Department of Computer Methods
Nicholas Copernicus University
Grudziądzka 5, 87-100 Toruń, Poland

COMPLEX SYSTEMS, INFORMATION THEORY AND NEURAL NETWORKS

Abstract

Włodzisław Duch and Norbert Jankowski[1]

In this paper relation between complex systems, information theory and the simplest models of neural networks are elucidated. Two different types of complex systems are distinguished, new complexity measure based on the graph theory defined, hierarchy of the correlation matrices introduced and connection with the correlation matrix memories and other types of neural models explained.

1 Two Types of Complex Systems

Complex system theory is a new field of science that emerged at the end of the last decade. A working definition of complex systems was given at the 1989 conference devoted to this subject [1]:

> ... systems that exhibit complicated behavior but for which there is some hope that the underlying structure is simple in the sense of being governed by a small number of degrees of freedom.

Another working definition is [2]

> A system is loosely defined as complex if it is composed of a large number of elements, interacting with each other, and the emergent global dynamics is qualitatively different from the dynamics of each one of the parts.

Examples of complex systems include fractals, snow flakes, cellular automata, games, Ising and spin glass models, artificial neural nets and many dynamical systems that exhibit complex behavior starting from simple dynamics. Other complex systems and objects, such as language (structure of words and sentences), proteins, genes, visual data, market analysis data, do not fit to these definitions. We do not even know if a simple dynamics responsible for their complexity exist and, in case of many-body systems such as proteins, we are convinced that a large number of degrees of freedom is necessary for their description.

It is useful to differentiate between two kinds of complex systems [3]:

Complex systems of the first kind, with known simple dynamics but unknown complex behavior. In this case we aim at the analysis and classification of possible behavior.

Complex systems of the second kind, or essentially complex systems, with unknown dynamics and partially known complex behavior. In this case our goal is to simplify description of these systems and, if possible, to find the dynamics.

[1] duch@phys.uni.torun.pl, norbert@phys.uni.torun.pl

So far only systems of the first kind have been considered in the literature, in particular the chaos theory deals with such systems. Molecular complex structures of the second kind are too small and irregular for statistical mechanics and too large for fundamental theories to tackle. Another example of complex system not covered by the quoted definition is the structure of natural language. Words and sentences have some regularity but it is very hard to find the "deep grammatical structure" that will allow us to parse complex sentences. Vocabularies are complex, relatively large systems of information hard to analyze via mathematical means and apparently without simple underlying mechanism that could generate them. Problem solving in artificial intelligence leads to the representation spaces and decision trees that show combinatorial explosion, thus leading to complex behavior or complex structure of their solution spaces.

Of course it may be that only trivially complex systems of the first kind exist in nature and the essentially complex systems of the second kind are just artificial constructions of the human mind. Nevertheless, at the present stage of scientific inquiry it seems appropriate to develop also an approach that should allow for characterization of complex systems of the second kind, where the dynamics is completely unknown or govern by too many degrees of freedom to handle it explicitly. The goal of such theory would be to simplify the description of complex systems by finding a series of simpler descriptions, approximations converging at the full complexity. One source of inspiration for such theory comes from the theory of information, devised by Shannon [4] and others to measure the amount of information in an arbitrary data system. Another approach is offered by statistics, in particular statistical theories of language. The latest approach comes from the distributed storage of patterns in simplified neural networks.

Some interesting connections of complex systems of the second kind with information theory, statistical approach and simplest models of neural networks are described below.

2 Information And Complexity Measures.

More than 40 years after the definition of information appeared in the landmark paper of Claude Shannon [4] we still do not have a satisfactory definition of information that would be in accord with our intuition and that could unambiguously be applied to such concepts as biological information or linguistic information. Different definitions or measures of information exist now, including axiomatic definition of Shannon information, algorithmic information, pragmatic information and cyclomatic information (for details see [5]).

