UDC 519.95

ON A CLASS OF PERCEPTRONS

V. N. Vapnik and A. Ya. Chervonenkis

(Moscow)

Translated from Avtomatika i Telemekhanika, Vol. 25, No. 1,
pp. 112-120, January, 1964
Original article submitted February 21, 1963

A class of perceptrons distinguished from existing perceptrons by the method of learing is considered. The block diagram of the perceptron is given and the learning method described in detail. The operating algorithms of various classes of perceptrons are compared with the theory of pattern recognition based on generalized portraits.

1. Introduction

In 1957 a group of American scientists under the leadership of Rosenblatt proposed a perceiving machine scheme which was named perceptron.

The basic considerations for the scheme were that the receptor field switching and the neurons of living organisms cannot be strictly determined, since this would lead, in the first place, to unreliability of the system, and, in the second place, to the need for transmitting a great quantity of genetic information.

Many perceptron schemes have been described in the literature since 1957. The basic aspect of all perceptrons has been that the switching of the receptor field and the neurons was random. The block diagram of the perceptron is given in Fig. 1.

The second part of the perceptron, connected with the recognition and learning block has been solved in various ways by different authors.

Thus different perceptron reward schemes were sought, for altering the weights with which the excitations of individual neurons were summed. The learning methods previously proposed for perceptrons did not guarantee that the weights $\lambda_1, \ldots, \lambda_n$, permitting patterns to be recognized, would be found if they existed. The method proposed in the present article gives this guarantee. The principle of the perceptron is founded on the axiomatic work [1]. In [1] a recognition and distinction scheme was described, connected with finding a generalized portrait and a recognition threshold. It was proposed that the machine be given a set of images of the set of objects on the unit sphere in Hilbert space, which is strictly defined for a given machine. It was shown that assignment of the set U is equivalent to the assignment of a single image

$$\mathcal{F}^* = c_1\mathcal{F}_1 \oplus c_2\mathcal{F}_2 \oplus \ldots \oplus c_n\mathcal{F}_n,$$

where c is a constant, $\mathcal{F}_i$ an image. The vector φ, the generalized portrait of the pattern, was to be found.

After development of the generalized portraits of the homogeneous system $\varphi_1^l, \ldots, \varphi_n^l$ (the upper index gives the number of the homogeneous system) and the corresponding recognition thresholds $c_1^l, \ldots, c_n^l$ the distinction problem is defined in terms of finding the vector φ_i^l for which among the numbers $(\varphi_1^l f) \ldots (\varphi_n^l f)$ the number $(\varphi_i^l f)$ be maximum (f is the unit vector in the space, corresponding to the object presented).

The problem of recognition consisted in determining the existence of and finding the vector φ_i^l for which

$$\bar{x}_1^l \ldots \bar{x}_{i-1}^l x_i^l \bar{x}_{i+1}^l \ldots \bar{x}_n^l = 1,$$

where

$$x_k^l = \theta\,[(f\varphi_k^l) - c_k^l], \qquad \theta\,(z) = \begin{cases} 1, & z \geqslant 0, \\ 0, & z < 0. \end{cases}$$

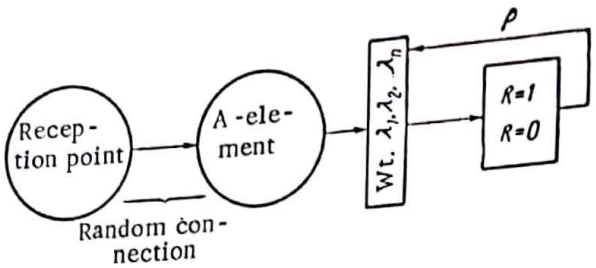

Fig. 1.

For a perceptron the prescription of an image $\mathcal{F}$ is equivalent to prescribing a commutation of the receptor field and the neurons. Using the image $\mathcal{F}$ a certain sequence $\alpha_1, \ldots, \alpha_n$ characterizing the degree of excitation of the receptor field is put into correspondance with a vector in the space E_m, where the dimensionality m of the space is defined by the number of neurons. For perceptrons finding the generalized portrait is equivalent to defining the weights λ_i:

$$\varphi = (\lambda_1, \ldots, \lambda_m),$$

with which the neuron excitations are summed in the R-cell:

$$\sum_i \lambda_i x_i = (\varphi X),$$

where λ_i is the amplification factor, and x_i is the neuron outputs.

