ISSN 1407-7493 **DATORZINĀTNE**
COMPUTER SCIENCE **2009-7493**

INFORMATION TECHNOLOGY AND MANAGEMENT SCIENCE
INFORMĀCIJAS TEHNOLOĢIJA UN VADĪBAS ZINĀTNE

PERCEPTRON ARCHITECTURE ENSURING PATTERN DESCRIPTION COMPACTNESS

TĒLU APRAKSTU KOMPAKTUMU NODROŠINOŠĀ PERCEPTRONA ARHITEKTŪRA

Sergejs Jakovlevs is currently PhD researcher at the Department of Modelling and Simulation of Riga Technical University. He received his master's degree in Information Technology from Riga Technical University in 2004. His research interests include artificial neural networks and bioinformatics.

Keywords: perceptron, pattern recognition

Abstract - This paper examines conditions a neural network has to meet in order to ensure the formation of a space of features satisfying the compactness hypothesis. The formulation of compactness hypothesis is defined in more detail as applied to neural networks. It is shown that despite the fact that the first layer of connections is formed randomly, the presence of more than 30 elements in the middle network layer guarantees a 100% probability that the G-matrix of the perceptron will not be special. It means that under additional mathematical calculations made by Rosenblatt, the perceptron will with guaranty form a space of features that could be then linearly divided. Indeed, Cover's theorem only says that separation probability increases when the initial space is transformed into a higher dimensional space in the non-linear case. It however does not point when this probability is 100%. In the Rosenblatt's perceptron, the non-linear transformation is carried out in the first layer which is generated randomly. The paper provides practical conditions under which the probability is very close to 100%. For comparison, in the Rumelhart's multilayer perceptron this kind of analysis is not performed.

Introduction

The hypothesis of compactness frequently employed in artificial neural networks (ANN) states that "in the used space of features, object representations belonging to one and the same class are close (have small distances) but those belonging to different classes are well separable from each other". Such a formulation seems to be too general to characterize the space of features with regard to the tasks being solved. First, it is not clear what it means to be well separable; due to that, in what follows we will consider it linearly separable. Secondly, from the hypothesis of compactness it is not clear which primary recognition tasks were stated. It is apparent that they are then reduced to a classification task but for analysis purposes we need to know how the task of recognition was formulated.

Bongard [1] wrote: "Sometimes a system that makes classification using a certain constant principle is called a recognizing device. A system of this kind can actually solve a single task quite well, for example, a cash machine checks the sizes, weight and material the inserted coin is made of, and recognizes the coin the machine is designed for". ANNs differ from that cash machine in that they can be trained to discover classification principle for diverse tasks but not for simultaneously different ones. That means that if we need to solve a second classification task not related to the first one, we have to take another network and train it independently for that second task. Instead, if we wish that classification is made by one network both for the first and the second task, we have to guarantee a compatible formulation of these two tasks. This is not as trivial as might appear, since the task being solved might contain two incompatible subtasks.

Even when we talk about these tasks independently, except for the first task, there is no sense to speak about the compactness hypothesis. This is due to the mentioned tasks are not reduced to classification task but are multi compound and relate to the problem of invariant and abstract representation of the data. At the moment being, this problem cannot be solved in general form.

Thus the formulation of compactness hypothesis as applied to ANNs is acquiring a more precise sense: "in the space of features obtained with the help of ANNs objects belonging to the same class of non-variant and non-abstract patterns are close to each other but measurements belonging to different such classes are linearly separable from each other".

To prove that such a space of features does not coincide with the space of initial data, the famous XOR problem can be used in which the space of initial data is linearly non-separable. Also, it is sometimes erroneously assumed that the presence of the hidden unit in the ANN in which a space of features is formed on the basis of the space of initial data, automatically meets the compactness hypothesis (even in the above clarified sense). It can be shown that at sufficiently high thresholds in elements of the hidden layer (A-elements), they simply stop sending signals to the output elements (R-elements) and vice versa, under low thresholds and high completeness of the input elements (S-elements) the signals are becoming so intensive that stimuli cannot be distinguished anymore. Besides, certain combinations of the weight coefficients are possible at

which it is impossible to construct a space of features satisfying the compactness hypothesis.

