduced by means of its set of operations. Since the set of operations can easily be modified, the corresponding continuous structure would also be modified, leading to the concept of evolving transformation system, an irreversible model briefly outlined in section 5.2.

5.1 Transformation systems

A *transformation system* (TS) is a 3-tuple

$$T = (\mathbf{O}, \mathbf{S}, \mathbf{D}) , \tag{3}$$

where $\mathbf{O} = \{o_1 , o_2 , \dots \}$ is a set of (analogously) *structured objects*, $\mathbf{S} = \{S_i\}_{i=1}^{m}$ is a finite set of reversible *substitution operations* that can transform one object into another, and $\mathbf{D} = \{\Delta_\omega\}_{\omega \in \Omega}$ is *competing family of distance functions* defined on $\mathbf{O}$ whose parameter set Ω is (m - 1)-dimensional unit simplex

$$\Omega = \{ \ \omega = (w^1 , \dots , w^m) \ \ | \ \ w^i \geq 0 , \ \sum_{i=1}^{m} w^i = 1 \ \}. \tag{4}$$

Each of the distance functions Δ_ω is defined as follows. Weight w^i is assigned to the operation S_i and

$$\Delta_\omega (o_1 , o_2) \quad = \quad \min_{\mathbf{s}_j = (S_1^{(j)},\dots,S_k^{(j)})} \left\{ \sum_{i=1}^{k} w^i_{(j)} \text{, where } \mathbf{s}_j \text{ is a sequence of operations that transforms } o_1 \text{ into } o_2 \right\}. \tag{5}$$

In other words, the minimum is taken over all possible sequences of operations that can transform structured object o_1 into structured object o_2. In order to ensure that Δ_ω is defined for every pair of objects, we shall assume that the set $\mathbf{S}$ of substitution operations is *complete*, i.e., there is enough operations to transform any object o_1 into any other object o_2. We cannot emphasize enough the critical role of the above definition for understanding of the model, since the above distances are new in mathematics.

To compute the above distance the system must *use its set of operations in a cooperative and competitive manner*. Thus, all properties of the system resulting from this definition should be viewed as *emergent* properties.

One of the basic examples of the transformation system is the *string* transformation system, in which: **O** is a subset of a set of strings over some finite alphabet; **S** is finite set of substitution operations, e.g. **aba** $\leftrightarrow$ **bc**, including deletion/insertion operations, e.g. **a** $\leftrightarrow \theta$, **ab** $\leftrightarrow \theta$ (θ is the null string); family **D** is defined as above. When the set **S** consists of single-letter deletion/insertion and substitution operations, the distance Δ_ω is called weighted edit, or Levenshtein, distance. These and some of their applications are considered at great length in [41] (see also [8], [29], [3]). There are efficient sequential and parallel algorithms for computing the Levenshtein distance between two strings, based on the idea of dynamic programming. Although for general set **S** the distance Δ_ω may not be computable, we shall not be concerned with this possibility, since the cases we are interested in correspond to the set **S** that contain all single-letter deletion/insertion operations, which result in computable Δ_ω.

It is very instructive to compare the above definition of the distance function Δ_ω, for a fixed weight vector $\omega = (w^1, \dots, w^m)$, with the most celebrated of physical principles, *the principle of least action* ([50], in particular, pp. 29, 47, 163-64). One of the most general forms of this principle, Hamilton's principle, can be stated as follows: a system moves from one configuration to another in such a way that the variation of the integral[5] $\int Ldt$ between the path taken and a neighbouring virtual path, co-terminous in space and time with the actual path, is zero. In other words, $\int Ldt$ is stationary [50, p.47].

One can show that the currently popular neural net systems are a special case of the TS, in which the corresponding operations are coupled in a fixed manner ([15], [20]).

An example of learning in the TS is given in [15]. Concept learning in a TS reduces to the following optimization problem:

$$\max_{\omega \in \Omega} f(\omega) , \qquad f(\omega) = \frac{f_1(\omega)}{c + f_2(\omega)} \tag{6}$$

where $f_1(\omega)$ is the Δ_ω-distance *between* C^+ and C^- (see section 2), $f_2(\omega)$ is the average Δ_ω-distance *within* C^+, and c is a small positive constant to prevent the overflow condition (when the values of f_2 approach 0).

Let us denote by Ω_{max} a subset of Ω consisting of all the (global) maximums of f on Ω. It is easy to see that, for the given concept C and given set **S** of operations, every weighting scheme $\omega^* \in \Omega_{max}$ generates the "best geometric configuration" of the training examples: under Δ_{ω^*} positive examples form the most compact set relative to the negative examples. Thus we are justified in calling function f the *quality of the (learning) class perception* ([11]).