The simplest approach to the quantitative definition of information is based on combinatorics [5]. Interesting applications of combinatorial information to the estimation of "entropy of a language" have been reported (Kolmogorov 1968). The entropy of words in a dictionary is considerably higher than the entropy of words in a literary text, indicating that there are some constrains (grammatical and stylistic) in literary texts.

The second approach is based on probability. It was introduced in the theory of information transmission by C. Shannon (1949) and is based on his formula

$$I_P = -\sum_i p_i \log p_i$$

where p_i is the probability of item i of the data. Shannon considered a question: what is the minimal number of bits needed to transmit data?

Another measure of information, called algorithmic or Chaitin-Kolmogorov information is in use in computer science (Kolmogorov 1965,1968, Chaitin 1966,1990).

Algorithmic information or the relative complexity of an object y with a given object x is defined as the minimal length of the program p for obtaining y from x. Algorithmic information captures some intuitive features of information: a binary string obtained by truly random process cannot be compressed and carries the amount of information equal to the number of its digits. Algorithmic complexity has found interesting applications in theoretical computer science to estimate the number of steps necessary to solve certain classes of mathematical problems.

Shannon I_P measures the number of bits per item needed to code this item. For example, in English texts 26 letters and a blank space is used; for equiprobable letters $I_0 = \log_2 27 = 4.75$ bits per character, but the real amount of information is lower due to interletter correlations. Taking real probabilities for English texts Shannon obtained $I_P = I_1 = 4.03$ bits. Second order information includes correlations between pairs of letters and is computed from the formula:

$$I_2 = I_{\alpha_1}(\alpha_2) = I(\alpha_1\alpha_2) - I_{\alpha_1} = -\sum_{l_2=a}^{z}\left(\sum_{l_1=a}^{z} p(l_1 l_2)\log_2 p(l_1 l_2) - p(l_2)\log_2 p(l_2)\right)$$

One may define higher order correlations and the infinite order limit. Redundancy is defined as $R = 1 - I_\infty/I_0$. For English language estimated redundancy is around 80%.

Shannon information is useful for estimation of data transmission and efficiency of data compression, but it does not estimate data complexity. Algorithmic information I_A of a random set of binary strings with N bits is of the order of N. Algorithmic information for all possible binary strings or other well-structured sets of strings is small. Although more intuitive and useful in complexity theory then the Shannon definition the concept of algorithmic information has problems:

1. Algorithmic information is hard to compute.

2. All "iterative structures" like fractals or cellular automata are equivalent, even some of these structures are obviously much more complex than the others.

Minimal graph complexity measure I_G [5] defines complexity of a given data structure to be equal to the number of arcs in the minimal graph that contains all the data. Let us take all n-digit binary string as an example. For $n = 5$ all binary strings, from 00000 to 11111, are contained in the minimal graph with 10 arcs. In general minimal graph for n-digit strings has $2n$ arcs, with one string removed $4n-3$ arcs (see Fig. 1). Minimal graph represents a set of data items.

I_G has properties of *pragmatic information* defined in a qualitative way by von Weizsäcker [6]: it grows quickly when novel information is given and it shrinks when the information confirms general pattern; for repeated data it does not change. It is similar to the algorithmic information but is easier to compute.

Minimal graph complexity I_G has applications for finite systems, such as lexicographical structures, proteins, genes, game theory. Semantic contents or meaning is relevant only if we have some cognitive system. Words and ideas have different meanings and different information contents for different people therefore it is not possible to give a universal definition. The meaning of the same information is different for different people because their internal representation of the world is different. We must refer to some representation of the world to define semantic information. Such representation may be based on a set of rules stored in a knowledge base of an expert system. This knowledge base, together with the rules of inference, define our universum of facts, representing knowledge or some model of the world M.

There is an analogy between the theory of **complex systems** and **linguistics**. If the system is not completely chaotic we can find an alphabet, a list of substructures or elements of behavior, which, due to some interactions, generate complexity. Interactions in this case are analogous to grammatical rules.