To form a space of features satisfying the compactness hypothesis (at least in the above mentioned narrow sense), a neural network has to meet certain conditions. The presence of two layers of connections is only a necessary not a sufficient condition. This paper describes conditions that are sufficient in a practical sense. It will help understand basic minimal requirements a neural network must satisfy even when it has a more complicated architecture necessary for solving more complicated tasks.

Rosenblatt's Perceptron

Rosenblatt's perceptron [2, 3] consists of three types of elements (see Fig. 1): S-elements, A-elements and R-elements (for simplicity, a case is considered below when there is only R-element). S-elements are a layer of receptors. These receptors are connected with A-elements by means of excitatory and inhibitory connections. Each receptor can be in one of two states - activated or deactivated. A-elements are summators with a threshold, i.e. formal neurons. It means that A-element is activated if the algebraic sum of activations coming to it from receptors exceeds a certain value, its threshold. The signals from activated A-elements are passed to summator R; the signal from the i-th associative element being passed with the weight coefficient v_i.

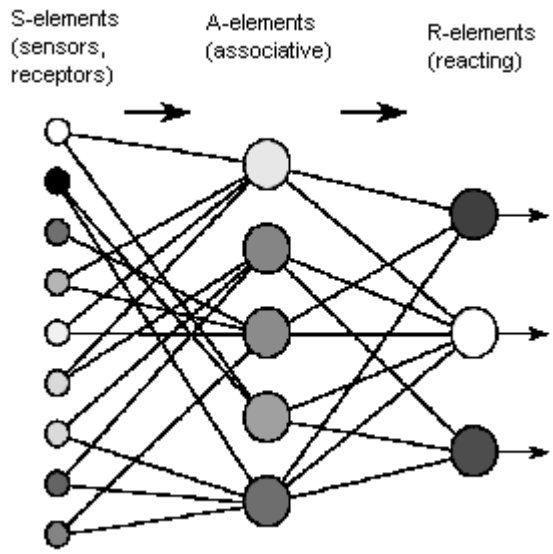

Fig. 1. Architecture of Rosenblatt's perceptron

The system of connections between receptors of S- and A-elements as well as the thresholds for A-elements are chosen in a certain random though fixed way, and training consists in changing coefficients v_i .

Say, we wish to train the perceptron to separate two classes of objects and we require the output to be positive when the objects of the first class are shown and to be negative when the objects of the second class are shown. The initial coefficients v_i are assumed to be equal to zero. Then a training set, i.e., objects (say, circles or squares) are shown, with indication to which class they belong. The perceptron is shown an object of the first class. With this, some A-elements will become activated. Coefficients v_i , corresponding to these activated elements will be increased by 1. Then an object of the second class is shown; coefficients v_i of those A-elements that become activated during that show are reduced by 1. The process is continued for the whole training set. As a result of learning, the values of weights of connections v_i are obtained.

As soon as the perceptron is trained, it is ready for working in the recognition mode. In this mode, the perceptron is shown unknown objects; it has to determine to which class they belong. The procedure is as follows: when an object is shown, the activated A-elements pass to the R-element a signal equal to the sum of the corresponding coefficients v_i. If the sum is positive, a decision is made that this object belongs to the first class, otherwise the object is ascribed to the second class.

Reasoning Made by Rosenblatt

Let us first discuss basic propositions proved by Rosenblatt in [2], which are necessary for further analysis. Following is the theorem proved by Rosenblatt:

«*Theorem 3.* Suppose there is given an elementary perceptron, a space of stimuli W and a classification $C(W)$. In order for the solution for $C(W)$ exist, it is necessary and sufficient that a vector u exists that lies in the same orthant as a vector x such that $Gx = u$ »

as well as two direct corollaries from that theorem:

"*Corollary 1.* Suppose an elementary perceptron and a space of stimuli W is given. Then if G is a special matrix (whose determinant is equal to zero), then there exists a certain classification for which there is no solution".

