DECISIONS ON SELECTING THE TRAINING ALGORITHM OF THE NEURAL NETWORK WITH A SET OF ITS BEING CONTROLLED PARAMETERS

Vadim Romanuke

Abstract: *There is predicated a decision making problem with the finite set of neural network training algorithms as alternatives against a set of that neural network parameters, being controlled and estimated. The rules for making decisions on selecting the training algorithms optimally are formulated.*

Keywords: *neural network, training algorithm, decision making problem, point-evaluation, interval-evaluation, Bayes — Laplace rule, optimal strategy.*

ACM Classification Keywords: *H.4.2 Types of Systems — Decision support, I.2.6 Learning — Connectionism and neural nets.*

Introduction

A primitive set of the neural network (NN) parameters contains the volume of the used resources (URV), erroneous performance rate (EPR), and the averaged period for teaching or training (ATP) the NN under an adjusted configuration [Haykin, 1999]. These three parameters, however, may be supplemented and split each into their subparameters [Borovikov, 2008], reflecting specific features of the functioning NN. But NN with a set of its being controlled parameters can be configured in several ways. Sometimes an NN configuration is associated with the algorithm of training the NN [Hagan, 1994], after which NN acquires converged values of its adjusted parameters [Vogl, 1988]. So, there stands an actual task to select that training algorithm (TA) which could ensure the desired values for the being controlled parameters of NN.

Training algorithms against the neural network parameters

The task of selecting the TA of NN with a set of its being controlled parameters is formulated as a decision making problem [Chernorutskiy, 2005] with the finite set of TA as alternatives against the set of NN parameters. Certainly, the values of elements of the corresponding decision matrix (DM) should be standardized or normalized [Trukhayev, 1981] so, that difference in dimension units of NN parameters would be disregarded [Bolshakov, 2007]. Although, there is a wide many of situations, when DM elements can be evaluated only as some intervals [Trukhayev, 1981]. For instance, such an evaluation is typical [Borovikov, 2008] for EPR, ATP, and rarely for URV. It causes wants to map these interval evaluations into points [Trukhayev, 1981] and then resolve the task of selection.

Target

Will formulate rules for making decisions on selecting the TA of NN with a set of its being controlled parameters. For hitting this target there should be generalized the selection task as choice task for cases when a probability measure (PM) is defined over NN parameters and when it is not.

Convention

May there be NN with M parameters $\{r_i\}_{i=1}^{M}$, each of which is evaluated as a point value or an interval of real values. This NN can be trained with one of N TA $\{a_j\}_{j=1}^{N}$. The value of the i-th parameter of NN, trained with the j-th TA, is u_{ij}, though it belongs to the defined segment as

$$u_{ij} \in \left[a_{ij};\, b_{ij}\right] \text{ by } a_{ij} < b_{ij} \tag{1}$$

if its definition is made intervally. May it also be there a convention that the parameters of NN were constituted so, that the lesser u_{ij} the better the i-th parameter of NN, trained with the j-th TA. Hence here is $N \times M$ matrix $\mathbf{U} = \left(u_{ji}\right)_{N\times M}$ of a decision making problem to select the optimal TA, making NN produce the desired values of its parameters.

1. Appropriateness check and standardization

If the value of u_{ij} is defined or determined as a point $u_{ij} = u_{ij}^{*}$, then firstly there comes an appropriateness check of

$$u_{ij}^{*} \leqslant u_i^{(\max)} \quad \forall\; i = \overline{1, M} \text{ for } j = \overline{1, N}. \tag{2}$$

Suppose that (2) is true for a set of indexes $J \subseteq \left\{\overline{1, N}\right\}$, that is

$$u_{ij}^{*} \leqslant u_i^{(\max)} \quad \forall\; i = \overline{1, M} \text{ and } \forall\; j \in J \subseteq \left\{\overline{1, N}\right\}. \tag{3}$$

