American Journal of Mathematical and Computer Modelling
2020; 5(3): 83-88
http://www.sciencepublishinggroup.com/j/ajmcm
doi: 10.11648/j.ajmcm.20200503.14
ISSN: 2578-8272 (Print); ISSN: 2578-8280 (Online)

One Approach to the Problem of the Existence of a Solution in Neural Networks

Sargsyan Siranush[1, 2], Hovakimyan Anna[1, 2, *]

[1]Department of Programming and Information Technologies, Erevan State University, Yerevan, Armenia

[2]Department of System Programming, Russian-Armenian University, Yerevan, Armenia

Email address:
siranushs@ysu.am (S. Siranush), ahovakimyan@ysu.am (H. Anna)
[*]Corresponding author

To cite this article:
Sargsyan Siranush, Hovakimyan Anna. One Approach to the Problem of the Existence of a Solution in Neural Networks. *American Journal of Mathematical and Computer Modelling*. Vol. 5, No. 3, 2020, pp. 83-88. doi: 10.11648/j.ajmcm.20200503.14

Received: August 7, 2020; **Accepted**: August 21, 2020; **Published**: September 16, 2020

Abstract: Artificial neural networks are widely used to solve various applied problems. For the successful application of artificial neural networks, it is necessary to choose the correct network architecture, to select its parameters, threshold values of the elements, activation functions, etc. The problem of evaluating the neural network parameters, based on a study of the probabilistic behavior of the network is much promising. The study in the direction of developing probabilistic methods for perceptron-type pattern recognition systems is considered in different works. The concept of the characteristic function of the perceptron introduced by S. V Dayan was used by him to prove theorems on the existence of a perceptron solution. At the same time, issues of choosing a network architecture, theoretical assessment, and optimization of neural network parameters remain relevant. In this paper, we propose a mathematical apparatus for studying the relationship between the probability of correct classification of input data and the number of elements of hidden layers of a neural network. To evaluate the network performance and to estimate some parameters of the neural network such as the number of associative elements depending on the number of classification classes the mathematical expectation and variance of weights at the input of the output layer are considered. A theorem on the necessary and sufficient condition for the existence of a solution for a neural network is proved. By a solution of neural networks, the ability to recognize images with a probability other than zero is meant. The results of the proved theorem and its corollaries coincide with the results obtained by F. Rosenblat and S. Dayan for the perceptron in a different way.

Keywords: Neural Networks, Parameters of Neural Network, Probability of Recognition, Solution in Neural Network, Characteristic Function

1. Introduction

Artificial neural networks have been developed for a long time. They are widely used for solving various applied problems. Currently, there is a significant increase in interest in artificial intelligence, caused by both the development of technical means and the demand of the software market for a qualitatively new product.

Against this process, numerous attempts are being made to apply various models of neural networks. Artificial neural networks are becoming more common due to such factors as the ability to solve difficultly formalized tasks, perform parallel data processing, use large amounts of data, etc. [1-6].

For the successful application of artificial neural networks, it is necessary to choose the correct network architecture, select its parameters, threshold values of the elements, activation functions, etc. [2-5, 27].

Research on the successful construction and the use of artificial neural networks is conducted mainly in the following areas: the selection of optimal learning algorithms, selection and optimization of neural network parameters (such number of layers, number of neurons in each layer, activation functions, etc.), as well as research on problems related to the convergence of the neural network. Since these tasks are interrelated, research on their solution was mainly conducted in parallel.

The issues of developing the optimal learning algorithms for neural networks are considered in the works of F. Rosenblatt. V. M. Glushkov, R. D. Joseph, A. G. Ivakhnenko, N. Nielson and others [4-8, 26-29].

Perceptron convergence issues have an important role in research related to the modeling and creation of recognition systems. These tasks are considered in the works of F. Rosenblatt and his colleagues (Nilson, Blok, Joseph, Kesten, etc.) [1, 4, 10-13].

F. Rosenblatt proved a convergence theorem for the perceptron, which states that an elementary perceptron, regardless of the initial state of the weight coefficients and the sequence of occurrence of stimuli, will always lead to a solution in a finite time. F. Rosenblatt also presented the proves of some concomitant theorems and their consequences that showed what requirements the architecture of artificial neural networks and the methods of their training should meet [1].