It is not difficult to see that if for $\omega^* \in \Omega_{max}$

$$f_2(\omega^*) = 0 \quad \text{and} \quad f_1(\omega^*) \neq 0 , \tag{7}$$

the set **S** of operations is sufficient to produce a *complete separation* of C^+ and C^-. This is "often" not the case, since then the learning agent has no need to acquire any new operations, or new

[5] The lagrangian L is the difference between the system's kinetic and potential energies.

"features".

It turns out that in some cases the above continuous optimization problem can be reduced to a relatively simple discrete optimization problem on a small subset of Ω - the set of its vertices plus the gravity centers, or barycenters, of some of its subsimplexes.

5.2 Evolving transformation systems

When the set **S** of substitution operations is not sufficient to achieve a *complete separation* of C^+ and C^- (see the last subsection), the structure of the model allows for a quite natural modification of the set **S**: all that one has to do is to add successively appropriate new operations to the set **S**. With every addition of a new operation, the former TS changes into a new TS, resulting in a sequence of TS.

More formally, by *evolving transformation system* (ETS) [11] we shall mean a finite or infinite sequence of TS's with a common set of structured objects

$$T_i = (\mathbf{O} , \mathbf{S}_i , \mathbf{D}_i , \mathbf{CR}) \qquad i = 0, 1, 2, \ldots \qquad (8)$$

in which each set of operations $\mathbf{S}_i$, except $\mathbf{S}_0$, is obtained from $\mathbf{S}_{i-1}$ by adding to it one or several operations that are constructed from the operations in $\mathbf{S}_{i-1}$ with the help of a small set **CR** of *composition rules.* Each rule $R_i \in$ **CR** allows one to construct new operations from the existing operations.

For example, for the above string TS, from two deletion/insertion operations $\mathbf{a} \leftrightarrow \theta$, $\mathbf{b} \leftrightarrow \theta$ one rule may allow us to build a new operation $\mathbf{ab} \leftrightarrow \theta$ by concatenating two left-hand sides of the given operations.

From the above definition it follows that at a stage t of the ETS

$$\mathbf{S}_0 \subset \mathbf{S}_1 \subset \mathbf{S}_2 \subset \ldots \subset \mathbf{S}_t \qquad (9)$$

and for $i = 0, 1, \ldots , t - 1$ we have

$$\forall\, o_1, o_2 \in \mathbf{O}, \quad \forall\, \Delta_{\omega_1} \in \mathbf{D}_i \quad \exists\, \Delta_{\omega_2} \in \mathbf{D}_{i+1} \qquad \Delta_{\omega_2} (o_1, o_2) \le \Delta_{\omega_1} (o_1, o_2). \qquad (10)$$

Of course, $\omega_1 \in \Omega_i$ and $\omega_2 \in \Omega_{i+1}$, where the dimensions of simplex Ω_i is smaller than the dimension of simplex Ω_{i+1}. In fact, Ω_i is a subsimplex of Ω_{i+1}.

Roughly speaking, the inductive learning process for the ETS, illustrated in section 5.4, proceeds by constructing a sequence of $\mathbf{S}_i$'s in such a way that for the corresponding T_i the minimum value of f_2 decreases (while making sure that the respective value of f_1 is not zero), i.e., the interdistances in C^+ gradually shrinks to 0 while the distance between C^+ and C^- remains non-zero.

Returning to section 2, we can now answer the following question: How does the above model allows one to capture the structure of the learning class? The structure of the class (concept) is captured by the following 3-tuple

$$(\overline{C} , \mathbf{S}_t , \Omega^{(t)}_{max}) , \tag{11}$$

where $\overline{C}$ is a reduced training set, $\mathbf{S}_t$ is the final set of operations (at the termination of the learning process), and $\Omega^{(t)}_{max}$ is the set of weight vectors corresponding to the global maximums of f at the termination stage. We shall not discuss here the question of how to obtain set $\overline{C}$ from $C^+ \cup C^-$ (see, however, [15, section 3.3]).

In connection with the above form of (inductive) concept representation or generalization, it is important to note that this representation is in complete agreement with, probably, the most accepted theory of concept learning - Rosch's exemplar theory: "Exemplar theory claims that concept-learning is accomplished by memorizing specific instances and by using some measure of instance similarity"[1, p. 160].

5.3 Structurally unbounded environments

In the next section we illustrate the application of the model by presenting example of the learning process in a structurally unbounded environment. "The history of AI and pattern recognition suggests that learning in some environments is 'easy' and thus little insight into the most interesting parts of the learning process is gained, while learning in some other environments is simply impossible on computational grounds" [12]. The requirements that was proposed in [12] is called *structural unboundedness*: no finite set of features or parameters is sufficient for specifying all potential classes (concepts) in the learning environment. This requirement ensures that the family of all admissible classes cannot be enumerated by varying some finite number of numeric parameters or features, i.e., that no "closed form" description of the environment exists. Note that the existence of infinitely many admissible classes is a necessary but not a sufficient property of the structurally unbounded environment, as the following family of classes illustrates

$$\{(x, y, z) \mid x - z = m, \; x + 2y = 2n \}_{m, n \in \mathbf{R}} \tag{12}$$

The main motivation for introducing the above requirement is to ensure that the structure of the environment cannot be encoded in some form into the structure of the learning machine or in the learning program. This, in turn, implies that to learn in such environment the agent would need to be able to construct during the learning process new features from some internal finite set of features. We believe that if such a minimum requirement is not imposed on the structure of the learning environment, one cannot be sure that the learning model captures the most critical part of the inductive learning process: efficient construction of new features, which is the most characteristic attribute of the inductive learning process.