The case with $|J| = \varnothing$ speaks that NN cannot be taught with the set of TA $\{a_j\}_{j=1}^{N}$. The case with $|J| = 1$ proclaims the decision on selecting the TA of NN to be already made: there is the single j_*-th TA, $j_* \in J$, for teaching the NN with the set $\{r_i\}_{i=1}^{M}$ of its being controlled parameters. In the case with $|J| = N_J > 1$ the values $\left\{\{u_{ij}^{*}\}_{j\in J}\right\}_{i=1}^{M}$ are standardized through

$$k_{ij} = \frac{u_{ij}^{*}}{\max\limits_{n\in J} u_{in}^{*}} \quad \forall\; i = \overline{1, M} \text{ and } \forall\; j \in J \tag{4}$$

or

$$k_{ij} = \frac{u_{ij}^{*}}{u_i^{(\max)}} \quad \forall\; i = \overline{1, M} \text{ and } \forall\; j \in J \tag{5}$$

by $J = \{j_l\}_{l=1}^{N_J} \subseteq \left\{\overline{1, N}\right\}$.

2. Point-evaluated DM elements and PM over NN parameters

If the value of u_{ij} is a point $u_{ij} = u_{ij}^{*}$ and PM over NN parameters is defined as weights

$$\{\mu_i^{*}\}_{i=1}^{M},\;\; \mu_i^{*} \in [0;\, 1] \quad \forall\; i = \overline{1, M} \text{ by } \sum_{i=1}^{M} \mu_i^{*} = 1, \tag{6}$$

then after (3) and (4) or (5) the optimal TA a_{j_*} has the number

$$j_* \in J_* = \arg\min_{j \in J} \sum_{i=1}^{M} \mu_i^* k_{ij} \tag{7}$$

due to the Bayes — Laplace rule, where the set J_* is always nonempty, though it is not excluded that $|J_*| > 1$. And in the case with $|J_*| > 1$ the intra-optimal TA $a_{j_{**}}$ number may be found by one of the following rules:

$$j_{**} \in J_{**} = \arg\min_{j_* \in J_*} \max_{i=\overline{1,\, N}} \left\{ k_{ij_*} \right\}, \tag{8}$$

$$j_{**} \in J_{**} = \arg\min_{j_* \in J_*} \max_{i=\overline{1,\, N}} \left\{ \mu_i^* k_{ij_*} \right\}, \tag{9}$$

$$j_{**} \in J_{**} = \arg\min_{j_* \in J_*} \left\{ \prod_{i=1}^{M} k_{ij_*} \right\}, \tag{10}$$

$$j_{**} \in J_{**} = \arg\min_{j_* \in J_*} \left\{ \prod_{i=1}^{M} \mu_i^* k_{ij_*} \right\}, \tag{11}$$

where the set J_{**} is always nonempty. Deeper, if one of the rules (8) — (11) outputs the set $|J_{**}| > 1$, then this set can be put into an non-empowered rule from those ones under minimization until the set J_{**} contains the single element or all the four rules (8) — (11) are applied. However, an arbitrary choice of TA with the number from the set $|J_*| > 1$ or $|J_{**}| > 1$ may be allowable too.

3. Point-evaluated DM elements and partially uncertain PM over NN parameters

If the value of u_{ij} is a point $u_{ij} = u_{ij}^*$ and PM $\{\mu_i\}_{i=1}^{M}$ over NN parameters is partially defined uncertain as weights

$$\{\mu_i\}_{i=1}^{M},\ \mu_i \in \left[\mu_i^{(\min)};\, \mu_i^{(\max)} \right],\ 0 < \mu_i^{(\min)} < \mu_i^{(\max)} < 1,\ \sum_{i=1}^{M} \mu_i = 1, \tag{12}$$

then before applying a making decision rule for TA choice the segments in (12) must be mapped properly into points (6), where in mapping the i-th segment $\left[\mu_i^{(\min)};\, \mu_i^{(\max)} \right]$ into the point μ_i^* there must be $\mu_i^* \in \left[\mu_i^{(\min)};\, \mu_i^{(\max)} \right]$ necessarily. Having got (3) and (4) or (5), here is two ways of accomplishing this mapping. The first is antagonistic (minimax) one: if an optimal strategy

