ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Assessment of Feature Saliency of MLP using Analytic Sensitivity

Tommi Kärkkäinen

Department of Mathematical Information Technology
P.O. Box 35, 40014 University of Jyväskylä - Finland

Abstract. A novel technique to determine the saliency of features for the multilayer perceptron (MLP) neural network is presented. It is based on the analytic derivative of the feedforward mapping with respect to inputs, which is then integrated over the training data using the mean of the absolute values. Experiments demonstrating the viability of the approach are given with small benchmark data sets. The cross-validation based framework for reliable determination of MLP that has been used in the experiments was introduced in [1, 2].

1 Introduction

Feature or variable selection (FS) refers to selecting the most important input variables for a data-based model. According to [3], the objective of feature selection is three-fold: improving the prediction performance of the model, providing faster and more cost-effective models, and providing a better understanding of the data generation process (simplifying the data collection process later). In [4], a given set of features is proposed to be divided into *irrelevant, weakly relevant,* and *strongly relevant* subsets. To determine such classification between features is, in general, a search problem, and many techniques and approaches for this purpose can be distinguished [5]. However, it was already argued in [6] that finding an optimal subset of features is usually an intractable problem, even NP-hard. The two main approaches (with the so-called embedded approach in between) for feature selection are the filter approach and the wrapper approach [4], where the current work is of the latter type, i.e., it involves the MLP model in the assessment of the features.

Feature selection techniques have their origins in variable selection for regression related to statistical modelling. This is described, for example, in [7] with a clear description of the starting point: for linear models with parameters estimated using least squares fit, adding more variables to the model means trading off reduced bias against increased variance. For the MLP, different variable contribution (ranking) methods in an ecological prediction context were tested in [8], using BP as the training method and a fixed size (10-5-1) network. The seven methods compared were i) the Partial Derivatives' (PaD) method that uses derivatives of the output with respect to input (see [9]), ii) the Connection Weights method, iii) the Input Perturbation method, (iv) the Profile method, and v-vii) three versions of the Classical and Improved Stepwise methods. For predicting the density of brown trout spawning redds using habitat characteristics, the PaD method was found as the most useful of the tested methods, due to proving most information and computational coherency in the form of stability. Precisely same conclusion was drawn from the similar comparison of four of these methods, for the application to connect friction stir welding parameters with the mechanical properties of the aluminum [10]. However, both of these tests were based on a special data set. Moreover, a thorough study of stopping criterion and need of network retraining in relation to backward FS for the MLP was performed in [11]. The basic observation there was that when computing saliency of a feature for the MLP by leaving it aside from input, retraining improves the selection performance. Numerical experiments were performed with stratified double five-four-folded (outer division to 5 sets, inner to 4 sets) cross-validation.

In this paper, we develop an assessment method for the feature saliency of MLP using a PaD-like approach, and provide its brief empirical evaluation with benchmark data sets through the comparison to all possible MLP models resulting from any feature subset. The approach as a whole is strictly novel, even if many individual parts have been introduced and tested in earlier work.

2 The Method

Assume that the data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ is given, where $\mathbf{x}_i \in \mathbb{R}^{n_0}$ denote the input and $\mathbf{y}_i \in \mathbb{R}^{n_2}$ the output vectors, respectively. Determine the weight matrices $\mathbf{W} \in \mathbb{R}^{n_2 \times (n_1+1)}$ and $\mathbf{V} \in \mathbb{R}^{n_1 \times (n_0+1)}$ for the MLP with single hidden layer of the size n_1 by minimizing the cost function

$$\mathcal{J}(\mathbf{W}, \mathbf{V}) = \frac{1}{2N} \sum_{i=1}^{N} \|\mathcal{N}(\mathbf{W}, \mathbf{V})(\mathbf{x}_i) - \mathbf{y}_i\|^2 + \frac{\beta}{2n_1} \sum_{(i,j)} \left(|\mathbf{V}_{i,j}|^2 + |(\mathbf{W}_1)_{i,j}|^2 \right) \tag{1}$$

for the penalization parameter $\beta \geq 0$ and the feedforward mapping $\mathcal{N}(\mathbf{W}, \mathbf{V})(\mathbf{x}_i) = \mathbf{W}\tilde{\mathcal{F}}(\mathbf{V}\tilde{\mathbf{x}_i})$. By $\tilde{}$ we denote the enlargement of a vector by constant 1 to include biases, so that both weight matrices have the block structure $\mathbf{W} = [\mathbf{W}_0 \; \mathbf{W}_1]$ and $\mathbf{V} = [\mathbf{V}_0 \; \mathbf{V}_1]$ for $\mathbf{W}_0$ and $\mathbf{V}_0$ containing the bias weights. The special form of penalization omitting the bias column $\mathbf{W}_0$ is due to Corollary 1 in [12]: *Every locally optimal solution to* (1) *provides an unbiased regression estimate having zero mean error over the training data.* Because the MLP model has the linear final layer, this refines the basic bias-variance interpretation of the linear models as described in Section 1 due to [7]. Furthermore, the unbiasedness is also a constraint, i.e., it is always satisfied by the locally optimal MLPs for (1).