The studies on neural networks were intensified in the 70s of the last century.

In 1970 A. G. Ivakhnenko developed a group method of data handling, which allows not only to calculate the weights of connections between neurons but also to determine the number of layers in the network and the neurons in them depending on the needs of the applied task [5-9].

In 1989, some of authors obtained a result stating that a perceptron with one hidden layer is an universal approximator, that is, it can approximate any continuous function if a continuous, monotonously increasing and limited function as an activation function of neural elements of the hidden layer is used [10-12]. Moreover, the accuracy of the approximation of the function depends on the number of neurons in the hidden layer. Thus, a perceptron with one hidden layer and an activation function of the aforementioned type is a universal classifier. In [12], it was also stated that for a network with $(n - m - p)$ architecture, to solve the problem of pattern classification (that is, perceptron convergence), there is to be inequality

$$log_2 p < m < (L - p)/(n + p - 1),$$

where n is the number of elements in the input layer, m is the number of neurons in the hidden layer, p is the number of classes into which it is necessary to split the input space of images, L is the volume of the training sample.

To select the network structure, aspects of the use of genetic algorithms have also been investigated. It should be noted that the conditions for the convergence of such algorithms are not well studied, even less is known about the rate of convergence [13].

In 1992 the architecture of cresceptron neural networks appeared [14, 15]. Cresceptron changes its topology during training, by analogy with networks using a group method of data handling [5]. An important idea proposed in the cresceptron is the use of max-pooling layers instead of layers with average. Layers of maximum choice are now widely used in convolutional neural networks. However, for training modern convolutional networks, the error backpropagation algorithm is used, which is more efficient [16, 17].

In 2006-2007, the development of deep learning convolutional networks based on training with a teacher took place. The work [16] describes the application of the error backpropagation algorithm for training a deep neural network with an architecture similar to a neocognitron and a cresceptron, consisting of alternating layers of convolution and maximum choice. This architecture of neural networks is actively used to date.

Sergey Ioffe and Christian Szegedy in 2015 proposed to use in neural networks special layers of batch normalization [18-21]. In [22], it was shown that the backpropagation error algorithm converges faster if the input data are normalized. It was noticed that when a signal propagates through a neural network, its math. expectation and disperse change from layer to layer, which negatively affects the learning process. Joffe and Zhegedy proposed to perform normalization not only at the entrance to the neural network but also before each layer of the network.

Some scientists considered probabilistic neural networks (PNN) widely used in classification problems. The essence of such networks is that the outputs of the network can be interpreted as estimates of the probability that an element belongs to a certain class, and the network actually learns to evaluate the probability density function [23, 24].

The task of estimating probability density according to data belongs to the field of Bayesian statistics. In contrast to Bayesian statistics, conventional statistics on a given model determines the probability of an outcome. In this case, the density has a certain definite form and the model parameters are estimated analytically. Bayesian statistics make it possible to evaluate the correctness of a model from available reliable data, that is, it makes it possible to estimate the probability density of distributions of model parameters from available data [25].

Another approach to estimating the probability density is based on nuclear estimates [26]. In this case, if there are a sufficient number of training examples, then the method gives a fairly good approximation to the true probability density.

Work on the creation of perceptron-type pattern recognition systems was also carried out in a different direction, namely, in the direction of developing probabilistic methods for perceptron studying [27]. The basis of this approach is the concept of the characteristic function of the perceptron (CFP), introduced by S. V Dayan [28-30]. Using CFP, theorems on the existence of a perceptron solution, on the choice of the number of elements of the hidden layer, on the length of the training sequence, etc., are proved [27, 31]. This direction of research continues to be developed by colleagues and students of S. V. Dayan [27, 31-36]. At the same time, issues of choosing a network architecture, theoretical assessment, and optimization of neural network parameters remain relevant.

In this paper we propose a mathematical apparatus for studying the relationship between the probability of correct classification of input data and the number of elements of hidden layers of a neural network. A necessary and sufficient condition for the existence of a solution of a neural network is proved. By a solution of neural network the ability to recognize images with a probability other than zero is meant. As a consequence of the proved theorem, the results obtained by F. Rosenblat [1] and S. Dayan [27] for the perceptron were obtained.