Analyzing the perceptrons in the literature, it is easily remarked that they can be classified in two types: recognizing and distinguishing perceptrons.

Thus the perceptron of type Mark-1 [2] can be classed as a perceptron of the recognizing type, although its operation was described by the authors in the distinction regime.

Each R-element of a perceptron realizes the operation

$$R_i = \theta\left[\sum_k \lambda_{ki} x_k - 0\right].$$

It divides the set of objects into two classes: in the first are objects for which $R_i = 1$, in the second those for which $R_i = 0$. Objects belonging to the first pattern should be placed in the first class, those which belong to the second pattern, in the second class:

$$(X\varphi) \geqslant 0, \text{ if X is an element of the first pattern,}$$

$$(X'\varphi) < 0, \text{ if X' is an element of the second pattern.}$$

This mode of operation of a perceptron can be interpreted as the recognition of the first class of objects. And since it is known that an object of the first or second class will be presented a recognized object is assigned to the first class, an unrecognized one to the second.

The perceptrons described by V. M. Glushkov [3] operate in the distinction regime. In these perceptrons the index of the vector φ_i was found for which $(\varphi_i f)$ was the largest of the numbers $(\varphi_1 f), \ldots, (\varphi_n f)$.

The class of perceptrons described in the present article can operate both in the recognition and in the distinction regimes.

2. Imaging

The first portion of a perceptron realizes the translation of a sequence of numbers $\alpha_1, \ldots, \alpha_n$ $(0 \leq \alpha \leq 1)$ characterizing the degree of excitation of the receptors, into a vector in the space E_m by means of an imaging $\mathcal{F}$. (In our case this vector is unique.)

In the perceptrons described in the literature [2, 4] the imagings employed were

$$y_i = \operatorname{sgn}\left(\sum_j r_{ij}\alpha_j - \theta_j\right),$$

$$y_i = \sum r_{ij}\alpha_{jr},$$

$$y_i = \left(\sum r_{ij}\alpha_j\right)^2,$$

where y_j are the neuron outputs.

Summation was carried out each time over all the α_j, and the r_{ij} were taken at random from the possible values 1, 0, +1.

In [5] the imaging proposed was

$$y_i = \operatorname{sgn}\left[\sum_i r_{ij}x_i - \sum_k z_{kj}y_k - \theta_j\right].$$

Technically imaging is realized by the connection of a certain random set of receptors to a single summing point, a neuron. The signal of the degree of excitation of the ith receptor is received with the sign plus or minus, in dependence on the r_{ij}. The neuron has a threshold of excitation θ_j; the output of the neuron has the value 0 or 1 depending on whether the sum of signals arriving at the neuron exceeds the threshold θ_j or not.

In the perceptron class proposed here any of the above imagings can be used. The most convenient, in the authors' opinion, is the imaging where together with all the values of y_i, there participate the values $\bar{y}_i$. In this case the vectors $Y(y_1, \ldots, y_n, \bar{y}_1, \ldots, \bar{y}_n)$ have a single norm.

3. Block Diagram and Principle of Operation

The block diagram of the proposed class of perceptrons is given in Fig. 2. We find there: the receptor field RF, the neuron field NF, the connection field between them, the unit for calculating the scalar product with the generalized portrait and for comparison with the recognition threshold SPC, the memory M, the learning unit L and the control system C.

The first four units are the usual parts of a perceptron. They are present in all known perceptron designs. The unit for calculation of the scalar product with the generalized portrait is usually the R-cell, and the generalized portrait itself is the vector composed of the amplification coefficients.

This class of perceptrons is distinguished mainly by its learning principle and the learning unit. The learning unit will be described in detail below. In case of an incorrect response during operation, the vector corresponding to the unrecognized object is sent to the memory and the learning unit, and is utilized to correct the generalized portrait. The control unit realizes this function as well as the coordination of the machine.

The machine can operate in three regimes: learning, recognition, and distinction. These regimes are described in detail below.

4. Learning Regime. Calculation of the Generalized Portrait

The generalized portrait is calculated by successive approximations. Each approximation is calculated on the basis of a finite number of objects of a given pattern presented to the machine, and a finite number of objects of other patterns of the same system of homogeneous patterns (a total of n objects). We denote the set of vectors corresponding to the presented objects of a given pattern by K_1, and the set of vectors corresponding to the objects presented of the other patterns by K_2.