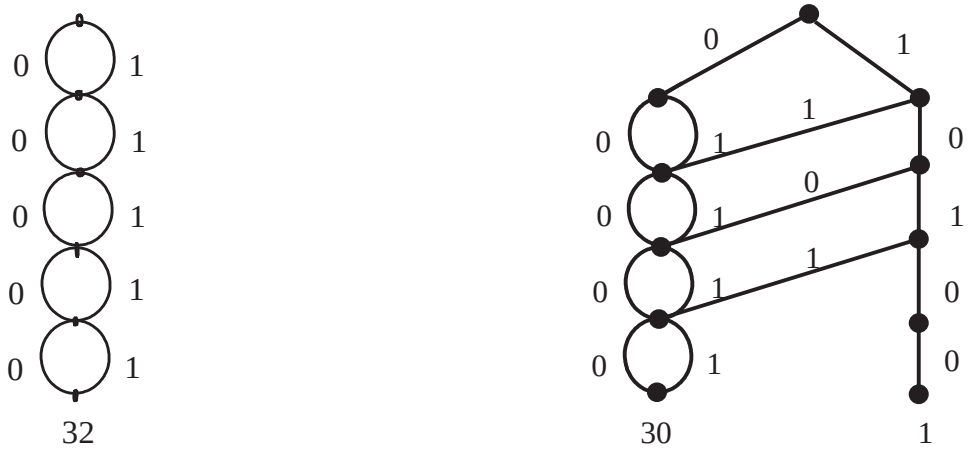

Figure 1: Example of completely folded graphs obtained from trees of binary strings for all 32 strings of 5 digits (left) and with one string, 10101, missing.

3 Hierarchy Of Correlation Matrices.

Calculation of the N-th order Shannon information requires knowledge of the N-th order probabilities $I_N \rightarrow p_N(l_1, l_2, \ldots, l_N)$. For lower order probabilities only statistical properties of the data will be captured, but for N large enough p_N should be equivalent to the original data. In particular a list of all non-zero probabilities for words of the length N is equivalent to the original list. The statistical approach to generation of music or generation of texts was considered, but the low order samples are very different from the real language. For example, third order sample, i.e. taking most probable (l_1, l_2, l_3) triples of letters for polish language, gives *okopomenta tyka wszcza spełniergi czniesz ach*. There is a systematic way of going from such low order statistical description to the original data.

To measure the information content of a lexicon we present first the list of words in a graphical form and than fold this graph into minimal graph.

$$\text{List of words} \rightarrow \text{Graph} \rightarrow \text{MinGraph}$$

The number of paths $N_p(MinGraph)$ in the minimal graph is equal to the number of paths $N_p(Graph)$ and to the number of words in the original list; each word may be recreated from this graph. The structure of the minimal graph reflects the structure present in the list of words. For example, minimal graph for the subset of polish language shows clearly the prefixes, roots and suffixes.

Instead of a minimal graph larger structure may be generated, with the same statistical properties as the original list. The list of words has a very complex structure. Is there any regularity in it? One approach to find it is to derive probabilities or statistical relations among the characters composing the words belonging to the list and than create a graph that stores all strings of characters with non-zero probability.

$$\text{List of words} \rightarrow p_N \rightarrow \text{Graph\_p} \rightarrow MinGraph\_p$$

Although the number of paths (words) in the Graph_p is much higher than in the original list $N_p(MinGraph)$ the minimal graph MinGraph_p itself is much smaller and simpler. It shows general structure of the complex system. Since 2-nd order $p_2(l_1, l_2)$ probabilities for adjacent letters give poor representation of real words in a dictionary and higher order $p_N(l_1, l_2, \ldots, l_N)$ lead to huge matrices we shall

take hierarchical approach and consider a supermatrix of second order probabilities $p(l_i, l_j)$ for $i > j$. Diagonal blocks of this matrix contain $p_2(l_i, l_i) = p_1(l_i)$, probabilities of single characters. The simplest and the least accurate approach is to take $p_2(l_i, l_i + 1)$; next step is to add $p_2(l_i, l_i + 2)$, a second sub-block of the supermatrix.

