A Representational MDL Framework for Improving Learning Power of Neural Network Formalisms

Alexey Potapov[1] , Maxim Peterson[1],

[1] St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Kronverkskiy pr. 49, 197101 St. Petersburg, Russia
{pas.aicv, maxim.peterson}@gmail.com

Abstract. Minimum description length (MDL) principle is one of the well-known solutions for overlearning problem, specifically for artificial neural networks (ANNs). Its extension is called representational MDL (RMDL) principle and takes into account that models in machine learning are always constructed within some representation. In this paper, the optimization of ANNs formalisms as information representations using the RMDL principle is considered. A novel type of ANNs is proposed by extending linear recurrent ANNs with nonlinear "synapse to synapse" connections. Most of the elementary functions are representable with these networks (in contrast to classical ANNs) and that makes them easily learnable from training datasets according to a developed method of ANN architecture optimization. Methodology for comparing quality of different representations is illustrated by applying developed method in time series prediction and robot control.

Keywords: representational minimum description length, artificial neural networks, machine learning

1 Introduction

Machine learning is one of the key paradigms in artificial intelligence, and ANNs pretend to be one of the universal approaches to learning. Such properties as self-adaptability, ability to generalize, and others are usually ascribed to ANNs as opposed to the traditional algorithms [1]. But the origin of these properties is rarely strictly explained. They are frequently grounded only by the presence of analogy between ANNs and real neural networks. At the same time, it can be said from bionic point of view that formal and biological neurons have almost nothing in common. Existing biophysically detailed neuronal models imitate their behavior much more precisely. However, learning is the very feature of neurons that has no plausible biophysical models, and thus cannot be really borrowed by ANNs.

As the result, there is no general theory of ANN learning. Particular training algorithms are proposed for all specific ANN architectures. Moreover, these

algorithms are quite classical and external to ANNs. Neuroglial networks with learning algorithms encoded within astrocytic nets [2] are rather interesting, because different learning rules can be made internal for them. However, the origin and structure of these rules remain unclear. Thereupon, recurring declarations about self-learning capabilities of ANNs and their distinctions from the traditional algorithms seem to be paradoxical. Apparently, ANNs don't solve the problem of machine learning. In particular, one difficult issue for them is overlearning, which has no good explanation in the ANN theory. Indeed, overlearning is typically prevented by restricting the training time. Nevertheless, popularity of ANNs is not accidental, but one should strictly describe their benefits in order to improve them further.

In this paper, ANNs are analyzed in the inductive inference framework. Induction is interpreted as construction of a model optimally describing the given data. Such the inductive inference methods include a model quality criterion, a model space, and a search strategy as their components [3].

The Kolmogorov complexity [4] with the set of all algorithms as the model space is known to be the correct basis for universal inductive inference [5]. However, the search problem is unsolvable here. Even computable analogs such as the Minimum Description Length (MDL) principle [6] and its counterparts [7, 8] are non-strictly applied in practice. Even so, they help to solve overlearning problem [9]. Here, information-theoretic analysis of ANNs is deepened on the base of the recent Representational MDL (RDML) principle [10], which takes into account that models are always constructed within certain representations. ANNs can also be interpreted as a specific algorithmic representation. This interpretation helped us to explain their good properties, which have an effect on all the component of induction.

The main contribution of this work is an extension of the information-theoretic approach from construction of a single ANN with some architecture to optimizing some ANN formalism as a data representation for a given set of learning tasks. Application of the RMDL framework for comparing quality of different ANN-based representations is illustrated with the use of a developed particular ANN type tested on the tasks of time series prediction and robot control.

2 Previous Works

Let's consider ANN learning as the induction problem, which requires introduction of a model space, a model quality criterion, and a search procedure.

The most widely used model selection criterion is MAP (maximum a posterior probability) calculated on the base of the Bayes' rule. However, its usage leads to the fundamental problem of prior probabilities [11], which cannot be inferred within statistical methods. These probabilities are sometimes ignored resulting in the maximum likelihood approach that leads to overlearning.