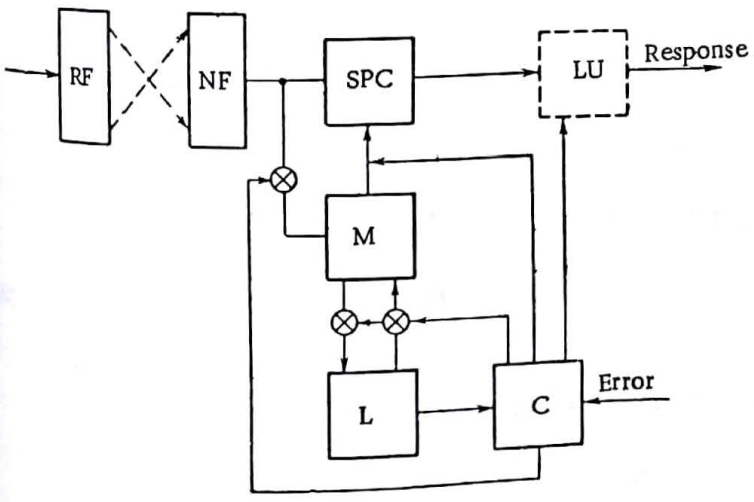

Fig. 2.

The approximation for the generalized portrait should satisfy the following condition:

$$(X\varphi) \geqslant (Y\varphi) \qquad (1)$$

(where $X \in K_1$, $Y \in K_2$, φ is the approximate generalized portrait) or,

$$(X\varphi) \geqslant c, \quad (Y\varphi) \leqslant c,$$

where c is a constant.

Among all the vectors satisfying condition (1) we find a unit vector $\varphi = \varphi_0$ such that the function $C(\varphi) = \min_{X \in K_1} (X, \varphi)$ attains a maximum, we term it the optimal approximation.

Further, we shall assume that among the acceptable vectors φ [satisfying condition (1)] there is such that $(\varphi X) > 0$ for all $X \in K_1$. Then for the optimal vector this condition is satisfied identically.

Proposition 1. The optimal vector always exists.

If the set of acceptable vectors is not empty, the proposition is obvious, since the function $C(\varphi)$ is continuous and is defined on a nonempty, closed, bounded set. This set is not empty since the generalized portrait belongs to it and exists by definition of the pattern.

Proposition 2. There exists an optimal vector φ_0 satisfying the following conditions:

$$\varphi_0 = \sum_i \alpha_i X_i + \sum_k \beta_k Y_k. \tag{2}$$

Here $C(\varphi_0) > 0$; $\alpha_i \geqslant 0$ for all i for which $(X_i \varphi_0) = C(\varphi_0)$, and $\alpha_i = 0$ for all remaining i; $\beta_k \leq 0$ for all k for which $(Y_k, \varphi_0) = C(\varphi_0)$, and $\beta_k = 0$ for all remaining k.

Further we shall term the system of vectors X_i and Y_k for which $(X_i\varphi) = (Y_k\varphi) = C(\varphi)$ the system of extreme vectors $S(\varphi)$. It is always possible to choose such small $\Delta\varphi$ ($\| \Delta\varphi \| > 0$), that in passage from φ to $\varphi + \Delta\varphi$ we have $S(\varphi) \supset S(\varphi + \Delta\varphi)$. Therefore, in the investigation of the behavior of the function $C(\varphi + \Delta\varphi)$ for small variations $\Delta\varphi$ it can be assumed that $C(\varphi + \Delta\varphi) = \min(X, \varphi + \Delta\varphi)$, $X \in S$, where X runs only over the extreme vectors $X \in S(\varphi)$. Satisfaction of condition (1) can be tested only on the extreme vectors $S(\varphi)$.

The optimal vector φ_0 should be a linear combination of the extreme vectors, since in the contrary case a small rotation of the vector, reducing the angle between the vector and the space drawn by the extreme vectors can yield an increase of $C(\varphi_0)$ without violating conditions (1) [$C(\varphi_0) > 0$].