"*Corollary 2.* If the number of stimuli in space W is larger than the number of A-elements of an elementary perceptron, then a certain classification $C(W)$ exists for which there is no solution."

Rosenblatt calls a space of stimuli the space of initial data. Special sole in the theorem and its corollaries is played by the so-called G - matrix; it looks as follows:

$G = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1n} \\ g_{21} & g_{22} & \cdots & g_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{pmatrix}$, where n - number of stimuli, but g_{ij} - generalisation coefficient. The generalisation coefficient indicates a relative number of A-elements reacting both to stimulus St_i and stimulus St_j. For example, when solving an XOR task, G - matrix could look like this: $G = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$, from which it can be seen that there are three stimuli (rows-stimuli, columns – A-elements) and, for example, to stimulus St_1 and stimulus St_3 only one common A-element reacts, whereas separately to stimulus St_3 two A-elements react but to stimulus St_1 a single A-element reacts. Now it can be seen that generalisation coefficients g_{ij} are intersection measure for sets of A-elements reacting to stimuli St_i and St_j. With this, a G - matrix can be derived from a simpler A-matrix. They are connected by relationship $G = AA^T$ (where A^T is the transposed matrix). The A-matrix is of size $n \times N_a$, where n is the number of stimuli, N_a is the number of A-elements but its elements are the following: $a_{ij} = 1$, if A – element a_j reacts to stimulus St_i, and $a_{ij} = 0$ otherwise.

Keeping in mind that in the perceptron the weight coefficients are fixed in the first layer, we can see that the A-matrix does not vary over time. With this, it indicates which of A-elements will be active when the perceptron will be shown a certain stimulus. Say, for example, when solving the XOR problem the A-matrix may look like this: $A = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, from which is follows that at stimulus No.1 the third A-element is active; when stimulus No.2 is shown, the first and the third A-elements are active but at stimulus No.3 the second and the third A-elements are active.

For further analysis it will be also necessary to use the so-called Q - functions of the perceptron introduced by Rosenblatt. In the simpler case they indicate the probability that the A-element of the perceptron of the given class reacts to stimulus St_i. A more detailed explanation of these functions will be provided below as the material is given.

Conditions at which the Perceptron Forms a Space that Satisfies the Compactness Hypothesis

From Corollary 2 it can be seen that the number of A-elements has to be equal to the number of stimuli, i.e., the matrix must be square, but from Corollary 1 one can see that G - matrix must not be special. By satisfying these two requirements we get sufficient conditions for Rosenblatt's perceptron to form a space satisfying the compactness hypothesis.

But Rosenblatt has not conducted a complete analysis of what it can mean in practice. As a result, insufficiently validated assertion of Minsky [4] appeared stating that "the perceptron is only working perfectly if the set of initial data is linearly separable".

Theoretically, there are quite a lot of proved Rosenblatt theorems to disprove that assertion by Minsky and point out that the perceptron is able to work on any dataset. But, as the first layer of connections in the perceptron is chosen at random and is not trained, an opinion frequently arises that the perceptron with equal probability may both work and do not work under linearly non-separable initial data and this is linear initial data only that ensure its perfect performance. In other words, G - matrix of the perceptron may with equal probability be both special and non-special.

Below it will be shown that this opinion is wrong. Besides, conditions will be formulated that have to be met to ensure that the G - matrix is not special, which, in turn, proves helpful in analysing other ANN architectures.

Connection of G Matrix and A Matrix of Perceptron

Let us first pass from matrix G to matrix A (in what follows, the matrix A has the size $n \times n$ so as to satisfy Corollary 2), since it proves to be more convenient for further analysis:

1. Let matrix $G = AA^T$ be special, that is $|G| = 0$. Then $|G| = |AA^T| = |A| \times |A^T| = |A| \times |A| = |A|^2$, hence $|A|^2 = 0$; from this it follows that $|A| = 0$, i.e. matrix A is special.
2. Let matrix $G = AA^T$ not be special that is $|G| = \zeta \neq 0$. Then $|G| = |A|^2$ we arrive at $|A|^2 = \zeta \neq 0$, from this it follows that $|A| \neq 0$, i.e., matrix A is not special.