Another way to increase the accuracy is to include partially higher order correlations by changing the representation of each letter, making it "sensitive to the environment", i.e. to other letters adjacent to it. Selecting new characters as $L_i = (l_i, l_i + 1)$ full pair-pair correlations $p_2(L_i, L_i + 1)$, equivalent to the full third order correlations are obtained. Smaller number of new characters L_i may be chosen by mapping different pairs or triples of characters into a single L. We can measure how effective is each step in bringing us closer to the exact description of the complex system by looking at the number of paths in the Graph_p; we can measure how much information is gained by counting the arcs in the MinGraph_p. A whole hierarchy of correlation matrices $p(L_i, L_j)$ is created in this way, describing complex system with increasing accuracy.

4 Connection With Neural Networks.

There is a unique correspondence between neural networks, correlation matrices of statistical models and graphs, with weights W of the network connections equal to the appropriate probabilities p. Statistical model based on pairs of adjacent letters $p_2(l_1, l_2)$ may be represented as a two-layered net with N_l (equal to the number of letters) inputs and outputs; second order model with $p_2(l_i, l_i + 1)$ is realized by the two-layered net with non-zero blocks near diagonal; full 2-nd order statistical model, with probabilities for all pairs of letters $p_2(l_i, l_j)$ corresponds to the fully connected two-layered network. Higher order statistical models correspond to networks with the intermediate layers of neurons.

This model is a generalization of Kohonen's CMM (Correlation Matrix Memory) [7]. Data (question) vectors x and answer vectors y are memorized in a correlation matrix

$$W_{xy} = \sum_k x_{(k)} y^T_{(k)}$$

If y vectors are orthogonal than

$$x_{(i)} = W_{xy} y(i) = \sum_k x_{(k)} y^T_{(k)} y_{(i)}$$

As an example of application of this approach to the real data the second-order model was used for dictionary of polish words. There are 35 characters in polish alphabet. Out of $35^3 = 42875$ possible combinations of 3 characters 624 words are found in the dictionary, from *abo* to *źli*. The network is composed from 2 layers, with 335=105 units in each row, i.e. each word is represented by 105 bits and is not orthogonalized. Weight matrix W_1 for correlation between (1,2), (2,3) and (1,3) letters has 3675 entries, replaced here by binary (0 or 1) values. Graph_p, generated by this network (i.e. from these correlation matrices), has 1692 paths, 1068 corresponding to wrong combinations of 3 letters (2.5% of errors) and 624 to the right combinations, i.e. to the words in the original list.

The third order model, with W_2 matrices for non-zero $p_3(l_1, l_2, l_3)$, corresponds to the correlations between $l_1 \rightarrow (l_2, l_3)$ and is equivalent to the original data. Another way of obtaining perfect representation of data by networks is to use a few second order nets, separating the words into orthogonal sets. In this case 6 networks or W_1 matrices are required.