We shall assume that the extreme vectors form a linearly independent system. Then the vector φ_0 is uniquely represented in the form

$$\varphi_0 = \sum_i \alpha_i X_i + \sum_k \beta_k Y_k.$$

If any of the α_i are less than 0, a rotation of the vector φ_0 reducing the angle between it and the hyperspace E drawn on the system of vectors $S(\varphi) \setminus X_i$ can increase $C(\varphi)$ without violating inequalities (1).

Similarly it is found that all $\beta_k \geq 0$. If the system of extreme vectors is linearly independent, it is possible to embed the space E_n in a space of dimensionality E_k such that for each vector X_i in the space E_n a vector X_i' in the space E_k is found, such that $0 < \| X_i' - X_i \| < \varepsilon$, and the system X_i' is linearly independent.

For the system X_i' proposition 2 is satisfied; the continuity of $\varphi_0(X_1', \ldots, X_n')$ permits the limiting passage $X_i' \to X_i$ to construct the vector $\varphi_0(X_1, \ldots, X_n)$ satisfying conditions (2).

Proposition 3. Conditions (2) are satisfied only by one single vector. Let now P be the vector satisfying (2) and $C(p) > 0$. We define the vector collinear with it:

$$P' = \frac{P}{C(P)}.$$

For the vector P' conditions (2) pass into the following conditions:

$$P' = \sum_i \alpha_i X_i + \sum_k \beta_k Y_k, \quad X \in K_1, \; Y \in K_2, \tag{3}$$

$$\alpha_i \geqslant 0 \text{ and } (X_i P') = 1, \; \beta_i \leqslant 0 \text{ and } (Y_k P') = 1,$$

or

$$\alpha_i = 0 \text{ and } (X_k P') > 1, \; \beta_k = 0 \text{ and } (Y_k P') < 1.$$

Let now two vectors P_0' and P_1' satisfy conditions (3). Then,

$$P_0' = \sum \alpha_i X_i + \sum \beta_k Y_k;$$

$$\alpha_i (P_0', X_i) = \alpha_i, \text{ since } \alpha_i = 0, \text{ or } (P_0' X_i) = 1;$$

$$\beta_k (P_0' Y_k) = \beta_k, \text{ since } \beta_k = 0, \text{ or } (P_0' Y_k) = 1;$$

$$\alpha_i (P_1' X_i) \geqslant \alpha_i, \ \beta_k (P_1' Y_k) \geqslant \beta_k, \text{ since}$$
$$(P_1' x_i) \geqslant 1 \text{ and } \alpha_i \geqslant 0, \ (P_1' Y_k) \leqslant 1 \text{ and } \beta_k \leqslant 0;$$

$$\| P_0' \|^2 = \left(P_0' \sum (\alpha_i x_i + \beta_k Y_k)\right) = \sum \alpha_i + \sum \beta_k;$$

$$(P_0 P_1) \geqslant \sum \alpha_k + \sum \beta_k; \ \| P_0' \|^2 \leqslant (P_0' P_1').$$

Similarly

$$\| P_1' \|^2 \leqslant (P_0' P_1'),$$

whence

$$P_1' = P_0'.$$

This demonstrates the uniqueness of the vector satisfying conditions (3) and the validity of proposition 3.

Propositions 1 and 2 prove the existence of a vector satisfying conditions (2). But since $C(\varphi_0) > 0$, then it is obvious that there always exists a vector satisfying conditions (3).

We now consider the following system of differential equations:

$$\frac{d\alpha_i}{dt} = -\varepsilon\alpha_i + F_1\left(1 - \sum_1^l k_{ij}\alpha_j - \sum_{l+1}^n k_{ij}\beta_j\right) \qquad (1 \leqslant i \leqslant l),$$

$$\frac{d\beta_i}{dt} = -\varepsilon\beta_i + F_2\left(1 - \sum_1^l k_{ij}\alpha_j - \sum_{l+1}^n k_{ij}\beta_j\right) \qquad (l+1 \leqslant i \leqslant n), \tag{4}$$

where the k_{ij} are pairwise scalar products of vectors from $K_1 \cup K_2$, where the number the vectors in K_1 from 1 to l, and those from K_2 from $l+1$ on. The functions

$$F_1(x) = \begin{cases} 0 & \text{for } x \leqslant 0, \\ x & \text{for } x \geqslant 0, \end{cases}$$

$$F_2(x) = \begin{cases} x & \text{for } x \leqslant 0, \\ 0 & \text{for } x \geqslant 0. \end{cases}$$