The environment we have chosen is the following infinite family $\boldsymbol{L}$ of regular languages over a finite alphabet. Each language L in $\boldsymbol{L}$ is defined by a string k, called the *kernel* of the language, and a finite set F of strings, called *features*, in which no string is a substring of another string. Language L is then constructed by forming all the strings that can be obtained from k by inserting in any place and in any order any number of strings from F: insertions into any of the already inserted strings from F is not allowed.

The structural unboundedness of the above environment is readily verified. The problem of learning in such environment is not trivial, since strings in F may have common sub-strings and may also have common substrings with the kernel k.

5.4 Example of learning with the evolving transformation system

Let us choose alphabet { a, b, c, d }, and the structurally unbounded environment described at the end of the last section. Let the positive and the negative training sets be as follows:

L^+	L^-
aaacbacd	abcd
aabbacdaac	aacab
abaaaccdab	dcbaabc
abababacd	abadcacb
abaabcaacd	bacdaabac
ababaabaaccdabaac	cababaccda

The initial set $\mathbf{S}_0$ of operations consists of the following deletion/insertion operations:

$$\mathbf{a} \leftrightarrow \theta \qquad \mathbf{b} \leftrightarrow \theta \qquad \mathbf{c} \leftrightarrow \theta \qquad \mathbf{d} \leftrightarrow \theta$$

where θ is the null string. To facilitate the understanding of the learning algorithm, instead of giving the pseudocode, the results of application of the corresponding learning algorithm to the above training set (L^+, L^-) are presented. For details see [34], [42]. All numbers are rounded to two decimal places.

The learning algorithm applied to (L^+, L^-)
----

1. Compute the value of function f_2 for the weight vector ω = (0.25, 0.25, 0.25, 0.25), which is the center of the initial 3-dimensional weight simplex Ω_0 in four-dimensional space $\mathbf{R}^4$:

$$f_2\,(0.25, 0.25, 0.25, 0.25) = 1.6.$$

Since the value is non-zero, continue.

The current L^+-interdistance matrix for the above weight vector is (entry *ij* is Δ_ω-distance between the corresponding pair of strings in L^+):

0.00	1.50	1.50	0.75	1.00	2.75
1.50	0.00	2.00	1.25	1.50	1.75
1.50	2.00	0.00	1.25	1.50	1.75
0.75	1.25	1.25	0.00	0.75	2.50
1.00	1.50	1.50	0.75	0.00	2.25
2.75	1.75	1.75	2.50	2.25	0.00

2. Compute the values of f_2 at the vertices of Ω_0

$$f_2\ (1, 0, 0, 0) = 2.2$$
$$f_2\ (0, 1, 0, 0) = 1.0$$
$$f_2\ (0, 0, 1, 0) = 0.67$$
$$f_2\ (0, 0, 0, 1) = 0.0\ .$$

Record letter **a** (i.e. the corresponding operation **a** $\leftrightarrow \theta$), for which the value of f_2 is maximal.

3. Check if any of the combinations of the letters recorded in step 2 (**a**) with each of the original 4 letters (**a**, **b**, **c**, **d**) are present as substrings in at least one of the strings from L+. Note that at this step, to obtain new operations ("combinations"), the only composition rule is applied, i.e. **CR** = $\{R_1\}$. There are 6 such combinations - **aa**, **ab**, **ba**, **ac**, **ca**, **da** - corresponding to the following deletion/insertion operations:

aa $\leftrightarrow \theta$ **ab** $\leftrightarrow \theta$ **ba** $\leftrightarrow \theta$ **ac** $\leftrightarrow \theta$ **ca** $\leftrightarrow \theta$ **da** $\leftrightarrow \theta$.