2. The Proposed Mathematical Model of Neural Network and Solution of the Problem

In this work, a neural network of direct propagation is investigated. This kind of neural network consists of a layer of input nodes, hidden layers, and an output layer. Neurons have unidirectional connections, do not contain connections between the elements inside the layer and feedback connections between the layers. The neurons of the input layer are connected to the neurons of the hidden layer by excitatory and inhibitory connections in a random way. The outputs of all the neurons of the hidden layer are connected to the neurons of the output layer. Neurons in each layer are referred to as input, hidden and output elements, respectively [1, 27, 31, 32, 34].

The input layer is represented by the receptor field S, the hidden layer consists of N associative elements forming the set A, the output layer consists of a finite number of reacting R-elements. The outputs of all associative elements are connected to reactive elements [34-36].

An image is formed in the receptor field, corresponding to external irritation. Under the image we mean a certain vector, the coordinates of which correspond to individual elements of the receptor field and can take the values 1 and 0, depending on whether the corresponding element is excited or not.

We consider an N-valued function f defined on some set $X$of vectors $x$and taking for each vector $x \in X$ the values $V_k\ (k = 1, 2, \dots N)$. The function f maps the receptor field to the associative layer with the value of V_k being the weight of the associative element A_k for the input vector x. If there are two different mappings f' and f'' then for vector x the following inequality holds

$$f'(x) > f''(x), if\ \sum_{k=1}^{N}\left(V_k' - V_k''\right) > 0,$$

where V_k' and V_k'' are the weights of the element $A_k\ (k = 1, 2, \dots, N)$ for mappings f' and f'' respectively.

Let there be some set X containing L vectors-pathogens $x_1, x_2, \dots, x_L$ in the receptor field S. We divide this set into d disjoint classes $X_1, X_2, \dots, X_d$

$$X = \bigcup_{i=1}^{d} X_i$$

A neural network has a solution for the set X, if and only if there are d mappings $f_1, f_2, \dots, f_d$ such that $f_i(x) > f_j(x), x \in X_i, j = 1, 2, \dots, d, i = 1, 2, \dots, d, i \neq j$.

For all pathogens x belonging to the same class X_i and for each element $A_k \in A$ the functions η^{ki} are introduced, taking values 0 and 1 and characterizing the activity of the element A_k under the influence of pathogens from class X_i [27]. If the external environment is divided into d classes $X_1, X_2, \dots, X_d$ and numbers of allocated in the classes representatives are $l_1, l_2, \dots, l_d$, respectively, then in the k^th A-element A_k, using the mapping f, the weight V_k is accumulated and calculated by the formula [27, 34, 35]:

$$V_k = \sum_{i=1}^{d}(\delta_i \sum_{j=1}^{li} \eta^{kj}) + V_0^k \qquad (1)$$

where V_0^k is the initial weight of the k^th A-element, δi is increment of the weight of the A-element, when one pathogen is shown from the i^{th} class of pathogens, η^{kj} is activity of A_k, $A_k \in A$ under the influence of pathogens from class X_j.

When an image appears on the receptor field the A-element can either be excited or remain unexcited. Let us denote by P_i the probability that an A-element is excited when a single image from the class X_i appears, $i = 1, 2, \dots, d$.

As a measure of the quality of recognition Dayan S. V. has introduced the characteristic function of the perceptron-type neural network (CFP) [27-30]. For each class X_i the characteristic function ζ_i has a form

$$\zeta_i = P_i - \sum_{\substack{j \\ i \neq j}} P_{ij} + \dots + (-1)^d P_{12\dots d} \qquad (2)$$

where $P_{12\dots k}\ (k = 1, 2, \dots, d)$ is the probability of A-element excitation from pathogens $S_1, S_2, \dots, S_d$. Note that the CFP characterizes the probability that the A-element is excited when a pathogen belonging to a certain class is shown and is not excited by a pathogen not belonging to the same class. If the pathogens excite A-elements with the same probability, then the expectation of the weight at the outputs of the A-elements (or, equivalently, at the input of R-elements) has the form [27, p. 305, Theorem 3]

$$\mu_i = N\delta_i \sum_{j=1}^{l} \zeta_i^j,\ l_1 = l_2 = \dots = l_d = l \qquad (3)$$

where l is the number of consecutively shown pathogens, N is the number of A-elements, δ_i is the increment of the weight of the A-element when showing one pathogen from the i^{th}class of pathogens.