Theoretical solution of this problem was given quite long ago by several authors [6, 7, 12] on the base of Kolmogorov complexity introduced in the algorithmic information theory. Kolmogorov complexity of the given string (data) D is defined as

the length $l(H)$ of the shortest program (model) H for the Universal Turing Machine (UTM) U reproducing the given string $U(H)=D$:

$$K(D) = \min_{H}\left[l(H) \mid U(H) = D\right] . \quad (1)$$

Here, the model space contains all the algorithms. It appeared to be useful to divide each model into two parts: regular and random components. The best model can be defined using the conditional Kolmogorov complexity:

$$H^{*} = \arg\min_{H}\left[l(H) + K(D \mid H)\right] , \quad (2)$$

where $K(D \mid H) = \min_{R}\left[l(R) \mid U(HR) = D\right]$, and the string R is the input to the program H necessary to reconstruct the data D (or the data description encoded with the model H). This leads to the MDL principle, in which the best model is determined by minimizing the sum $l(H)+K(D|H)$. Using connection between information quantity and probability, one can write [13]: $-\log_2 P(H)=l(H)$ and $-\log_2 P(D|H)=K(D|H)$.

Prior probabilities of models are defined on the base of lengths of corresponding UTM programs leading to the universal priors, which are still under discussion [14]. Besides the search problem, these priors are inapplicable, because particular induction tasks require large amount of prior information influencing the model selection criterion. For this reason, the MDL principle is used in its inexact verbal form, and heuristic coding schemes are contrived in order to compute description lengths.

In particular, the MDL principle was applied in this heuristic way to solve the problem of ANN architecture optimization [15-17]. In these works, components of the description length are calculated within some ANN coding schemes. Partial solution of the overlearning problem is achieved here, because increase of the model precision at the cost of increase of its complexity (the number of neurons and connections) is allowed only in the case, when it decreases the total description length [9, 18].

However, heuristic coding schemes are ungrounded. They introduce non-optimal inductive bias into model selection criteria, and specify arbitrary model space containing regularities, which can be inadequate for the given learning task. For example, activation functions are rarely considered as model components that also should be optimized. Moreover, ANNs with restricted architecture (e.g., radial basis function networks [19] or multilayer perceptrons [9]) are typically considered. The possibility of inclusion of ANN formalisms themselves into the optimization process on the base of information-theoretic criteria has not been considered yet, despite the fact that such the optimization can be more significant than optimization of a single ANN within particular formalisms.

3 Methodology

Particular ANN formalisms define algorithmically incomplete model spaces, which adequacy is not analyzed. Consequently, regularities in the data can be inexpressible within the selected formalism leading to bad learning capabilities. This fact is typically ignored because of the well-known proof that ANNs are universal

approximators [20]. This proof is cited even in the papers devoted to the MDL-based ANNs [17]. This contradiction arises from a lack of understanding that approximation of any function with preset accuracy is insufficient in machine learning. If some regularity in the data is not expressible in the given model space, overlearning cannot be completely avoided: Ptolemy's epicycles can approximate planetary orbits only with precision restricted by observation errors, because they don't capture underlying regularities and thus cannot generalize in contrast to Kepler's model. Arbitrary regularities are expressible only with the use of the algorithmically complete model space, but the search in this space is currently unachievable. Thus, narrowing the model space is unavoidable in applied tasks. But this just means that special attention should be paid to the problem of the model space selection (also, in the case of ANNs). Heuristic coding schemes not only specify restricted model spaces, but also introduce inductive bias assigning different complexities for models.

The formalized notion of representation replaces heuristic coding schemes within the RMDL principle that focuses attention on the selection of an optimal representation for any given class of inductive inference tasks. This principle was applied to construction of "essentially" learnable computer vision methods [21], but its significance can be extended on the fundamental issues of machine learning.

It can be noted that ANNs of the same type are trained on different data samples independently. Thus, we can set a mass problem of induction. Let a set $\mathbf{D}=\{D_i\}$ of data samples is given, and the best model for each sample D_i should be constructed independently. Data samples can contain mutual information (at least in the form of similar regularities). Consequently, the sum of their individual complexities $K(D_i)$ will be much higher than the complexity of their concatenation $K(D_1 \ldots D_n)$.

Thus, the universal criterion (1) is not applicable in this case, because individual models of D_i will be much worse than their common model. However, we can include some prior information S into the method developed specially for the mass problem **D**, and each data sample D_i will be described conditionally with the given S.