3. Let $|A| = 0$. Let us find $|G|$. We have $|G| = |A|^2 = 0 \cdot 0 = 0$, hence, matrix G is special.
4. Let $|A| = \zeta \neq 0$. Let us find $|G|$. $|G| = |A|^2 = \zeta \cdot \zeta = \zeta^2 \neq 0$, thus matrix G is not special.

So we have that matrix $G = AA^T$ is special if and only if matrix A is special.

Activity Probability of A-elements

From the definition of A-matrix it can be seen that it is binary: it assumes the meaning 1 when a corresponding A-element is active. We will be interested to know the probability of occurring of 1 and, respectively, the probability of occurring of 0 in the matrix. These events may not obligatorily be equal probable. As was mentioned above, for that purpose Rosenblatt studied the so-called Q-functions. Here we will only discuss the basic points.

Let us consider a binomial model of connections in the first layer. This is the case when there is the fixed number of connections from S -elements to each A-element. These connections consist of x excitatory connections (the weight +1) and y inhibitory (the weight -1). Threshold θ is fixed and is the same for all A – elements. The beginnings of connections to A-elements are selected irrespective of each other and are distributed with equal probability over the whole set of S - elements.

In this model Q - functions do not depend on the full number of sensor elements; they only depend on the relative number of illuminated S - elements. Fig. 2. shows how the probability of A-elements activity , Q_i depends on the size of the illuminated area of the retina. Note that for models that have only excitatory connections ($y = 0$), at a large count of illuminated S - elements the value Q_i is approaching 1. It means that the probability of A-elements activity directly depends on the number of illuminated S - elements, which badly influences recognition with regard to small and large images. As will be shown later, it occurs because of a large probability that A-matrix will be special. Due to that, to ensure pattern recognition stability for any geometrical sizes, relation $x : y$ has to be selected so that the probability of A-elements activity would possibly less depend on the number of illuminated S - elements.

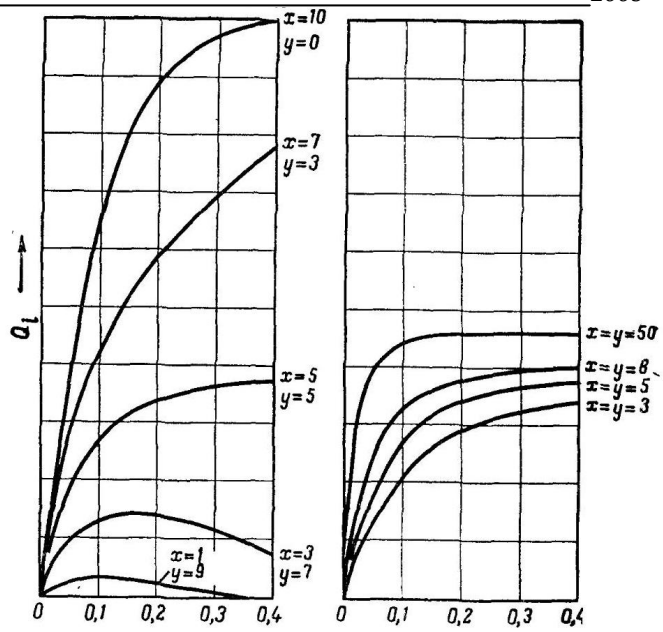

Fig. 2. Dependence of probability of A-elements activity. (a) impact of relation $x : y$ at $\theta = 1$; (b) impact of connection number variation at $x = y$ and $\theta = 1$ (borrowed from [2] Fig. 7).

From Fig. 2a it can be seen that at $x = y$, Q_i remains nearly constant over the whole region, except for a very small or very large number of illuminated S - elements, whereas Fig. 2b shows that if the total number of connections is enlarged, the region in which Q_i remains constant is becoming larger. At small θ and equal x and y Q_i is approaching the value 0.5, i.e., is equal to possible appearance of activity of A-element in reply to arbitrary stimuli. Thus conditions are found at which the appearance of unit and zero in the matrix A is equal probable.