Each of these operations is, in turn, added to the initial set $\mathbf{S}_0$ to obtain six expanded sets of operations (each containing 5 operations): $\mathbf{S}_1^1$, $\mathbf{S}_1^2$, $\mathbf{S}_1^3$, $\mathbf{S}_1^4$, $\mathbf{S}_1^5$, $\mathbf{S}_1^6$. Six values of f_2 (for each of the six sets of operations) is computed for the same weight vector corresponding to the center of the original simplex (all constructed new operations, i.e. defining features, carry weight 0 throughout the algorithm):

$\mathbf{S}_1^1$: $f_2\ (0.25, 0.25, 0.25, 0.25, 0) = 1.1$

$\mathbf{S}_1^2$: $f_2\ (0.25, 0.25, 0.25, 0.25, 0) = 0.9$

$\mathbf{S}_1^3$: $f_2\ (0.25, 0.25, 0.25, 0.25, 0) = 1.03$

$\mathbf{S}_1^4$: $f_2\ (0.25, 0.25, 0.25, 0.25, 0) = 1.07$

$\mathbf{S}_1^5$: $f_2\ (0.25, 0.25, 0.25, 0.25, 0) = 1.5$

$\mathbf{S}_1^6$: $f_2\ (0.25, 0.25, 0.25, 0.25, 0) = 1.47$.

Record combination **ab** (i.e. the corresponding operation), for which the value of f_2 is minimal.

4. As at the beginning of step 3, check if any combinations of the recorded string **ab** with each of the original 4 letters, i.e. operations from $\mathbf{S}_0$, are present as substrings in at least one of the strings from L+. In what follows we will call this step *search for the candidate expanded combinations*. There are 6 candidate operations:

aba $\leftrightarrow \theta$ **aab** $\leftrightarrow \theta$ **abb** $\leftrightarrow \theta$ **bab** $\leftrightarrow \theta$ **abc** $\leftrightarrow \theta$ **dab** $\leftrightarrow \theta$.

As in step 3, we obtain six new expanded sets of operations each containing 5 operations: $\mathbf{S}_2^1$, $\mathbf{S}_2^2$,

$\mathbf{S}_2^3$, $\mathbf{S}_2^4$, $\mathbf{S}_2^5$, $\mathbf{S}_2^6$. Again, as above, value of f_2 (for each of the six sets of operations) is computed for the same weight vector:

$$
\begin{array}{ll}
\mathbf{S}_2^1: & f_2\,(0.25,\ 0.25,\ 0.25,\ 0.25,\ 0) = 1.15 \\
\mathbf{S}_2^2: & f_2\,(0.25,\ 0.25,\ 0.25,\ 0.25,\ 0) = 1.33 \\
\mathbf{S}_2^3: & f_2\,(0.25,\ 0.25,\ 0.25,\ 0.25,\ 0) = 1.55 \\
\mathbf{S}_2^4: & f_2\,(0.25,\ 0.25,\ 0.25,\ 0.25,\ 0) = 1.48 \\
\mathbf{S}_2^5: & f_2\,(0.25,\ 0.25,\ 0.25,\ 0.25,\ 0) = 1.5 \\
\mathbf{S}_2^6: & f_2\,(0.25,\ 0.25,\ 0.25,\ 0.25,\ 0) = 1.4\ .
\end{array}
$$

Since the minimum of the last six values is larger than the previous minimum of f_2 (for $\mathbf{S}_1^2$), stop the process of new operation construction and replace the initial set $\mathbf{S}_0$ by the set $\mathbf{S}_1 = \mathbf{S}_1^2$. The current L+-interdistance matrix for $\mathbf{S}_1$ and the above weight vector (0.25, 0.25, 0.25, 0.25, 0) is:

$$
\begin{array}{cccccc}
0.00 & 1.50 & 0.50 & 0.75 & 1.00 & 1.25 \\
1.50 & 0.00 & 1.00 & 0.75 & 0.50 & 0.75 \\
0.50 & 1.00 & 0.00 & 0.75 & 0.50 & 0.75 \\
0.75 & 0.75 & 0.75 & 0.00 & 0.75 & 1.50 \\
1.00 & 0.50 & 0.50 & 0.75 & 0.00 & 1.25 \\
1.25 & 0.75 & 0.75 & 1.50 & 1.25 & 0.00
\end{array}
\quad .
$$

2´. Go to step 2 with $\mathbf{S}_0$ replaced by $\mathbf{S}_1$ and keeping in mind that the weight of the new operation is always fixed to be zero:

$$
\begin{array}{l}
f_2\,(1,\ 0,\ 0,\ 0,\ 0) = 0.33 \\
f_2\,(0,\ 1,\ 0,\ 0,\ 0) = 0 \\
f_2\,(0,\ 0,\ 1,\ 0,\ 0) = 0.67 \\
f_2\,(0,\ 0,\ 0,\ 1,\ 0) = 0.0\ .
\end{array}
$$

Record letter **c**, for which the value of f_2 is maximal.