Independent descriptions of D_i with the given S can be almost as efficient as their common description. Here, one should require that $(\forall D \in \mathbf{D})(\exists H, R) U(SHR) = D$. Such program S can be called *representation*, within which a model H can be constructed for each data D. Optimal representation can be chosen using

$$K(D_1 D_2 \ldots D_n) \approx \min_S \left(\sum_{i=1}^{n} K(D_i \mid S) + l(S) \right), \; S^* = \arg\min_S \left(\sum_{i=1}^{n} K(D_i \mid S) + l(S) \right). \qquad (3)$$

This gives the mentioned RMDL principle for choosing the best representation for the given set **D**, which minimizes the summed description length of all data samples and the representation itself. Each ANN formalism or architecture has its own executing algorithm, which precisely corresponds to the definition of the representation. These algorithms aren't affected during ANN learning on the specific data D_i, but they can be optimized for each mass problem **D** using the criterion (3).

4 A Model Representation with Dynamic ANNs

The criterion (3) gives quantitative evaluation of representation quality, but it can also be estimated qualitatively by analysis of a set of expressible regularities. For example, outputs of a feedforward network with linear activation function are linear combinations of its inputs independent from the number of hidden neurons. Thus, such the ANN can represent only linear models. Because of this severe limitation, nonlinear activation functions are typically introduced. However, let's consider linear dynamic (recurrent with continuous time) ANN containing M neurons, which activities $x_i(t)$ follow the law:

$$x_i'(t) = \frac{dx_i(t)}{dt} = \sum_{j=1}^{M} w_{ji} x_j(t) \ , \tag{4}$$

where w_{ji} are connection weights constituting a matrix $\mathbf{W}$.

Starting from some initial values $x_i(0)$, activities $x_i(t)$ will evolve producing some functions as an output. Well-known general solution of the system of homogeneous linear differential equation has the form $\exp(\mathbf{W}t)$ corresponding to the mixture of harmonic, exponential, and polynomial functions, which appear to be representable by such ANNs. Consequently, even linear dynamic ANNs can be called "universal approximators" that can fit any regular function. One interesting application is time series forecasting, in which the data $D=\{\mathbf{y}(t_1),\ldots,\mathbf{y}(t_n)\}$ is given, where the values $\mathbf{y}(t_i)=(y_1(t_i),\ldots, y_N(t_i))$ of N-dimensional vector are observed at some moments of time $t_i \in [0,T_{\max}]$. The task is to predict values $\mathbf{y}(t)$ for $t > T_{\max}$.

Such the connection weights w_{ij} and such the initial activities $x_i(0)$ should be found that the activities $x_i(t)$ are most precisely correspond to the values $y_i(t)$. Naïve approach leads to minimization of the mean-square error:

$$E^2 = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \left(y_j(t_i) - x_j(t_i) \right)^2 \ . \tag{5}$$

The number of neurons M should be not less than the dimension N of the vector $\mathbf{y}$, but it can be larger. In this case, additional neurons can be treated as hidden dynamic variables. They are not included into the MSE criterion (5). Apparently, increase of the number of additional neurons will result in decrease of the MSE as well as in overfitting. In accordance with the MDL principle, the model complexity should also be taken into account in addition to the description length of the data encoded within the model that can be estimated as $nN\log_2 E$ (accurate to a constant).

ANN model description includes information about the number of neurons (roughly $\log_2 M$ bits), established connections (roughly $\log_2 K + \log_2 C(K, M^2)$ bits), their weights ($0.5K\log_2 n$ bits), and initial values of activity ($0.5M\log_2 n$ bits). Total MDL criterion for the ANN with M neurons and K connections can be roughly estimated as

$$L = nN \log_2 E + \log_2 M + \log_2 K + \log_2 C_{M^2}^{K} + 0.5(M+K)\log_2 n \ . \tag{6}$$

To find the best ANN, one should consider and optimize ANNs with different number of neurons and connections. In order to reduce computational complexity of this process, we utilized an iterative scheme, in which new neurons are consequently added and redundant connections are removed if these operations result in reduction of the description length criterion (6). We considered and implemented a combination of several optimization techniques (stochastic gradient descent, genetic algorithms, and simulated annealing) for optimizing ANNs with fixed architecture. Unfortunately, detailed analysis of this search problem goes beyond the scope of the paper.