When the control pathogen S_y is presented, summing formula (1) overall A- elements, we get the total weight of the associative layer U_y at the output of the associative layer. Then the dispersion of the weights is represented by the formula

$$\sigma^2 U_y = \mu(U_y^2) - (\mu(U_y))^2$$

Using the above concepts of the characteristic function and the math. expectation and disperse of weights, the following theorem is proved.

Theorem.

If a set of neural networks and a classification of the external environment are given, then for the existence of a solution it is necessary and sufficient that there be an inequality

$$N\zeta \geq P + \sigma^2/\mu^2 \qquad (4)$$

Sufficiency.

If the condition (4) is satisfied, then $N\zeta \geq 0$. Let us show that $N\zeta > 0$.

Using the law of large numbers, one can find the dependence of the probability of correct identification on the expectation and variance of the input quantity, i.e.

$$P(x > 0) \geq 1 - \sigma^2/\mu^2, \text{ if } \mu > 0$$

$$P(x < 0) \geq 1 - \sigma^2/\mu^2, \text{ if } \mu < 0 \qquad (5)$$

Consequently, for large μ and small σ, the correct separation of pathogens occurs with a probability close to one, so

$$if\ \sigma^2/\mu^2 \to 0, then\ P \to 1$$

$$if\ P \to 0, then\ \sigma^2/\mu^2 \to 1$$

So P and σ^2/μ^2 can't be zero at the same time, then $N\zeta > 0$. Hence $\zeta \neq 0$. In this case, by the Dayan theorem ([27], p. 310, Theorem 8), it follows that there exists a solution of neural network.

Necessity.

If a set of neural networks and the classification of the external environment are given, then if $\zeta \geq 1/N$ then there are solutions for any classification of the external environment [27].

In that case, we have

$$N\zeta \geq 1, \qquad N\zeta - \sigma^2/\mu^2 \geq 1 - \sigma^2/\mu^2 \leq P$$

Considering formula (5), for minimal P we get

$$N\zeta - \sigma^2/\mu^2 \geq P, N\zeta \geq P + \sigma^2/\mu^2$$

Q. E. D.

Corollary 1. For pattern recognition with a probability other than zero, it is necessary to have at least N associative elements, where

$$N \geq (\mu^2 P + \sigma^2)/(\mu^2\,\zeta) \qquad (6)$$

Corollary 2. If $L = 2$ then to classify images the number of associative elements $N \geq 4$.

Proof.

Since $\sigma^2/\mu^2 \geq 0$ we get

$$N\zeta \geq P + \sigma^2/\mu^2 \geq P, N\zeta \geq P, N \geq\ P/\zeta$$

If $L = 2$, then max ζ=0.25=1/4 [27, p. 307, Theorem 5, Corollary 3].

Consequently, $N \geq 4P$ and since $max\ P = 1$, then $N \geq 4$.

Corollary 3. For correct classification, the number of identifiable classes must be less than the number of associative elements.

Proof.

For a fixed i^{th}class formula (4) looks as follows

$$N\zeta_i\ \geq P_i + \sigma_i^2/\mu_i^2.$$

So

$$\zeta_i \geq (P_i + \sigma_i^2/\mu_i^2)/N \text{ and } \zeta_i \geq P_i/N + \sigma_i^2/(\mu_i^2 N)$$

Summing up by classes, we find

$$\sum_{i=1}^{d} \zeta_i\ \geq\ 1/N * \sum_{i=1}^{d} P_i\ + 1/N * \sum_{i=1}^{d}(\sigma_i^2/\mu_i^2) \qquad (7)$$

where d is the number of classes.