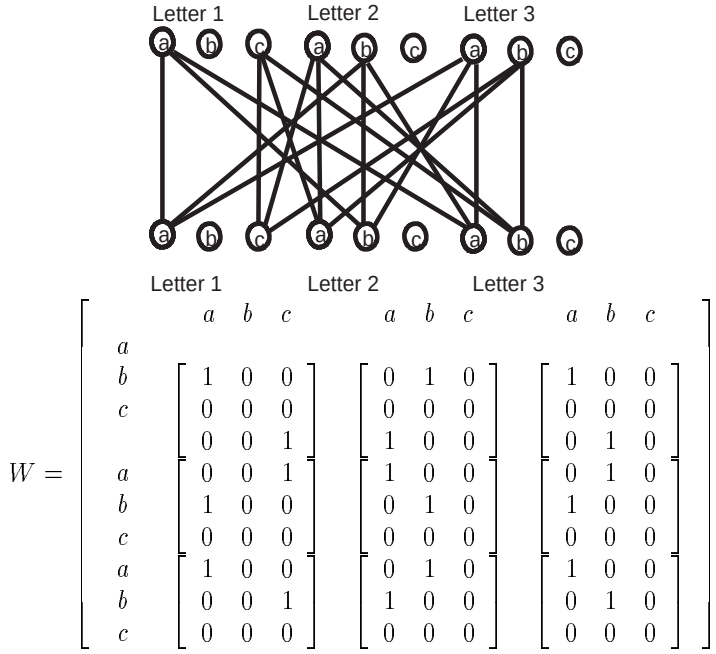

Figure 2: Fragment of a 2-layered neural net with inputs for letters at a given position and non-zero connections created for two words: *aba* and *cab*. Binary matrix W shows non-zero connections between input and output neurons.

References

[1] E. Jen. Lectures in complex systems. Addison–Weslay Publishing Corporation, 1990.

[2] M. S. Garrido and R. V. Mendes. Complexity in physics and technology. *World Scientific*, Singapore 1993.

[3] Włodzisław Duch. Towards statistical theory of complex systems. Technical report, Department of Computer Methods, 1992.

[4] C. E. Shannon. A mathematical theory of communication. (27):379–423, 623–656, 1948.

[5] Włodzisław Duch. Syntactic and semantic information in finite systems. In *Open Systems and Information Dynamics (submitted).*

[6] E. von Weizsäcker. *Erstmaligkeit und Bestätigung als komponenten der pragmatischen Information*, pages 82–113. Ernest Klett Verlag, Stutgart 1974.

[7] T. Kohonen. *Self-organization and associative memory.* Springer–Verlag, Heidelberg, 1984.

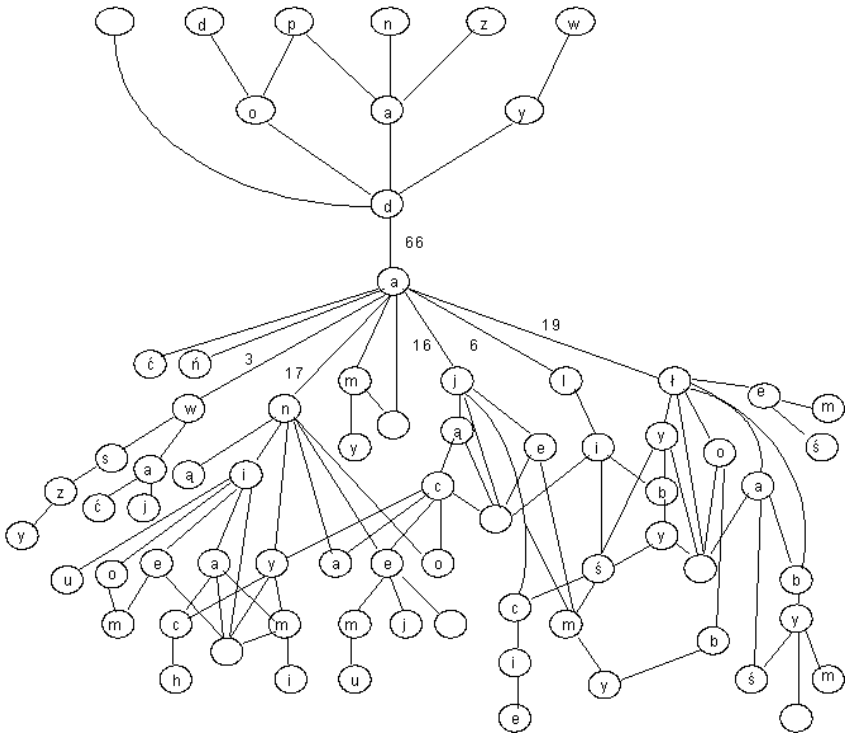

Figure 3: Fig. 3. An example of a minimal graph for a subset of Polish language. The root is *da*, there are 7 prefixes and 65 suffixes. Very high compression of data is achieved.