3´. Search for the candidate expanded combinations of the recorded letter **c** with each of the five operations in $\mathbf{S}_1$. There are 6 such combinations:

$$
\mathbf{ca} \leftrightarrow \theta \qquad \mathbf{ac} \leftrightarrow \theta \qquad \mathbf{cb} \leftrightarrow \theta \qquad \mathbf{bc} \leftrightarrow \theta \qquad \mathbf{cc} \leftrightarrow \theta \qquad \mathbf{cd} \leftrightarrow \theta\ .
$$

As in step 3 (and step 4), we obtain six expanded new sets of operations each containing 6 operations: $\mathbf{S}_3^1$, $\mathbf{S}_3^2$, $\mathbf{S}_3^3$, $\mathbf{S}_3^4$, $\mathbf{S}_3^5$, $\mathbf{S}_3^6$. Again, as above, we compute

$$\mathbf{S}_3^1 : \quad f_2\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.83$$
$$\mathbf{S}_3^2 : \quad f_2\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.4$$
$$\mathbf{S}_3^3 : \quad f_2\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.83$$
$$\mathbf{S}_3^4 : \quad f_2\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.87$$
$$\mathbf{S}_3^5 : \quad f_2\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.9$$
$$\mathbf{S}_3^6 : \quad f_2\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.73\ .$$

Record combination **ac**, for which the value of f_2 is minimal.

4´. Check for the candidate expanded combinations of the recorded string **ac** with each of the five operations in $\mathbf{S}_1$. There are 6 such combinations:

$$\mathbf{aac} \leftrightarrow \theta \qquad \mathbf{acb} \leftrightarrow \theta \qquad \mathbf{bac} \leftrightarrow \theta \qquad \mathbf{acc} \leftrightarrow \theta \qquad \mathbf{acd} \leftrightarrow \theta \qquad \mathbf{abac} \leftrightarrow \theta\ .$$

We obtain six expanded new sets of operations (each containing 6 operations): $\mathbf{S}_4^1$, $\mathbf{S}_4^2$, $\mathbf{S}_4^3$, $\mathbf{S}_4^4$, $\mathbf{S}_4^5$, $\mathbf{S}_4^6$. Again, as above, we begin to compute

$$\mathbf{S}_4^1 : \quad f_2\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.$$

Since the value of f_2 is zero, compute the value of f_1 for the corresponding weight vector:

$$\mathbf{S}_4^1 : \quad f_1\,(0.25, 0.25, 0.25, 0.25, 0, 0) = 0.25\ .$$

Since we found the set of operations and the weight vector ω* such that

$$f_2\,(\omega^*) = 0 \qquad\qquad f_1\,(\omega^*) \neq 0\ ,$$

the algorithm terminates with the latter solution. In other words, the final set of operations, $\mathbf{S}_2$, is:

S_1: **a** ↔ θ S_2: **b** ↔ θ S_3: **c** ↔ θ S_4: **d** ↔ θ
S_5: **ab** ↔ θ S_6: **aac** ↔ θ

and the optimum weighting scheme is (0.25, 0.25, 0.25, 0.25, 0, 0). In other words, the set F of strings, which was used to generate the learning class, is {**ab**, **aac**}.

This means, in particular, that as the result of inductive learning *the structure of the class* represented to the system by the training set (L^+, L^-) can be encoded by the following 3-tuple

$$(L^+\,,\ \mathbf{S}_2\,,\ \omega^*). \tag{13}$$

Thus the system can now *exactly* identify the elements of the language (concept) L in the environment ***L***:

$$o \in L \quad \text{if and only if} \quad \Delta_{\omega^*}(o, L^+) = 0. \qquad (14)$$

This scheme for concept learning, as mentioned in section 5.2, turns out to be a formal realization of the exemplar theory of concept learning.

It is also interesting to note that the above learning process is consistent with the following suggestion of Johnson-Laird [25, p. 143]:

> One way in which to think of learning is as a procedure that discovers a way in which to combine old functions so as to create new ones. The system starts with an inborn set of functions and an inborn set of procedures for combining them, and learning consists in constructing functions that the system did not hitherto possess out of the function that it did possess.

Extensive experiments conducted by us for much larger problems in ***L*** also demonstrated that the presence of a relatively substantial "noise" in L^+ does not prevent the inductive learning process to discover the *exact* structure of the concept [34].

6. On the role of symbols and numbers and their relationship in the proposed model

Which object representation (symbolic or numeric) would allow the agent to capture the maximum possible structural information about the concept being learned? By "symbol" we mean some sign that stands for a *physical* atomic feature. We will call it p-symbol, or simply symbol. By "number" we mean an element of a set that satisfies the axioms for real numbers [6]. There are infinitely many ways to represent a set of numbers (Arabic numerals, Roman numerals, binary numerals, etc.). The critical difference between p-symbols and numbers arises from the fact that numbers satisfy a fixed set of axioms - e.g. those for addition, multiplication, order - while p-symbols do not have any such predefined relationships. Generalization of real numbers to vector spaces does not essentially change the above situation (for axiomatic definition of vector spaces see, for example, [21]). Thus, the symbolic representation allows much greater "freedom" in capturing the concept structure and structural dependencies than the numeric representation.