Experimental validation of the developed algorithm showed that low-sized ANNs are automatically chosen if the data D is generated using combinations of harmonic, polynomial, and exponential functions. These ANNs extrapolated the given functions with relative errors less than 2% on interval $[T_{\max}, 2T_{\max}]$. Such precision is difficult to achieve with the use of conventional ANNs with nonlinear activation function, because all these elementary functions are not simultaneously representable by such ANNs. But they are representable by linear dynamic ANNs, which can extract these regularities from few data points and can make good predictions following from their high efficiency in terms of the RMDL principle.

Even linear dynamic ANNs can be rather useful, but still they define very restricted model space. Only extension of representable regularities can help to increase their learning power (and extrapolation capabilities) further. Thus, some type of nonlinearity should be introduced. However, typically used nonlinear activation functions violate the representability of the mentioned elementary functions.

In this context, it is not surprising that hybrid systems gain growing popularity. They include methods with different representable regularities, e.g., nonlinear ANNs and linear auto-regressive models [1]. However, the search problem in the heterogeneous model spaces is more difficult. Here, we propose a homogeneous representation, within which both linear and nonlinear models can be described.

It is natural to construct such the extension of the linear ANNs that will also incorporate models of nonlinear dynamic theory. These models are typically described with differential equations, which can be linear or can contain nonlinearity.

We propose to introduce optional nonlinearity by adding connections from neurons to other connections ("synapses on synapses"). The 2nd-order connections exert nonlinear influence on signals propagating through ordinary connections, but don't change the connection weights themselves. Neurophysiologic prototypes are the modulating neurons. These connections can be introduced in the following way. Consider the system containing 3 neurons shown on the figure 1a.

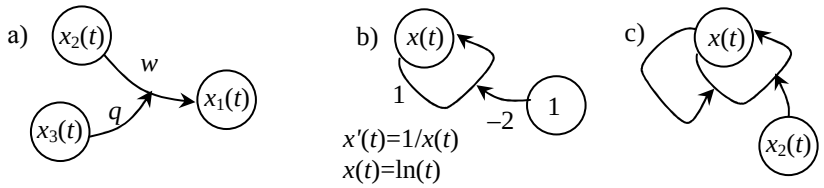

Fig. 1. General form of "connection on connection" (a); minimal (b) and automatically constructed (c) ANNs reproducing logarithmic function.

Let the postsynaptic neuron activity be described by the following equation:

$$x_1'(t) = w x_2^{q x_3(t)+1}(t) \ . \tag{7}$$

ANNs with this structure can simply represent power functions as well as logarithmic functions (Fig. 1b). Chaotic modes of the dynamic systems are also representable. For example, a network reproducing the Lorenz attractor can be easily (manually) constructed. It can be seen that these ANNs define the wide model space containing perspicuous regularities. The mentioned above learning (search) procedure can be applied to these extended dynamic ANNs almost without modifications.

Our experiments showed that simple non-chaotic functions are automatically recovered rather reliably. However, even for the basic elementary functions the best network is not always constructed, because of complexity of the search problem. For example, the network on the figure 1c was reconstructed for the logarithmic function. This network contains two unnecessary connections. Nevertheless, its extrapolation error on the doubled time interval appeared to be less than 1%.

The search problem becomes very difficult in the case of chaotic time series. Although different chaotic sequences are representable, the necessary ANN can hardly be found by the direct approximation of the data points. Apparently, this difficulty is connected with instability of chaotic trajectories of the dynamic systems that results in very non-monotonic landscape of the quality criterion under optimization. Since the individual trajectories of chaotic dynamic systems are almost irreproducible, it is more reasonable to reconstruct their invariant measure. This also can be done within the RMDL framework applied to the dynamic ANNs, if one uses such the representation that encodes the values $\mathbf{y}(t_i)$ in accordance with the hypothesized invariant measure (defined by a particular ANN) instead of encoding the deviations of the output of this ANN from $\mathbf{y}(t_i)$. Unfortunately, discussion of this method goes beyond the scope of the paper and requires additional research.

5 Experiments

At first, the developed ANN type and the method for its optimization were tested on the well-known Wolf annual sunspot time series. Wolf numbers till 1979 were used as the training sample. The constructed ANN contained 4 neurons, 11 connections, and 2 second-order connections. Obtained prediction MSE value for 1980–1988 years equals to 220. The other methods mentioned in [18] show the MSE between 214 and 625. Thus, the proposed ANN type is usable. Judging by the prediction accuracy and the ANN size, it can be concluded that overlearning is avoided.