In this case the number of 1s in the matrix A will be described by the binomial distribution:

$$\frac{l!}{k!(l-k)!} \Big/ 2^l , \qquad (1)$$

where l - the count of elements in the matrix, k - count of 1s in the matrix. In case if the appearance of 1s and 0s in the matrix A is not equal probable, the count of 1s in the matrix A will be described by Bernoulli's equation

$$\frac{l!}{k!(l-k)!} p^k (1-p)^{l-k} , \qquad (2)$$

where p - the probability of appearance of a 1.

Fig. 3 demonstrates the distribution for non-special matrices of size 5x5 depending on the count of 1s in the matrix.

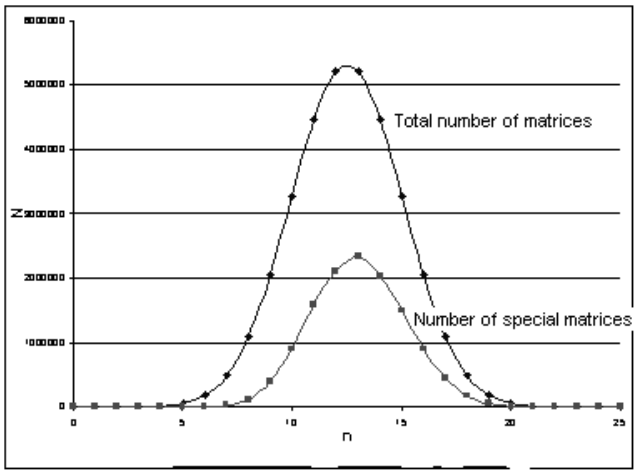

Fig. 3. Distribution of the count of matrices, N in dependence of the count of 1s, n, in the matrix of size 5x5

Probability that Matrix A is Special

From binomial distribution or Bernoulli distribution it is possible to learn how the total number of matrices with different number of 1s in each is distributed (see Fig. 3, the upper curve). However, for the count of non-special matrices (see Fig. 3, the lower curve) there is no mathematical formula by now yet.

Though, there is available the calculated sequence A002884 [5] that provides the lower boundary of the probability that matrix of size $n \times n$ will not be special (the determiner of these matrices is equal to 1). For cases $n > 30$ the probability is nearly stable and is going down practically insignificantly; it constitutes 28.8788 %. There is also a sequence A055165 [6] providing the probability that the matrix of size $n \times n$ is not special, though it can only be calculated by exhaustive search, which is made up to the cases $n \leq 8$.

The results of analysis of the case of equal probability of appearance of 1 and 0 in the matrix are given in Table 1. In its turn, Fig. 4 shows the dependence of the probability that the matrix is non-special, on its size. The first 8 values of (P) – are precise calculated data obtained by using a sequence: A055165. The next values are obtained using the Monte-Carlo method; more specifically, 1000 matrices are selected at random and the percentage of special matrices is then calculated out of them. Thus, 100 experiments (checks) are performed and the lower and the upper probability boundaries are derived.

Table 1

Dependence of probabilities of special matrices appearance on their size

n	Min, %	P, %	Max, %
2	32,5	37,5	42,1
3	31,1	33,9	38,2
4	31,3	34,4	38,8
5	33,9	37,2	40,2
6	38,2	41,9	45,2
7	43,9	48,0	52,9
8	51,2	55,0	59,1
9	59,2		65,7
10	66,4		74,2
11	73,8		79,8
12	80,8		86,9
13	86,0		91,3
14	90,5		94,9
15	93,5		97,2
16	95,9		98,3
17	97,5		99,1
18	98,3		99,8
19	99,0		99,9
20	99,4		100,0
21	99,5		100,0
22	99,7		100,0
23	99,8		100,0
24	99,8		100,0
25	99,9		100,0
26	99,9		100,0
27	99,9		100,0
28	100,0		100,0
29	100,0		100,0
30	100,0		100,0