Jean Piaget stressed the important role of operations on objects:

> In order to know objects, the subject must act upon them and therefore transform them [both physically and mentally]: he must displace, connect, combine, take apart, and reassemble them. … Knowledge is constantly linked with actions or operations, that is, with *transformations*. [36, p.704]

The operations that can be carried out on p-symbols are quite different from the operations that can be carried out on numbers or vectors. For example, new symbols can be created from the old ones by concatenating them. Insertion, deletion, and substitution of symbols and symbolic structures are the fundamental operations that can be carried out on symbols. The algebraic operations of addition and multiplication are not defined on symbols. For example, symbol **nose** can be concatenated with symbol **mouth** to form symbol **nose and mouth** that conveys sensible information in the real world. However, **nose** and **mouth** cannot be multiplied or added to produce some third sensible symbol. The algebraic operations are defined on numbers and vectors, but the insertion, deletion and substitution operations are not defined on numbers. Moreover, by definition, p-symbols denote some physical entities, while numbers do not.

Thus it should be clear that p-symbols and numbers play very different roles in all models, and it is important to keep in mind the issue of their *legitimate use*. The main consequences of the above to an inductive learning model are related to the fact that the symbolic and numeric "geometries" are fundamentally different, since these "geometries" are defined in terms of the corresponding sets of operations. In connection with the latter, we note that the treatment of symbolic representations in connectionist models is absolutely inappropriate.

On the other hand, due to the axiomatic restrictions imposed on them, numeric models, including vector spaces, have very powerful analytical tools developed for them, while the symbolic models, in principle, cannot have such tools (see also section 3.1).

Returning to the model outlined in section 5, we note that to introduce semantics into the classical syntactic model, it uses numbers in two different ways:

(1) Each operation is assigned a real weight such that the sum of weights is equal to 1. The weights capture the interrelationship between the operations (symbolic structures). The weight space (vector space) is separate from the object space. Their fusion, however, leads to the creation of *emergent properties* of the set of operations, both mathematical and "physical" (see also section 5.1).

(2) Since the purpose of carrying out operations on objects is to discover their similarities (in order to produce inductive generalization), the dissimilarity, or distance measure is a very natural way to evaluate the effect of an operation. New operations are selected and analyzed on the basis of the function f_2, the average dissimilarity of positive examples. The concept of distance appears to be the only way to introduce the power of "abstract geometry" into a symbolic setting. Geometry, of course, offers a universal and most powerful way to guide the search.

To evaluate new (constructed) operations, the model relies only on "the relative" values of function f_2, which makes the learning process immune to random noise in the training set [34]. Since the random noise may effect any feature, the fact that the direction of the learning algorithm remains relatively stable is of considerable importance. As a consequence, the model can extract exact class description even in the presence of noise.

To summarize, one can say that the proposed inductive learning model offers a fundamentally new and very natural symbiosis of the symbolic and the numeric paradigms.

7. Conclusion

We believe that the inductive learning process is the most fundamental part of intelligent processes, since it is the only process that allows the agent to capture and encode in a compact form inductive generalization of a class, an object, or an event. Once constructed, this generalization may be applicable to an infinite class of objects or events. Without such mechanisms, the agent cannot succeed, or even survive, in a reasonably complex dynamic environment.

We agree with the tacit AI claim about the critical role of the symbolic representation of objects or events in the environment. This role is, essentially, due to a much greater universality and flexibility (power) of the inductive learning (generalization) mechanisms based on the symbolic representation as compared to those based on the numeric representation.

Inductive learning process would then involve discovery and encoding of the symbolic structures (features) and their interrelationships that make the positive training objects similar to each other and, at the same time, different from the negative training objects. Mathematically, the inductive learning process involves synthesis of the distance measure that captures the structure of the class.

As was already pointed out by von Neumann (section 3.1), a purely symbolic (discrete) mathematical model lacks the analytical power offered by the classical (continuous) mathematical models. Although purely symbolic models, in particular the present AI models, can *heuristically construct* some symbolic object features, they do not have sufficiently powerful tools to efficiently analyze and utilize the extracted information. Thus, Newell and Simon's physical symbol system hypothesis is incomplete, simply because the classical discrete models are not adequate for modelling of inductive learning process.

At the same time, as was noted by Searle ([45], [44]) and others [24, p. 334], a purely symbolic model cannot have semantics, simply because it requires another (symbolic) system to guide the corresponding process. It should be clear from the example in section 5.4 (and those in [34]) that only due to its numeric "side" does the proposed inductive learning model - which is a symbiosis of the discrete and the continuous - gains access to the semantic information.