Then, we performed comparison of the ANN-based representations on mass problems. These representations included linear ANNs, ANNs with non-linear activation function, and ANNs with second-order connections. The data samples D_i were taken from a number of financial time series. The complexity of representations under comparison is similar, so we ignored $l(S)$ term in the criterion (3); however, this term can be crucial in more advanced cases in order to avoid overlearning on the level

of representations. Table 1 shows the value of the RMDL criterion (divided on the number of data samples), and the relative prediction error (10 points ahead).

Table 1. The values of the RMDL criterion and the relative error for different types of ANNs

ANN type	RMDL, bits	error, %
Linear	651	15,8
Activation function	617	10,1
2^{nd}–order connections	608	9,9

Although we obtained an agreement between the short-term prediction precision and the RMDL criterion in average, one can agree with the statement: "MSE and NMSE are not very good measures of how well the model captures the dynamics" [18]. One can hope that the MDL criterion is the better measure of how well the model captures the underlying regularities, and the RMDL principle helps to extend this criterion on representations.

Another considered mass problem for the ANN-based representations is the robotic control. Here, we used a wheel robot with two motors and sonar for measuring the distance to the obstacles. In this case, additional sensory neuron was included into the network. The training data samples D_i were obtained by recoding the sensory input and the motor commands from the robot under the manual control used for obstacle avoidance. It should be pointed out that the quality criterion included only the approximation precision of motor commands (not the sensory input), and such the network could be constructed that directly approximates commands ignoring the sensory data. An example of successful extrapolation of a sequence of motor commands is given on the figure 2. It can be seen that the prediction results are relatively successful (and the robot controlled by the trained network performs free roaming with adequate reaction to obstacles). This implies that such the ANN was constructed, in which the sensory neuron was connected to the rest network in such the way that it helped to increase both approximation quality (in terms of the MDL criterion) and prediction accuracy.

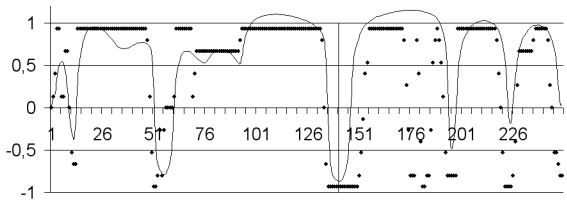

Fig. 2. Robot control commands reproduction (extrapolation is after vertical solid line).

The results of estimation of the RMDL criterion on a set of data samples were 1826 bits for linear ANNs, 1793 bits for ANNs with the nonlinear activation function, and 1798 bits for ANNs with 2^{nd}-order connections.

In this case, the extrapolation precision is meaningless, because human chooses direction of movement during the obstacle avoidance randomly. Sometimes, the choice made by an ANN precisely corresponds to the human choice; but it cannot be guaranteed. So, one can rely only on the RMDL criterion that will hopefully reflect general adequacy of the robot movement.

The RMDL criterion values for both ANN types are compatible, but the ANN with the 2^{nd}-order connections showed more interesting behavior. This result is understandable, because the control command sequences (as non-smooth functions) are not expressible within all the ANN-based representations under comparison.

5 Conclusions

The problem of comparison of learning power of ANN formalisms was considered as the optimization of model representations in the tasks of inductive inference. Model spaces defined by different ANN types are subsets of the set of all algorithms, so they can be optimized within the approach based on the algorithmic information theory. The simplicity of the descriptions of regularities, which presence is expected in the datasets, specifies the inductive bias defining prior probabilities of corresponding models and thus necessary amount of information for their reconstructions.

Such new modification of the ANN formalism was proposed, within which regularities corresponding to the elementary functions are representable as opposed to the ANNs with nonlinear activation functions. The method for optimization of such ANNs was developed. The number of neurons and connections between them is also controlled by the information-theoretic criterion in order to avoid overlearning. The methodology based on the RMDL principle for comparing quality of different representations was proposed and experimentally verified with the use of the developed method on tasks of time series prediction and robot control. Different representations appeared to be more efficient depending on the task.

Our research showed that the RMDL principle can be used to compare the quality of representations while solving inductive mass problems. However, efficiency of representations cannot be reduced only to their RMDL quality. Even in the case of very simple nonlinear representations, it is very difficult to find the best model even if it exists in the specified model space. Representations should give not only the optimal inductive bias for some mass problem, but also should make the model search process more efficient. The speed priors are known [22], but one can expect that they also depend on the representation. Thus, the speed priors can be extended with the notion of representation, or equivalently the RMDL principle should incorporate the model search speed.