The conditions for equilibrium of the system as $\varepsilon \to 0$ are expressed in the form

$$\alpha_i \geqslant 0 \text{ and } \sum_1^l k_{ij}\alpha_j + \sum_{l+1}^n k_{ij}\beta_j = 1; \qquad (1 \leqslant i \leqslant l)$$

or

$$\alpha_i = 0 \text{ and } \sum_1^l k_{ij}\alpha_j + \sum_{l+1}^n k_{ij}\beta_j > 1;$$

$$\beta_i \leqslant 0 \text{ and } \sum_1^l k_{ij}\alpha_j + \sum_{l+1}^n k_{ij}\beta_j = 1 \qquad (l+1 \leqslant i \leqslant n)$$

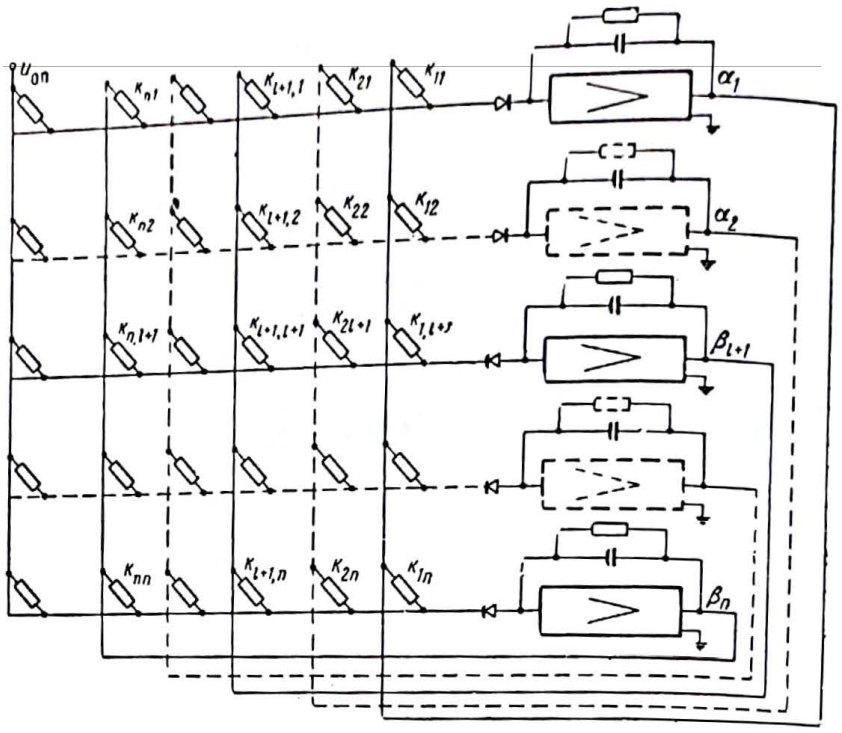

Fig. 3.

or

$$\beta_i = 0 \text{ and} \sum_1^l k_{ij}\alpha_j + \sum_{l+1}^n k_{ij}\beta_j < 1.$$

Introducing the vector

$$\varphi' = \sum_1^l \alpha_i X_i + \sum_{l+1}^n \beta_i Y_i, \quad X \in K_1, \quad Y \in K_2,$$

we obtain

$$\alpha_i \geqslant 0 \text{ and } (\varphi' X_i) = 1 \text{ or } \alpha_i = 0 \text{ and } (\varphi' X_i) > 1;$$

or

$$\beta_i \leqslant 0 \text{ and } (\varphi' Y_i) = 1$$

$$\beta_i = 0 \text{ and } (\varphi' Y_i) < 1.$$

But these are just conditions (3) applied to the vector φ'. This signifies that by virtue of uniqueness proven in proposition 3, the vector $\varphi_0 = \varphi'/\|\varphi'\|$ is the optimal approximation. The system (4) is easily modeled. A circuit of the model is given in Fig. 3. The quantity ε defines the precision of solution and is taken so small that the error arising due to the finite value of ε be commensurable with the other errors.