The concept of distance measure is the key concept of the new model. The resulting new mathematical structure - the (evolving) transformation system - is a very natural unification, or symbiosis, of the symbolic and the numeric paradigms. Within the proposed model, to understand the objects (or events) the learning mechanism performs operations on the objects and (simultaneously) analyses the effect of the operations by means of the corresponding numeric information: the effect of any symbolic operation is measured by means of the change it introduces in the geometric configuration of the training set (which, in turn, is measured by functions f_2 and f_1). Thus, the concept of distance appears to be the only candidate for the basic tool that can guide the inductive learning process.

Returning to the issue of semantics, it appears that the symbiosis of symbols and numbers, as suggested by the evolving transformation system (ETS), breaks the "wall between the syntax and semantics. It was this wall, present in the current symbolic AI models, that prompted Searle to conclude that AI was impossible" [10].

On the other hand, being purely numeric frameworks, vector space based learning models, including connectionist models, cannot discover new symbolic structures. Thus, according to our formulation of the inductive learning problem in section 2, connectionist models, in principle, cannot carry out reliable generalization from the training set. However, once the learning (generalization) stage is completed by the ETS, and the vector representation of the training set is constructed ([20], [16]), the neural network may be used as a possible tool for on-line recognition.

The ETS appears to be the first example of an irreversible mathematical model: as the system evolves with time, it changes in an irreversible manner, i.e., the perception of the environment differs from one stage to the next one.

Last but not the least, as a consequence of the symbiotic structure of the new model, it becomes clear how two fundamentally different types of mechanisms - the numeric recognition mechanisms and the symbolic propositional mechanisms - can function *on top* of the inductive learning model. Moreover, since the previous history of cognitive science, AI, and pattern recognition amply testifies that without such inductive learning model no substantial progress in modelling the intelligent processes is possible, we believe that the new model should be seriously considered as the foundation for cognitive science, AI, and pattern recognition.

References

1. BOURNE, L. E., DOMINOWSKI, R. L., LOFTUS, E. F., and HEALY, A. F. (1986). *Cognitive Processes*, second edition, Prentice-Hall, Englewood Cliffs, New Jersey.

2. BRIDGEMAN, B. (1988). *The Biology of Behavior and Mind*, Wiley, New York .

3. BUNKE, H., and SANFELIU, A. (eds.) (1990). *Syntactic and Structural Pattern Recognition Theory and Applications*, Word Scientific, Singapore.

4. CARELLO, C., TURVEY, M. T., KUGLER, P. N., and SHAW, R. E. (1984). Inadequacies of the computer metaphors, in GAZZANIGA, M.S. (ed.), *Handbook of Cognitive Neuroscience*, Plenum Press, New York.

5. CHAN, T. Y. T. (1991). *Learning As Optimization: An Overture*, Ph.D. dissertation, Faculty of Computer Science, Univ. of New Brunswick, Fredericton.

6. DIEUDONNE, J. (1960). *Foundations of Modern Analysis*, Academic Press, New York.

7. FEYNMAN, R. (1967). *The Character of Physical Law*, MIT Press, Cambridge, Massachusetts.

8. FU, K. S. (1982). *Syntactic Pattern Recognition and Applications*, Prentice-Hall, Englewood Cliffs, New Jersey.

9. GODEMENT, R. (1968). *Algebra*, Houghton Mifflin, Boston.

10. GOLDFARB, L. (1992). A cognitive theory without inductive learning, *Behavioral and Brain Sciences,* **15**, 446 - 447.

11. GOLDFARB, L. (1992). What is distance and why do we need the metric model for pattern learning, *Pattern Recognition,* **25**, 431 - 438.

12. GOLDFARB, L. (1991). Varifiable characterization of an intelligent process, in GOLDFARB, L., and NICKERSON, B.G. (eds.), *Proc. 4th UNB Artificial Intelligence Symposium*, 67 - 80.

13. GOLDFARB, L. (1990). A working characterization of intelligence and a new model, Technical Report TR90-062, Faculty of Computer Science, Univ. of New Brunswick, Fredericton.

14. GOLDFARB, L. (1990). Why do we need the auxiliary vector representation for the metric pattern recognition problem? Technical Report TR90-056, Faculty of Computer Science, Univ. of New Brunswick, Fredericton.

15. GOLDFARB, L. (1990). On the foundations of intelligent processes I: An evolving model for pattern learning, *Pattern Recognition,* **23**, 595 - 616.

16. GOLDFARB, L. (1985). A new approach to pattern recognition, in KANAL, L.N., and ROSENFELD, A. (eds.), *Progress in Pattern Recognition 2*, North-Holland, Amsterdam.

17. GOLDFARB, L. (1984). A unified approach to pattern recognition, *Pattern Recognition,* **17**, 575 - 582.

18. GOLDFARB, L. (1979). *Riemannian Representation of Finite Symmetric Spaces and Their Application to General Data Analysis*, Ph.D. dissertation, Department of Systems Design, Univ. of Waterloo.

19. GOLDFARB, L., and NIGAM, S., Inductive learning, generalization, symbols, and the resolution of the Searle's paradox (in preparation).