References

1. Khashei, M., Bijari, M.: An Artificial Neural Network (p, d, q) Model for Timeseries Forecasting. Expert Systems with Applications. 37, 479–489 (2010)

2. Pazos, A.B.P., Gonzalez, A.A., Pazos, F.M.: Artificial NeuroGlial Networks. In: Dopico, J.R.B., Calle, J.D., Sierra, A.P. (eds.) Encyclopedia of Artificial Intelligence. pp.167–171. Hershey, New York (2009)
3. Baxter, R.A.: Minimum Message Length Inference: Theory and Applications. Ph.D. thesis. Department of Computer Science. Monash University. Clayton. Australia. (1996)
4. Kolmogorov, A.N.: Logical Basis for Information Theory and Probability Theory. IEEE Trans. Inform. Theory. IT-14, 662–664 (1968)
5. Solomonoff, R.: Algorithmic Probability Solve the Problem of Induction. Oxbridge Research, P.O.B. 391887, Cambridge, Mass. (1997)
6. Rissanen, J.J.: Modeling by the Shortest Data Description. Automatica-J.IFAC. 14, 465–471 (1978)
7. Wallace, C.S., Boulton D.M.: An Information Measure for Classification. Comput. J. 11, 185–195 (1968)
8. Vitanyi, P., Li, M. Ideal MDL and Its Relation to Bayesianism. Proceeding of ISIS: Information, Statistics and Induction in Science. 282–291 (1996)
9. Zhao, Y., Small, M.: Minimum Description Length Criterion for Modeling of Chaotic Attractors with Multilayer Perceptron Networks. IEEE Transactions on Circuits and Systems I. 53(3), 722–732 (2006)
10. Potapov, A.S.: Comparative Analysis of Structural Representations of Images Based on the Principle of Representational Minimum Description Length. Journal of Optical Technology. 75(11), 715–720 (2008)
11. Li, M., Vitanyi, P.: Philosophical Issues in Kolmogorov Complexity (Invited Lecture). In Kuich W. (eds.) Proc. on Automata, Languages and Programming. vol.623, 1–15 (1992)
12. Solomonoff, R. A Formal Theory of Inductive Inference, parts 1-2. Information and Control. 7, 1–22, 224-254 (1964)
13. Vitanyi, P., Li, M.: Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity. IEEE Trans. on Information Theory. 2, 446–464 (2000)
14. Wood, I., Sunehag, P., Hutter, M.: (Non-)Equivalence of Universal Priors. Solomonoff 85th Memorial Conference. abs/1111.3854 (2011)
15. Lappalainen, H.: Using an MDL-Based Cost Function with Neural Networks. Proc. IJCNN. 2384–2389 (1998)
16. Wang, J.-S., Hsu Y.-L.: An MDL-Based Hammerstein Recurrent Neural Network for Control Applications. Neurocomputing. 74, 315–327 (2010)
17. Molkov, Y.I., Mukhin, D.N., Loskutov E.M., Feigin, A.M., Fidelin, G.A.: Using the Minimum Description Length Principle for Global Reconstruction of Dynamic Systems from Noisy Time Series. Physical Review E. 80, 046207(1-6) (2009)
18. Small, M., Tse C.K.,: Minimum Description Length Neural Networks for Time Series Prediction. Physical Review E. 66, 066701(1–12) (2002)
19. Leonardis, A., Bischof, H.: An Efficient MDL-Based Construction of RBF Networks. Neural Networks. 11(5), 963–973 (1998)
20. Hornik, K., Stinchcombe, M., White, H.: Multilayer Feedforward Networks are Universal Approximators. Neural Networks. 2(5), 359–366 (1989)
21. Potapov, A.S., Malyshev, I.A., Puysha, A.E., Averkin A.N.: New Paradigm of Learnable Computer Vision Algorithms Based on the Representational MDL Principle. Proceeding of SPIE. 7696, 769606 (2010)
22. Schmidhuber, J.: The Speed Prior: A New Simplicity Measure Yielding Near-Optimal Computable Predictions. Proceedings of the 15th Annual Conference on Computational Learning Theory. Sydney. Australia. LNAI. pp.216–228, Springer. (2002)