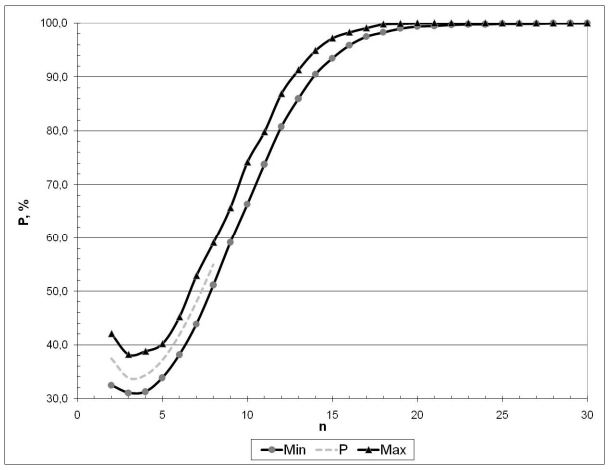

Fig. 4. The dependence of probabilities of special matrices appearance on their size

From Table 1 it can be seen that even at the matrix size $30x30$ one can speak about 100% confidence that occasionally obtained matrix will not be special. This is the final real condition at which Rosenblatt's perceptron will form a space satisfying the compactness hypothesis.

Conclusions

In 1965 Cover formulated his theorem known as Cover's theorem [7]: "The probability that classes are linearly separable increases when the features are nonlinearly mapped to a higher dimensional feature space". The theorem in essence describes the process taking place in the first layer of the perceptron. Yet, even before Cover has formulated his theorem, it was used to prove the corollary of Rosenblatt's Theorem 3 [2], which in fact proves the same but regarding the perceptron architecture. So it was Josef who first proved the theorem in 1960 [8, 9] but Rosenblatt employed it to specify perceptron characteristics. But it was not until 1965 that Cover independently formulated that theorem in the form known today.

One can find up-to-date applications of the theorem in several kinds of artificial neural networks, say in radial basis function (RBF) networks [10]. But taking into account that the perceptron has implemented the essence of that theorem long before these networks appeared, we can come to a conclusion that the RBF network is a particular case of the perceptron which only differs in a special activation function.

Thus the first conclusion we can draw is that the RBF network and some up-to-date ANNs are just subspecies of Rosenblatt's perceptron, which makes it possible to compare them not only analytically but also experimentally.

The second conclusion is that the analysis performed allows one to understand why in some cases researchers have drawn incorrect conclusions regarding perceptron's abilities. For example, in [11] it is pointed out that „experiments give evidence of the limited performance of the α-system of A-elements weight adaptation" as well as „the lack of convergence of the algorithm of A-element weight adaptation can be apparently explained by non-linearity of class boundaries in the space of A-elements". These conclusions are errroneous because the analysis provided indicates that under certain conditions (see above) a non-linear class boundary cannot be formed. Here the problem is neither in the perceptron architecture nor in weight adaptation algorithm; the problem is that experiment maker does not meet the conditions described in this paper. Say, in the case under consideration there were only 20 elements, whereas the theoretical evaluation shows that at least 30 A-elements have to participate in the perceptron. From the practical point of view, it is more reliable to have at least about 100 A-elements to exclude the case when G-matrix becomes special.

The major result of the study is that it the numerically evaluates under which conditions the probability of linear separability of the patterns increases according to Cover's theorem so that it practically constitutes 1, i.e., provides a 100% confidence that any pattern will be linearly separated by a device similar to Rosenblatt's perceptron.