20. GOLDFARB, L., and NIGAM, S., On the generalization capabilities of the classical vector space learning techniques including neural networks (in preparation).

21. HALMOS, P. (1958). *Finite-dimensional Vector Spaces*, Van Nostrand, Princeton.

22. HARNAD, S. (1992). Connecting object to symbol in modelling cognition, in CLARKE, A. and LUTZ, R. (eds.), *Connectionism in Context*, Springer-Verlag, New York.

23. HELMHOLTZ, VON H. (1971). The origin and correct interpretation of our sense impressions, in KAHL, R. (ed.) *Selected Writings of Hermann von Helmholtz*, Wesleyan University Press.

24. JOHNSON-LAIRD, P. N. (1988). *The Computer and the Mind*, Harvard University Press, Cambridge, Massachusetts.

25. JOHNSON-LAIRD, P. N. (1983). *Mental Models*, Harvard University Press, Cambridge, Massachusetts.

26. KOLMOGOROV, A. N. (1984). Automata and Life, in MAKAROV, I. M. (ed.) *Cybernetics Today: Achievements, Challenges, Prospects*, Mir Publishers, Moscow.

27. MADDOX, I. J. (1988). *Elements of Functional Analysis*, second edition, Cambridge University Press, Cambridge.

28. MASLOV, S. Y. (1987). *Theory of Deductive Systems and Its Applications*, MIT Press, Cambridge, Massachusetts.

29. MICLET, L. (1986). *Structural Methods in Pattern Recognition*, Springer-Verlag, New York.

30. NEUMANN, VON J. (1970). The general and logical theory of automata, in PYLYSHYN, Z. W. (ed.), *Perspectives on the Computer Revolution*, Prentice-Hall, Englewood Cliffs, New Jersey.

31. NEUMANN, VON J. (1966). *Theory of Self-Reproducing Automata*, Univ. of Illinois Press, Urbana, Illinois.

32. NEWELL, A. (1992). Precis of *Unified theories of cognition, Behavioral and Brain Sciences*, **15**, 425 - 492.

33. NEWELL, A. and SIMON, H. A. (1976). Computer science as empirical enquiry: Symbols and search, *Communications of the ACM,* **19**, 113 - 126.

34. NIGAM, S. (1993). *Metric Model Based Generalization and the Generalization Capabilities of the Connectionist Models*, M.Sc. dissertation, Faculty of Computer Science, Univ. of New Brunswick, Fredericton.

35. PARTEE, H. H. (1974). Discrete and continuous processes in computers and brains, in CONRAD, M., GUTTINGER, W. and DAL, M. (eds.), *Physics and Mathematics of the Nervous System*, Springer-Verlag, New York.

36. PIAGET, J. (1970). Piaget's theory, in MUSSEN, P. H. (ed.), *Carmichael's Handbook of Child Psychology*, Wiley, New York.

37. PIATTELLI- PALMARINI, M. (ed.) (1980). *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, Harvard University Press, Cambridge, Massachusetts.

38. RASTRIGIN, P. A., and MARKOV, V. A. (1976). *Cybernetic Models of Cognition*, Zinatne, Riga (in Russian).

39. RICH, E., and KNIGHT, K. (1991). *Artificial Intelligence*, second edition, McGraw-Hill, New York.

40. RUSSEL, B. (1986). *Outline of Philosophy*, Unwin Paperbacks, London.

41. SANKOFF, D., and KRUSKAL, J. B. (eds.) (1984). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons*, Addison-Wesley, Reading, Massachusetts.

42. SANTOSO, W. B. (1992). *Learning Algorithm for the Reconfigurable Learning Machine*, M.Sc. dissertation, Faculty of Computer Science, Univ. of New Brunswick, Fredericton.

43. SATRIAWAN, A. (1990). A Tree Transformation System for Pattern Learning, M.Sc. dissertation, Faculty of Computer Science, Univ. of New Brunswick, Fredericton.

44. SEARLE, J. R. (1984). *Minds, Brains and Science*, Harvard University Press, Cambridge, Massachusetts.

45. SEARLE, J. R. (1980). Minds, brains and programs, *Behavioral and Brain Sciences,* **3**, 417 - 457.

46. THOMPSON, R. F. (1985). *The Brain: An Introduction to Neuroscience*, Freeman, New York.

47. TRVERSKY, A. (1977). Features of similarity, *Psychological Review,* **84**, 327 - 352.

48. WATANABE, S. (1985). *Pattern Recognition: Human and Mechanical*, Wiley, New York.

49. WINSTON, P. H. (1992). *Artificial Intelligence*, third edition, Addison-Wesley, Reading, Massachusetts.

50. YOURGRAU, W., and MANDELSTAM, S. (1979). *Variational Principles in Dynamics and Quantum Theory*, Dover, New York.