ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

Extreme Minimal Learning Machine

Tommi Kärkkäinen

University of Jyvaskyla,
Faculty of Information Technology, Finland
tommi.karkkainen@jyu.fi

Abstract. Extreme Learning Machine (ELM) and Minimal Learning Machine (MLM) are nonlinear and scalable machine learning techniques with randomly generated basis. Both techniques share a step where a matrix of weights for the linear combination of the basis is recovered. In MLM, the kernel in this step corresponds to distance calculations between the training data and a set of reference points, whereas in ELM transformation with a sigmoidal activation function is most commonly used. MLM then needs additional interpolation step to estimate the actual distance-regression based output. A natural combination of these two techniques is proposed here, i.e., to use a distance-based kernel characteristic in MLM in ELM. The experimental results show promising potential of the proposed technique.

1 Introduction

Kernels or basis functions have a central role in machine learning. The appearance of Radial Basis Function networks (RBFN) (e.g., [1]) made it clear that universal approximation property of a neural network technique does not need a fully adaptable basis. With a priori fixed location and scatter parameters of radial basis functions, one could construct nonlinear approximators of unknown functions. Actually already in [2] something similar was suggested for the Multilayered Perceptron (MLP, e.g. [7]): first optimize all weights using the whole data and then freeze the hidden layer weights in the nonlinear cross-validation, by only adapting the weights in the outer layer.

In MLP and in deep learning (see [8] and articles therein), we might have a large pool of adaptation in the deeply layered basis. However, the Extreme Learning Machine (ELM) as proposed by Huang et al. [4, 5], established one of the key randomized neural network frameworks without kernel adaptation [6]. Probabilistic convergence analysis of ELM was presented in [3], where the necessity of the repeated sampling of the sigmoidal kernel and the advantage of the weight decay (ridge regression) were concluded.

Recently, a new supervised learning method, called Minimal Learning Machine (MLM, [9, 10]), emerged. MLM is based on the idea of the existence of a mapping between the geometric configurations of points. This configuration is seeked using a distance-based regression technique, where both input and output data are sampled and two distance-matrices are formed between these reference

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

points and the whole training data. For the output of any test observation, an interpolation problem using the computed distances needs to be solved. Both ELM and MLM contain only one hyperparameter: the size of the hidden layer in ELM or number of reference points in MLM. Actually both definitions are typically proportional to the number of observations available in the training set [4, 5, 10, 11].

This paper proposes and describes a natural combination of ELM and MLM: Extreme Minimal Learning Machine (EMLM). The technique uses distance-based kernel according to MLM to generate a random basis for a nonlinear approximation. Then, similarly to ELM (and to many other basically linear techniques [13]), regularized least-squares problem is solved to recover the matrix of weights. Compared to MLM, we then omit solution of the optimization problem to estimate the actual distance-regression based output.

2 The Extreme Minimal Learning Machine Method

Let $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^n$ and $\mathbf{y}_i \in \mathbb{R}^k$, be the training data of input-output samples. In ELM, we associate for each bias-enlarged input $\tilde{\mathbf{x}}_i = [1\ \mathbf{x}_i^T]^T \in \mathbb{R}^{n+1}$ the sigmoidal basis function $\mathbf{h}_i = \frac{1}{1+\exp(-\mathbf{G}\tilde{\mathbf{x}}_i)}$, where $\mathbf{G} \in \mathbb{R}^{m\times(n+1)}$ with $(\mathbf{G})_{ij} \in \mathcal{U}([-1, 1])$ (uniform distribution on $[-1, 1]$). Here m denotes the number of basis functions. To determine weights for the linear combination of this basis, let us consider the regularized least-squares optimization problem

$$\min_{\mathbf{V}\in\mathbb{R}^{k\times m}} \mathcal{J}(\mathbf{V}), \text{ where } \mathcal{J}(\mathbf{V}) = \frac{1}{2N}\sum_{i=1}^{N} \|\mathbf{V}\mathbf{h}_i - \mathbf{y}_i\|_2^2 + \frac{\alpha}{2m}\sum_{i=1}^{k}\sum_{j=1}^{m} |\mathbf{V}_{ij}|^2. \quad (1)$$