$$\hat{\mathbf{X}} = \left[\hat{\mu}_1 \quad \hat{\mu}_2 \quad \cdots \quad \hat{\mu}_M \right] \in \arg\max_{\mathbf{X} \in \mathcal{X}} \min_{\mathbf{Q} \in \mathcal{Q}} \left(\mathbf{X} \cdot \mathbf{K} \cdot \mathbf{Q}^{\mathrm{T}} \right) = \hat{\mathcal{X}} \subset \mathcal{X} =$$
$$= \left\{ \mathbf{X} = \left[\mu_1 \quad \mu_2 \quad \cdots \quad \mu_M \right] \in \mathbb{R}^M : \sum_{i=1}^{M} \mu_i = 1,\ \mu_i \in [0;\, 1]\ \forall\ i = \overline{1,\, M} \right\} \tag{13}$$

by

$$\mathbf{Q} = \left[q_{j_1} \quad q_{j_2} \quad \cdots \quad q_{j_{N_J}} \right] \in \mathcal{Q} = \left\{ \mathbf{Q} \in \mathbb{R}^{N_J} : \sum_{l=1}^{N_J} q_{j_l} = 1,\ q_{j_l} \in [0;\, 1]\ \forall\ l = \overline{1,\, N_J} \right\} \tag{14}$$

in the $\mathbf{K} = \left(k_{ij_l} \right)_{M \times N_J}$ game appears such, that

$$\widehat{\mu}_i \in \left[\mu_i^{(\min)};\, \mu_i^{(\max)}\right] \quad \forall\; i = \overline{1,\, M} \tag{15}$$

then

$$\mu_i^* = \widehat{\mu}_i \quad \forall\; i = \overline{1,\, M}. \tag{16}$$

Consequently, instead of the rules (7) — (11) the $\mathbf{K} = \left(k_{ij_l}\right)_{M \times N_J}$ game ensues: the second player optimal strategy

$$\breve{\mathbf{Q}} = \left[\breve{q}_{j_1} \quad \breve{q}_{j_2} \quad \cdots \quad \breve{q}_{j_{N_J}}\right] \in \operatorname{arg}\min_{\mathbf{Q} \in \mathcal{Q}} \max_{\mathbf{X} \in \mathcal{X}} \left(\mathbf{X} \cdot \mathbf{K} \cdot \mathbf{Q}^{\mathrm{T}}\right) = \breve{\mathcal{Q}} \subset \mathcal{Q} =$$

$$= \left\{\mathbf{Q} = \left[q_{j_1} \quad q_{j_2} \quad \cdots \quad q_{j_{N_J}}\right] \in \mathbb{R}^{N_J} : \sum_{l=1}^{N_J} q_{j_l} = 1,\, q_{j_l} \in [0;\, 1] \;\; \forall\; l = \overline{1,\, N_J}\right\} \tag{17}$$

in this game shows that it is optimal to apply the j_l-th TA with the probability (or weight) $\breve{q}_{j_l}$ by $l = \overline{1,\, N_J}$ inasmuch as

$$\breve{\mathcal{Q}} = \operatorname{arg}\min_{\mathbf{Q} \in \mathcal{Q}} \left(\widehat{\mathbf{X}} \cdot \mathbf{K} \cdot \mathbf{Q}^{\mathrm{T}}\right) \tag{18}$$

by (13). But if (15) is not true for any optimal strategy from the set $\widehat{\mathcal{X}}$ then the second way is in removing $\left\{\left[\mu_i^{(\min)};\, \mu_i^{(\max)}\right]\right\}_{i=1}^{M}$ uncertainties [Romanuke, 2011] independently of the optimal strategy (13):

$$\mu_k^* = \frac{\mu_k^{(\max)}}{1 + \sum\limits_{m=1}^{M-1} \mu_m^{(\max)} - \sum\limits_{m=1}^{M-1} \mu_m^{(\min)}} \quad \text{for} \quad k = \overline{1,\, M-1} \tag{19}$$