The input data for calculating the generalized portrait are thus the scalar products of the vectors k_{ij} from K_1 and K_2. The outputs are the coefficients of the expansion of the generalized portrait φ_0 in these vectors. In the calculation a system of essential representatives is automatically determined, corresponding to the system of extreme vectors. Only for these vectors are the coefficients α and β distinct from zero.

The recognition threshold for the vector φ' is 1, and for the generalized portrait $\varphi_0 = \varphi'/\|\varphi\|$, $c = 1/\|\varphi'\| = 1/\sqrt{\Sigma\alpha_i + \Sigma\beta_i}$. Thus calculation of the generalized portrait φ_0 and threshold c is realized by finding the equilibrium position of the system (4).

In the learning regime the machine as a whole operates in the following manner. Certain objects are presented to the machine, belonging to a given pattern. We obtain a vector at the output of the neuron field for each object, which serves as the latter's representative in the given random imaging. These vectors are stored in the

memory. Then in the learning unit the matrix of scalar products of these vectors is calculated and the approximate generalized portrait is found by the above-described method. The generalized portrait is found by the above-described method. The generalized portrait is stored in the memory, and of the vectors presented for learning only those are retained in the memory which enter into the generalized portrait with nonzero coefficients (extreme vectors). Further learning of the given pattern must be carried out during learning of other patterns. At each presentation it is verified whether the object is correctly recognized by the generalized portrait or not. The scalar product should exceed the threshold if the object belongs to the given pattern, and should not exceed it if the presented object belongs to another pattern of the system $[\mathcal{F}_k(\Phi)]$ of homogeneous patterns.

In case of error the operator introduces into the machine the signal "error." The vector corresponding to the incorrectly recognized object is placed in the memory with indication of membership or nonmembership of the object in a given pattern. The learning block further calculates the extended matrix of scalar products (a row and column corresponding to the new vector are added) and the new approximate generalized portrait and threshold are calculated. Their values are transferred to memory and simultaneously those vectors are removed from memory which enter into the new generalized portrait with zero coefficients.

Learning is continued until the probability of error is reduced to a prescribed value. Learning can be continued even further, during the recognition phase of the machine's operation, but now it will be necessary to store not only the generalized portraits and the thresholds, but also the system of extreme vectors composing the generalized portrait by their linear combination (to form the extended scalar product matrix).

5. Recognition Regime

During recognition a system of homogeneous patterns is shown to the machine and an object is shown. The machine should assign it to one of the patterns or decide that it does not belong to any of them. The machine forms the scalar product of the vector corresponding to the object shown with each of the generalized portraits of the given system of homogeneous patterns and compares them with the thresholds. Further a logical treatment of the obtained data occurs; if none or more than one scalar product has exceeded its threshold, the object does not belong to any of the patterns: if only a single product $(X\varphi_i)$ has exceeded its threshold, then the object belongs to the corresponding pattern.

In this method the object has to be compared with all of the generalized portraits of the system (n comparisons).

6. Distinction Regime

In distinction as well a system of homogeneous patterns is shown to the machine, and an object is presented, but it is further assumed that the object definitely belongs to one of the patterns of the system.

In distinction the operation of the machine differs from that in recognition only in the following two points.

There is no logical treatment required of the post-threshold data, since it is certain that one and only one scalar product exceeds its threshold.

A method of operation is also possible without threshold comparisons. The generalized portraits are normalized so that the thresholds are equal. Then that generalized portrait φ_i for which the maximum $(\varphi_k X)$ is obtained corresponds to the required pattern.

LITERATURE CITED

1. V. N. Vapnik and A. Ya. Lerner, Pattern Recognition by Means of Generalized Portraits, Avtomatika i Telemekhanika, 24, No. 6 (1963).
2. J. S. Hay, F. S. Martin, and S. V. Whitman, Perceptron Mark-1, Its Design and Characteristics, Cybernetic Collection, No. 4, Izd. Inostr. Liter. (1962).
3. V. M. Glushkov, Theory of Learning for a Class of Discrete Perceptrons, Zh. Vychislit. Matemat. i Matematich. Fiz., 2, No. 2 (1962).
4. L. J. Roberts, Pattern Recognition with Adaptive Systems, Cybernetic Collection, No. 4, Izd. Inostr. Liter. (1962).
5. F. Rosenblatt, Generalized Perception Over Transformation Groups, Cybernetic Collection, No. 4, Izd. Inostr. Lit. (1962).