The coefficients $\frac{1}{N}$ and $\frac{1}{m}$ in $\mathcal{J}(\mathbf{V})$ normalize the components with respect to the amount of data and the size of basis, respectively. $\alpha > 0$ is the Tykhonov regularization/weight decay parameter, which restricts the increase of the weights and, by enforcing strict coercivity, guarantees the unique solvability of (1). The solution $\mathbf{W} \in \mathbb{R}^{k\times m}$ of the problem satisfies as

$$\frac{1}{N}(\mathbf{W}\mathbf{H} - \mathbf{Y})\mathbf{H}^T + \frac{\alpha}{m}\mathbf{W} = \mathbf{0}, \quad (2)$$

where $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^N \in \mathbb{R}^{m\times N}$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^{k\times N}$.

In the minimal learning machine, construction of distance-based random basis starts by a selection of m reference points $\mathbf{R} = \{\mathbf{r}_i\}_{i=1}^m$ such that, for all i, $\mathbf{r}_i = \mathbf{x}_j$ for some j. Hence, $\{\mathbf{r}_i\}_{i=1}^m$ is a random subset of input vectors. With the two set of vectors, the set of reference points and the whole set of input vectors, we define the matrix $\mathbf{H} \in \mathbb{R}^{m\times N}$ as

$$(\mathbf{H})_{ij} = \|\mathbf{r}_i - \mathbf{x}_j\|_2. \quad (3)$$

A method referred as *Extreme Minimal Learning Machine*, EMLM, is obtained when (3) is used in (2).

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Dataname	Train$N$	N	NT	n	k	R-MCP
Overlap	$\lfloor N/2 \rfloor$	3 960	990	2	4	16.3
Outdoor	$\lfloor N/2 \rfloor$	2 400	1 600	21	40	29.0
COIL	$\lfloor N/2 \rfloor$	1 800	5 400	20	100	3.5
Satimage	$\lfloor N/2 \rfloor$	4 435	2 000	36	6	-
Letter	$\lfloor N/2 \rfloor$	16 000	4 000	16	26	3.0
USPS	N	7 291	2 007	256	10	4.6
Isolet	N	6 238	1 559	617	26	3.8
MNIST	$\lfloor 3N/4 \rfloor$	60 000	10 000	666	10	5.2

Table 1. Description of test datasets.

Let us briefly comment the proposed method. The basic ingredient is that the sigmoidal transformation of input vectors in ELM is replaced with the distance-based kernel underlying MLM. When compared to the classical forms of sigmoidal or gaussian basis, the distance-based form is mathematically of very different nature. In particular, the nonlinearity in (3) is not based on any transformation with a nonlinear function. Moreover, the proposed approach has certain lazy flavor, because output for a test observation needs re-estimation of the kernel. Therefore, it will be interesting to assess the approximation properties of such an approach in what follows.

3 Experiments

Reference versions of the techniques in Section 2 were implemented with Matlab (R2015b). Datasets for the tests mostly originate from the UCI machine learning repository. Because of the incremental flavor of MLM and for comparison, we used almost the same datasets as in [14]. The datasets are described in Table 1. There, Train$N$ refers to the maximum size of hidden layer m (wrt N), N denotes the number of training and NT the test observations, n refers to the dimension of the data, k gives number of classes, and "R-MCP" includes the best reference result from [14] as MisClassifications in Percentages, MCP.

Output vectors were formed using 1-of-k encoding and $\alpha = 10^{-6}$ was fixed throughout. As preprocessing, we removed constant variables and min-max scaled all features into $[0, 1]$. We also realized the Leave-One-Out cross-validation technique (TR-PRESS in [15], with the suggested efficient implementation) to identify the only metaparameter, m, needed in both techniques. Note that we then measure the least-squares approximation error of the methods on a complete different scale compared to MCP.