and

$$\mu_M^* = 1 - \sum_{k=1}^{M-1} \mu_k^* = 1 - \frac{\sum\limits_{k=1}^{M-1} \mu_k^{(\max)}}{1 + \sum\limits_{m=1}^{M-1} \mu_m^{(\max)} - \sum\limits_{m=1}^{M-1} \mu_m^{(\min)}} = \frac{1 - \sum\limits_{m=1}^{M-1} \mu_m^{(\min)}}{1 + \sum\limits_{m=1}^{M-1} \mu_m^{(\max)} - \sum\limits_{m=1}^{M-1} \mu_m^{(\min)}} \tag{20}$$

if

$$\frac{\mu_k^{(\max)}}{1 + \sum\limits_{m=1}^{M-1} \mu_m^{(\max)} - \sum\limits_{m=1}^{M-1} \mu_m^{(\min)}} \geqslant \mu_k^{(\min)} \quad \forall\; k = \overline{1,\, M-1}. \tag{21}$$

Nevertheless, let $\mathcal{M}_r$ be the set of those indexes j from the set $\left\{\overline{1,\, M-1}\right\}$ for which there is

$$\frac{\mu_j^{(\max)} \left(1 - \sum\limits_{l \in \bigcup\limits_{q=1}^{r-1} \mathcal{M}_q} \mu_l^{(\min)}\right)}{1 + \sum\limits_{i=1}^{M-1} \mu_i^{(\max)} - \sum\limits_{l \in \bigcup\limits_{q=1}^{r-1} \mathcal{M}_q} \mu_l^{(\max)} - \sum\limits_{i=1}^{M-1} \mu_i^{(\min)}} < \mu_j^{(\min)} \tag{22}$$

by $j \in \mathscr{M}_r$ and $r \in \left\{\overline{1,\, M-2}\right\}$. Then

$$\mu_j^* = \mu_j^{(\min)} \quad \forall\ j \in \mathscr{M}_r \quad \text{after} \quad \mu_j^* = \mu_j^{(\min)} \quad \forall\ j \in \bigcup_{q=1}^{r-1} \mathscr{M}_q \tag{23}$$

and

$$\mu_k^* = \frac{\mu_k^{(\max)}\left(1 - \sum\limits_{l \in \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_l^{(\min)}\right)}{1 + \sum\limits_{i=1}^{M-1} \mu_i^{(\max)} - \sum\limits_{l \in \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_l^{(\max)} - \sum\limits_{i=1}^{M-1} \mu_i^{(\min)}} \quad \text{for} \quad k \in \left\{\overline{1,\, M-1}\right\} \setminus \bigcup_{q=1}^{r} \mathscr{M}_q \tag{24}$$

by

$$\frac{\mu_k^{(\max)}\left(1 - \sum\limits_{l \in \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_l^{(\min)}\right)}{1 + \sum\limits_{i=1}^{M-1} \mu_i^{(\max)} - \sum\limits_{l \in \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_l^{(\max)} - \sum\limits_{i=1}^{M-1} \mu_i^{(\min)}} \geqslant \mu_k^{(\min)} \quad \forall\ k \in \left\{\overline{1,\, M-1}\right\} \setminus \bigcup_{q=1}^{r} \mathscr{M}_q \tag{25}$$

and

$$\mu_M^* = 1 - \sum_{k=1}^{M-1} \mu_k^* = 1 - \sum_{j \in \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_j^{(\min)} -$$

$$- \frac{1 - \sum\limits_{l \in \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_l^{(\min)}}{1 + \sum\limits_{i=1}^{M-1} \mu_i^{(\max)} - \sum\limits_{l \in \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_l^{(\max)} - \sum\limits_{i=1}^{M-1} \mu_i^{(\min)}} \cdot \sum_{k \in \left\{\overline{1,\, M-1}\right\} \setminus \bigcup\limits_{q=1}^{r} \mathscr{M}_q} \mu_k^{(\max)} . \tag{26}$$

After having mapped partially uncertain PM (12) over NN parameters into (19), (20) by (21) or (23) with $r \in \left\{\overline{1,\, M-2}\right\}$, (24), (26) by (25) there is applied the rule (7). If $\left|J_*\right| > 1$ then the set J_{**} is determined within (8) — (11) by the section 2 remarks.