Since in [27, p. 331] there is the following relationship

$$\sum_{i=1}^{d} \zeta_i < 1, \qquad (8)$$

then substituting the estimate (8) in (7), and strengthening it, we obtain

$$1/N * \sum_{i=1}^{d} P_i + \frac{1}{N} * \sum_{i=1}^{d}(\sigma_i^2/\mu_i^2) < 1.$$

So

$$1/N * (\sum_{i=1}^{d}(P_i + \sigma_i^2/\mu_i^2)) <\ 1 \qquad (9)$$

Considering inequality (5), we obtain

$$P \geq 1 - \sigma^2/\mu^2, \qquad P + \sigma^2/\mu^2 \geq 1$$

Substituting the last in (9), for all classes we get the inequalities

$$1/N * \sum_{i=1}^{d} 1 < 1, d/N < 1, d < N \qquad (10)$$

The result obtained coincides with the result of the F. Rosenblatt theorem [1, Theorem 3, Corollary 2, p. 101] and the theorem on the choice of the number of A-elements [27, Theorem 11, Corollary 9, p. 331].

The dependence (4) is a convenient mathematical apparatus for the study of the statistical characteristics of neural networks. The obtained estimates can be used in defining the network architecture for application in practice tasks.

3. Conclusion

In this work, a mathematical apparatus for studying the probabilistic behavior of a perceptron-type neural network is developed. This apparatus is based on the characteristic function of a perceptron.

A theorem on the necessary and sufficient condition for the existence of a solution of a neural network is proved.

As a consequence of this theorem, results are obtained that are consistent with the results of F. Rosenblatt and S. Dayan [1, 2].

The results obtained are of both theoretical and practical interest. Until now, in most cases, network parameters are selected empirically and refined as a result of experiments.

The results obtained in the work connect the network architecture with the probability of correct pattern recognition. These results give constructive recommendations for the construction of recognition systems based on neural networks.

Acknowledgements

We are thankful to our families and colleagues from our department for their patient attitude and for the time, that they

spent communicating with us to discuss the problems presented in the paper.

References

[1] Rosenblatt F. Principles of Neurodynamics. Perceptrons and Theory of Brain Mechanisms. M., Mir, 1965 (in Russian).

[2] Glushkov V. M. Introduction to cybernetics. Publishing House of the Academy of Sciences of the Ukrainian SSR, K., 1964. (in Russian).

[3] Ivakhnenko A. G. Self-learning recognition systems and automatic control. Technique, K., 1969. (in Russian).

[4] Nielson N. Learning machines. Mir, M., 1967. (in Russian).

[5] Ivakhnenko A. G. Heuristic Self-Organization in Problems of Engineering Cybernetics // Automatica. 1970. Vol. 6, No. 2. P. 207-219.

[6] Ivakhnenko A. G. Polynomial Theory of Complex Systems // IEEE Transactions on Systems, Man and Cybernetics. 1971. Vol. 4. P. 364-378.

[7] Ikeda S., Ochiai M., Sawaragi Y. Sequential GMDH Algorithm and Its Application to River Flow Prediction // IEEE Trans. on Systems, Man and Cybernetics. 1976. Vol. 7. P. 473-479.

[8] Witczak M, Korbicz J, Mrugalski M., et al. A GMDH Neural Network-Based Approach to Robust Fault Diagnosis: Application to the DAMADICS Benchmark Problem // Control Engineering Practice. 2006. Vol. 14, No. 6. P. 671-683.

[9] Kondo T., Ueno J. Multi-Layered GMDH-type Neural Network Self-Selecting Optimum Neural Network Architecture and Its Application to 3-Dimensional Medical Image Recognition of Blood Vessels // International Journal of Innovative Computing, Information and Control. 2008. Vol. 4, No. 1. P. 175-187.

[10] Cybenko G. Approximations by superpositions of a sigmoidal function // Mathematics of control, signals, systems. 1989. Vol. 2. P. 303–314.

[11] Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators // Neural networks. 1989. № 2 P. 359–366.

[12] Golovko, V. A. Krasnoproshin V. V. Neural network data processing technologies: textbook. Minsk: BSU, 2017, 263 p. (in Russian).

[13] Tarkhov D. A. Neural networks. Models and algorithms. Book 18. M. Radio Engineering, 2005, 256 p. (in Russian).