For the results, the minimum value of m was set to 50 and it was then incremented with stepsize 10 for the first five datasets, with small number of features, that were trained until $\lfloor N/2 \rfloor$. The incremental training was stopped before $\lfloor N/2 \rfloor$ when either 99.9% training accuracy was obtained or ten increments without training error improvement were faced by both of the two tech-

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

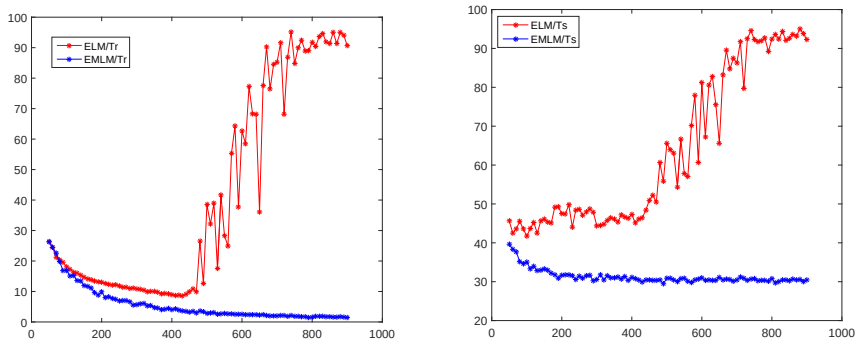

Fig. 1. Outdoor: training errors (left) and test errors (right).

niques. The three larger dimensional datasets were trained until N or $\lfloor 3N/4 \rfloor$, with the incrementation stepsize 1% of N.

Results of the experiments are given in Table 2. There, for ELM and EMLM, we report the smallest training set MCP-error "Tr", the LOO-CV/TR-PRESS-error "LO", and the separate test set error "Ts". All these are searched separately from all the tested values of m in each experiment, and this value is given before the error (i.e., format "m: Err").

Behavior of training errors and test errors for the two techniques are further illustrated in Figs. 1 and 2. They illustrate the common behavior: for Overlap, Outdoor, COIL, Satimage, Letter, and Isolet, the training of ELM deteriorated when m was increased yielding to the increase of the test error as well. This did not happen to USPS and MNIST. For EMLM, training failed for some values of m only for Letter and otherwise, for the tested values, both the training and test error showed generally a nonincreasing trend. Hence, the training error is readily useful for EMLM to search for an appropriate value of m. This is

Data	ELM/Tr	EMLM/Tr	ELM/LO	EMLM/LO	ELM/Ts	EMLM/Ts
Overlap	80: 29.2	300: 14.9	50: 2.9e-3	90: 1.3e-5	80: 28.9	180: 16.0
Outdoor	430: 8.5	790: 1.5	80: 8.4e-8	180: 7.1e-5	100: 41.8	490: 29.5
COIL	330: 12.8	730: 3.5	50: 1.7e-7	130: 1.5e-4	230: 20.0	840: 5.7
Satimage	1380: 4.0	2140: 2.6	260: 9.3e-8	530: 1.4e-5	480: 11.3	1920: 8.1
Letter	690: 7.2	1470: 4.8	140: 6.3e-8	530: 9.9e-5	690: 11.5	1480: 8.9
USPS	4649: 0.1	7131: 0.0	3481: 1.3e-6	707: 4.4e-5	2678: 4.2	5963: 4.1
Isolet	3515: 0.0	5972: 0.0	50: 4.6e-9	428: 8.8e-5	1688: 6.3	4460: 2.9
MNIST	21050: 0.0	40250: 0.3	29450: 2.7e-6	3650: 9.7e-6	21050: 1.8	28250: 1.6

Table 2. Experimental results.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

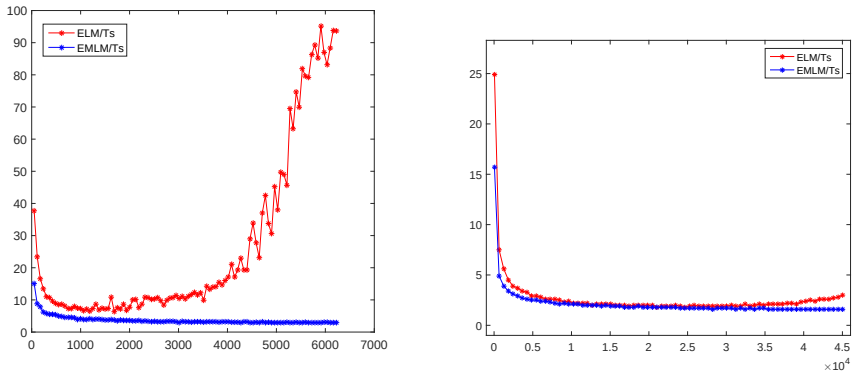

Fig. 2. Test errors for Isolet (left) and MNIST (right).

important, because we were not able to identify suitable values of m for either of the techniques with LOO-CV.