4. Interval-evaluated DM elements

If the value of u_{ij} is defined intervally as (1), then firstly there comes the set $J = \left\{j_l\right\}_{l=1}^{N_J} \subseteq \left\{\overline{1,\, N}\right\}$ of

$$b_{ij} \leqslant u_i^{(\max)} \quad \forall\ i = \overline{1,\, M} \quad \text{and} \quad \forall\ j \in J \subseteq \left\{\overline{1,\, N}\right\}. \tag{27}$$

The results of cases with $\left|J\right| = \varnothing$ and $\left|J\right| = 1$ are the same as in the section 1. In the case with $\left|J\right| = N_J > 1$ the values $\left\{\left\{u_{ij}^{*}\right\}_{j \in J}\right\}_{i=1}^{M}$ are determined as

$$u_{ij}^{*} = \sqrt{a_{ij} b_{ij}} \quad \forall\ i = \overline{1,\ M} \ \text{ by } \ j \in J \subseteq \left\{\overline{1,\ N}\right\} \tag{28}$$

and standardized through (4) or (5). If PM over NN parameters is defined as weights (6), then there is straight out the rule (7) and are the rules (8) — (11) by the section 2. But when PM $\left\{\mu_i\right\}_{i=1}^{M}$ over NN parameters is partially defined uncertain as weights (12), then if it is mapped as (16) by (15) from (13), there is counseled it to apply the j_l-th TA with the weight $\breve{q}_{j_l}$ by $l = \overline{1,\ N_J}$. Otherwise, if partially uncertain PM (12) over NN parameters is mapped into (19), (20) by (21) or (23) with $r \in \left\{\overline{1,\ M-2}\right\}$, (24), (26) by (25), then the optimal TA number is determined over again with rules (7), (8) — (11) by the section 2 remarks.

5. Unknown PM over NN parameters

What-and-how-ever defined the value of u_{ij} is, the unknown PM over NN parameters directs to use (16) from (13). Then instead of the rules (7) — (11) the $\mathbf{K} = \left(k_{ij_l}\right)_{M \times N_J}$ game ensues: the second player optimal strategy (17) in this game shows that it is optimal to apply the j_l-th TA with the weight $\breve{q}_{j_l}$ by $l = \overline{1,\ N_J}$ inasmuch as (18) by (13) is true. Before this, certainly, the values $\left\{\left\{u_{ij}^{*}\right\}_{j \in J}\right\}_{i=1}^{M}$ for the submatrix $\mathbf{K} = \left(k_{ij_l}\right)_{M \times N_J}$ of the matrix $\mathbf{U}^{\mathrm{T}} = \left(u_{ij}\right)_{M \times N}$ are drawn from either checking (2) into (4) or (5), or checking (27) with calculations in (28) into (4) or (5).

Conclusion

The formulated rules for making decisions on either selecting the optimal TA number by (7) — (11) or combining TA within their set $\left\{a_j\right\}_{j=1}^{N}$ optimally by (17) for NN with its being controlled parameters $\left\{r_i\right\}_{i=1}^{M}$ are valid only under the expounded convention. By that the point evaluations of DM elements and PM over NN parameters in (6) should be strongly reliable. The reliability of interval-evaluated DM elements by (1) and partially uncertain PM (12) is higher, though. And, naturally, the endpoints of segments in $\left\{\left\{\left[a_{ij};\ b_{ij}\right]\right\}_{j=1}^{N}\right\}_{i=1}^{M}$ and $\left\{\left[\mu_i^{(\min)};\ \mu_i^{(\max)}\right]\right\}_{i=1}^{N}$ are not required to be as strongly reliable as the point evaluations [Park, 2010]. Actually, transition from point evaluations into interval evaluations often is explained with low reliability of pre-defined points $\left\{\left\{u_{ij}^{*}\right\}_{j=1}^{N}\right\}_{i=1}^{M}$ and weights $\left\{\mu_i^{*}\right\}_{i=1}^{M}$ in (6).