[14] Weng J., Ahuja N., Huang T. S. Cresceptron: a Self-Organizing Neural Network Which Grows Adaptively // International Joint Conf. on Neural Networks. 1992. Vol. 1. P. 576-581.

[15] Weng J. J., Ahuja N., Huang T. S. Learning Recognition and Segmentation Using the Cresceptron // International Journal of Computer Vision. 1997. Vol. 25, No. 2. P. 109-143.

[16] Ranzato M. A., Huang F. J., Boureau Y. L., et al. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2007. P. 1-8.

[17] Scherer D., Muller A., Behnke S. Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition. Lect, Notes in Comp. Science. 2010. Vol. 6354, P. 92-101.

[18] Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift // JMLR Workshop and Conference Proceedings. Proceedings of International Conference on Machine Learning, 2015. Vol. 37. P. 448-456.

[19] Szegedy C., Liu W, Jia Y. et al. Going Deeper with Convolutions // IEEE Conference on Computer Vision and Pattern Recognition, 2015. P. 1-9.

[20] Szegedy C., Vanhoucke V., Ioffe S., et al. Rethinking the Inception Architecture for Computer Vision // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. P. 2818-2826.

[21] Szegedy C., Ioffe S., Vanhoucke V., et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), 2017. P. 4278-4284.

[22] LeCun Y., Bottou L., Orr G. B. Efficient Back Prop // Neural Networks: Tricks of the Trade. 1998. P. 9-50.

[23] Streit R. L. Maximum likelihood training of probabilistic neural networks. IEEE Trans. Neural Networks, V. 5, 199, №5, P. 764-783.

[24] Zaknich A. Introduction to the modified probabilistic neural network for general signal processing applications. IEEE Transactions on Signal Processing, V. 46, 1998, №7, P. 1980-1990.

[25] Romero R. David and Touretzky R. T., Optical Chinese Character Recognition using Probabilistic Neural Networks, 1996.

[26] Bishop C. Neural Networks for Pattern Recognition, Oxford Univ. Press, 1995.

[27] Perceptron - pattern recognition system, Edited by Ivahnenko A. G., Kiev, Naukova Dumka, 1975, p. 426. (in Russian).

[28] S. V. Dayan. The concept of the characteristic function of the perceptron. Proceedings of the scientific and technical conference ErNIIMM (Yerevan Scientific Research Institute of Mathematical Machines). Yerevan, Armenia, 1968. (in Russian).

[29] S. V. Dayan. Optimal learning of perceptron to recognize external situations. Reports of the 24th All-Union Session. Gorky, 1968. (in Russian).

[30] S. V. Dayan. Investigation of the probabilistic properties of the characteristic function of the perceptron. In the book: The problems of bionics. 3. Publishing of Kharkov University. Kharkov. 1970. (in Russian).

[31] Sargsyan S. G., Dayan S. V. Modeling of the learning process for pattern recognition on computers, YSU, Scientific notes, 1984 (in Russian).

[32] Sargsyan S. G. Determination of the probability characteristics of adaptive recognition system, Trans. of Intern. Conf. Adaptable software, Kishinev, 1990. pp. 46-51 (in Russian).

[33] Kharatyan A., Sargsyan S. G., Hovakimyan A. S., Time Series Forecasting Using Artificial Neural Networks, YSU, Faculty of Economics. Yearbook, Yerevan, 2014 (in Russian).

[34] S. Sargsyan, A. Hovakimyan, Probabilistic Methods for Neural Networks Study, Quarterly Journal of Mechanics and Applied Mathematics, Issue 4 (2), Volume 69, Oxford University Press, Nov. 2016, pp. 669-675.

[35] S. Sargsyan, A. Hovakimyan. Statistical Evaluation of the Performance of the Neural Network. London Journal of Research in Computer Science and Technology, Volume 17 | Issue 1 | Compilation 1.0, 2017, pp. 1-6.

[36] S. Sargsyan, A. Hovakimyan, M. Ziroyan. Hybrid Method for the Big Data Analysis Using Neural Networks. Engineering Studies, Issue 3 (2), Volume 10. Taylor & Francis, 2018. P. 519-526.