References

1. Бонгард М.М. Проблема узнавания. – Москва, «Наука», 1967.
2. Розенблат Ф. *Принципы нейродинамики (Перцептроны и теория механизмов мозга)*, Москва, «Мир», 1965.
3. Проект Википедия – энциклопедия в интернете http://ru.wikipedia.org/wiki/Перцептрон
4. Минский М. *Перцептроны*, Москва, «Мир», 1971.
5. Encyclopedia of Integer Sequences. "Number of nonsingular n x n matrices over GF(2)" http://www.research.att.com/~njas/sequences/A002884
6. Encyclopedia of Integer Sequences. "Number of regular n x n matrices with rational entries" equal to 0 or 1. http://www.research.att.com/~njas/sequences/A055165
7. Cover T.M. "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition", IEEE Transactions on Electronic Computers, Vol. EC-14, 1965, p.326-334.
8. Joseph R.D. "The number of orthants in n-space intersected by an s-dimensional subspace", Technical Memo 8, Project PARA, Cornell Aeronautical Lab., Buffalo, NY, 1960.
9. Joseph R.D. "Contributions to Perceptron Theory", Cornell Aeronautical Lab., № VG-1196-G-7, Buffalo, 1960.
10. Хайкин, С. «Нейронные сети: Полный курс», Neural Networks: A Comprehensive Foundation. — 2-е изд. — М.: «Вильямс», 2006. — ISBN 0-13-273350-1.
11. А.Н. Борисов, В.Е. Голендер, «Выбор прототипов перцептрона», Кибернетика и диагностика, Рига, 1968, с. 93-102.

Sergejs Jakovlevs. Tēlu aprakstu kompaktumu nodrošinošā perceptrona arhitektūra

Rakstā tiek apskatīti nosacījumi, kurus ir jāizpilda neironu tīklam, lai ar garantiju noformētu pazīmju telpu, kas apmierina kompaktuma hipotēzi. Ir precizēts kompaktuma hipotēzes formulējums attiecībā uz neironu tīkliem. Ir parādīts, ka perceptronam, neskatoties uz to, ka pirmais saišu slānis formējas gadījuma ceļā, vairāk kā 30 elementu klātesamība vidējā slānī garantē 100% varbūtību tam, ka perceptrona G-matrica nebūs īpaša. Bet tas nozīmē to, ka pie papildus matemātiskajiem pārveidojumiem kurus bija formulējis Rozenblats, perceptrons garantēti noformē pazīmju telpu, kuru pēc tam būs iespējams atdalīt lineāri. Tiešām, Kovera teorēmā ir apgalvots tikai tas, ka atdalāmības iespējamība pieaug, pārveidojot sākotnējo telpu lielāka skaita dimensiju telpā nelineārā gadījumā. Bet šī teorēma nenorāda, kad šī varbūtība būs 100%. Rozenblata perceptrona nelineārais pārveidojums tiek realizēts pirmajā slānī, kurš tiek ģenerēts gadījuma ceļā. Rakstā ir doti praktiski nosacījumi, pie kuriem varbūtība ir ļoti tuvu pie 100%. Līdzīgā gadījumā ar Rumelharta ML perceptronu tāda analīze netiek veikta.

Сергей Яковлев. Гарантированное формирование перцептроном пространства, которое удовлетворяет гипотезе компактности

В статье рассмотрены условия, которые должна выполнить нейросеть для того, чтобы гарантированно сформировать пространство признаков, удовлетворяющее гипотезе компактности. Уточнена формулировка гипотезы компактности по отношению к нейросетям. Для перцептрона показано, что несмотря на то, что первый слой связей формируется случайным образом, наличие в сети более 30 элементов в среднем слое гарантирует 100% вероятность того, что G-матрица перцептрона не будет особенной. При дополнительных математических выкладках, которые были сделаны Розенблаттом, это означает, что перцептрон гарантированно сформирует пространство признаков, которое затем сможет быть разделено линейно. Это показывает, также для других видов нейронных сетей, что наличие двух слоев связей между элементами не является достаточным условием, того, что сформированное пространство признаков будет линейно разделимо. Действительно, теорема Ковера говорит лишь о том, что вероятность разделения увеличивается при преобразовании исходного пространства в пространство большей размерности в нелинейном случае. Но она не указывает, когда эта вероятность будет 100%. Для перцептрона Розенблатта нелинейное преобразование осуществляется первым слоем, который генерируется случайным образом. В статье описаны практические условия, когда вероятность крайне близка к 100%. Для сравнения, в MLP Румельхарта такой анализ не проведен.