Unknown PM over NN parameters is a case, which is widespread at the start of functioning of NN, when potentialities, usage conditions and parameters of NN can be estimated only through a series of solving real practical problems. So no astonishing, that the decision to use PM over N_J TA $\left\{a_{j_l}\right\}_{l=1}^{N_J}$ as the distribution in (17) at the start of NN functioning appears the most reliable and substantiated, although the second player optimal strategy (17) must be realized properly through the running multistage process of training the NN.

Having finally solved the generalized choice task for TA, either the rules (7) — (11) or TA combination with weights in the second player optimal strategy (17) both ensure optimization of at least EPR, ATP, URV and other fundamental parameters of NN. That optimization is meant as the values of parameters of NN are converged to their averages [Moller, 1993], whose some convex combination is minimal or, for the case of the $\mathbf{K} = \left(k_{ij_l}\right)_{M\times N_J}$ game, is not greater than $\widehat{\mathbf{X}} \cdot \mathbf{K} \cdot \breve{\mathbf{Q}}^{\mathrm{T}}$.

Acknowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA Bulgaria www.ithea.org, and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine www.aduis.com.ua.

Bibliography

[Haykin, 1999] Haykin S. Neural Networks: A Comprehensive Foundation. — New Jersey: Prentice Hall, Inc., 1999.

[Borovikov, 2008] Borovikov V. P. Neural Networks. STATISTICA Neural Networks: Methodology and technologies of modern data analysis [reference edition] / Under the editorship of V. P. Borovikov. — [2nd edition, remade and expanded]. — Moscow: Hot Line — Telecom, 2008. — 392 p. (in Russian)

[Hagan, 1994] Hagan M. T., Menhaj M. Training feedforward networks with the Marquardt algorithm // IEEE Transactions on Neural Networks. — 1994. — Vol. 5, N. 6. — P. 989 — 993.

[Vogl, 1988] Vogl T. P., Mangis J. K., Rigler A. K., Zink W. T., Alkon D. L. Accelerating the convergence of the backpropagation method // Biological Cybernetics. — 1988. — Vol. 59. — P. 256 — 264.

[Chernorutskiy, 2005] Chernorutskiy I. G. Methods of decision making. — Saint-Petersburg: BHV-Petersburg, 2005. — 416 p. (in Russian)

[Trukhayev, 1981] Trukhayev R. I. Models of decision making under uncertainties. — Moscow: Nauka, 1981. — 258 p. (in Russian)

[Bolshakov, 2007] Bolshakov A. A., Karimov R. N. Methods of processing multidimensional data and time series. — Moscow: Hot Line — Telecom, 2007. — 520 p. (in Russian)

[Romanuke, 2011] Romanuke V. V. Generalized model of removing N partial uncertainties of the probabilistic type as continual antagonistic game on $(2N-2)$-dimensional parallelepiped with maximal disbalance minimization // Visnyk of Khmelnitskiy National University. Technical Sciences. — 2011. — N. 3. — P. 45 — 60. (in Ukrainian)

[Park, 2010] Park I., Amarchinta H. K., Grandhi R. V. A Bayesian approach for quantification of model uncertainty // Reliability Engineering & System Safety. — 2010. — Vol. 95, Issue 7. — P. 777 — 785.

[Moller, 1993] Moller M. F. A scaled conjugate gradient algorithm for fast supervised learning // Neural Networks. — 1993. — Vol. 6. — P. 525 — 533.

Author' Information

Vadim Romanuke *— c. t. s., associate prof., associate prof. of the Applied Mathematics and Social Informatics Department of Khmelnitskiy National University, 29016, Ukraine, Khmelnitskiy, Institutskaya str., 11; e-mail: romanukevadimv@mail.ru*

Major Fields of Scientific Research: Decision making, Game theory