EMLM training always gave smaller test error level compared to ELM. Compared to the R-MPC values in Table 1, the results were especially competitive for the largest three problems.

4 Conclusions

In this work, a combination of two scalable machine learning techniques, ELM and MLM, with random kernels was proposed. The straightforward idea was to use the distance-based kernel from MLM in ELM-like regularized least-squares framework. The obtained results indicate that the distance-based random basis is a viable option for a random kernel with such techniques. The EMLM appeared more stable than ELM in the experiments and, hence, it seems easier to find a good value of the only metaparameter m. This issue and more experiments should be carried out to understand better the behavior of the proposed technique. The results of EMLM for the three largest datasets, USPS, Isolet and especially MNIST, showing no indication of overlearning actually explain the success of deep learning with enriched learning data for these problems: the noise in classification is negligible so only the flexibility of basis for maximally descriptive learning data matters for the test result accuracy.

Acknowledgments

The work of TK has been supported by the Academy of Finland from the projects 311877 (Demo) and 315550 (HNP-AI). The author gratefully acknowledge the role of ESANN in facilitating a collaborative platform with other active research groups of the two methods [11, 12].

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

References

1. Powell, M. J. D.: Radial basis functions for multivariable interpolation: a review. Algorithms for Approximation. Clarendon Press, Oxford, 143–167, 1987
2. Kwok, T.-Y. and Yeung, D.-Y.: Efficient cross-validation for feedforward neural networks. In Proceedings of the IEEE International Conference on Neural Networks, 5: 2789–2794, 1995
3. Xia, L., Lin, S., Fang, J., and and Xu, Z.: Is extreme learning machine feasible? A theoretical assessment (Parts I–II). IEEE Transactions on Neural Networks and Learning Systems, 26(1): 7–34, 2015
4. Huang, G.-B., Zhu, Q.-Y., and Siew, C.-K.: Extreme learning machine: theory and applications. Neurocomputing, 70(1): 489–501, 2006
5. Huang, G.-B., Zhou, H., Ding, X., and Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(2): 513–529, 2012
6. Gallicchio, C., Martin-Guerrero, J. D., Micheli, A., and Soria-Olivas, E.: Randomized Machine Learning Approaches: Recent Developments and Challenges. Proceedings of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN-2017, 77-88, 2017
7. Kärkkäinen T. and Heikkola E.: Robust formulations for training multilayer perceptrons. Neural Computation, 16, 837–862, 2004
8. Angelov, P. and Sperduti, A.: Challenges in Deep Learning. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2016, 489–496, 2016
9. Souza Junior, A. H., Corona, F., Miché, Y., Lendasse, A., Barreto, G., and Simula, O.: Minimal Learning Machine: A New Distance-Based Method for Supervised Learning. Proceedings of the 12th International Work Conference on Artificial Neural Networks (IWANN'2013), 7902, 408–416, 2013
10. Souza Junior, A. H. S., Corona, F., Barreto, G. A., Miche, Y., and Lendasse, A.: Minimal Learning Machine: A novel supervised distance-based approach for regression and classification. Neurocomputing, 164:34–44, 2015
11. Gomes, J. P. P., Mesquita, D. P. P., Freire, A. L., Souza Junior, A. H., and Kärkkäinen, T.: A robust minimal learning machine based on the M-estimator. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2017, 383–388, 2017
12. Akusak, A., Saarela, M., Kärkkäinen, T., Björk, K.-M., and Lendasse, A.: Mislabel detection of Finnish publication ranks. In Proceedings of the 8th International Conference on Extreme Learning Machines - ELM2017 (9 pages, to appear)
13. Friedman, J., Hastie, T., and Tibshirani, R.: The elements of statistical learning. Vol. 1. New York: Springer series in statistics, 2001
14. Losing, V., Hammer, B., and Wersing, H.: Choosing the Best Algorithm for an Incremental On-line Learning Task. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2016, 369–374, 2016
15. Miche, Y., van Heeswijk, M., Bas, P., Simula, O., and Lendasse, A.: TROP-ELM: A double-regularized ELM using LARS and Tikhonov regularization. Neurocomputing 74: 2413–2421, 2011