This paper continues the work of [1, 2], where the ten-fold cross-validation (10-CV) has been used as the best-generalization-error seeking learning framework for the MLP. The grid-search of MLP complexity is carried out, with respect to the size of the hidden layer n_1 (structural complexity) and the amount of penalization β to favor small weights (functional complexity, see [13]). Moreover, as demonstrated in [2] (and the original references therein), the use of the so-called *Distribution Optimally Balanced* stratified CV (Dob-SCV) can stabilize the selection of the two metaparameters (n_1, β), especially in classification compared to the classical stratified CV. Namely, the folding approach in Dob-SCV is

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Algorithm 1 Feature saliency assessment using the AMAS approach.

```
Input: Data $(\mathbf{X}, \mathbf{Y}) = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ of inputs and desired outputs.
Output: Ranked list of MLP input variables with relative saliencies.
1: Determine $(n_1^*, \beta^*)$ as in [2] with 10-Dob-SCV
   $(n1max = 20,\ \vec{\beta} = \{10^{-4}, 10^{-5}, 10^{-6}\})$
2: Train $\mathcal{N}(\mathbf{W}, \mathbf{V})$ with full data using $(n_1^*, \beta^*)$ (5 restarts of which the best
   solution to (1) is stored)
3: Compute AMAS of $\mathcal{N}(\mathbf{W}, \mathbf{V})$ using formula (3)
4: Order input features in descending order with respect to AMAS and normalize the final saliencies to have their sum equal to unity
```

based on approximating the classwise density distributions (or the input density in regression) in the folds, in addition to approximating the class frequencies as with the classical stratified CV. To this end, also along the findings in [2], we compute the network errors for training and test sets in CV using MRSE (Mean-Root-Squared-Error) instead of the more common RMSE.

To assess the saliency of a feature of the MLP model, the approaches that were found most successful in the works reviewed in Section 1 were based on the analytical input-output derivative (the PaD method) and retraining of the MLP model when its inputs change. The analytic formula of the feedforward network's input sensitivity can be directly calculated from the layer-wise description of the network's action as given above. It reads as follows:

$$\frac{\partial \mathcal{N}}{\partial \mathbf{x}} = \mathbf{W}_1 \,\mathrm{Diag}\{(\mathcal{F})'(\mathbf{V}\tilde{\mathbf{x}})\}\, \mathbf{V}_1. \tag{2}$$

To take into account the given training data, (2) is applied to compute the *Mean Absolute Sensitivity, MAS*, of the network as follows:

$$\frac{1}{N} \sum_{i=1}^{N} \left| \frac{\partial \mathcal{N}}{\partial \mathbf{x}_i} \right|. \tag{3}$$

The approach as a whole is referred as AMAS (Analytic MAS). The actual input ranking is then obtained according to the rule: *the higher the AMAS, the more salient the feature is.* This is based on the well-known Taylor's theorem in calculus related to local approximation of smooth functions. Namely, if a function is locally constant, its gradient vector is zero and such a function could be (locally) represented and absorbed to the MLP bias. Note that exactly the same starting point to define variables which truly affect the output was taken in [14] when deriving the Delta Test for FS. Hence, the larger the mean sum of the absolute values of the local partial derivatives with respect to an input variable, the more important that input variable is for representing the variability of an unknown function approximated by the MLP.

The overall approach for obtaining feature saliencies of the MLP is given in Algorithm 1. This algorithm was preliminarily applied to educational data in [15] with promising results on a methodologically triangulated framework.

3 Experimental results

We carried out a set of experiments in Matlab, using *fminunc* to minimize (1), with small classification benchmark data sets from the UCI repository as described in Table 1. All features were scaled into the range $[0, 1]$ of the sigmoidal activation functions, and the class labels were encoded as target output using the *1-of-*n_2 coding scheme. For all data sets, all the best-generalization-error MLP models were determined using 10-Dob-SCV by testing all possible $2^{n_0} - 1$ variable subsets. This was done to properly assess the proposed technique, however, meaning that when combined with the grid search in Step 1 of Algorithm 1, the wall-clock time of the individual tests was hours or even days even with these small data sets. In the tables, we report the best models with their mean MRSE over the test folds (TsE) and the mean amount of misclassifications in percentages over the test folds (TsFCP). For some feature subsets there is only one result providing minimal solution for both of these quantities whereas in certain cases two different solutions are given. For each test data, the best TsE and TsFCP are indicated with the bold font. Ranking of all input features according to Algorithm 1 is provided in the last row of the corresponding table.

We draw the following conclusions from Tables 2–4: i) For “Iris”, the two most important features as determined by the experimented subsets are $\{3\}$ and $\{4\}$ which are ranked top by the AMAS-algorithm 1 as well. These results coincide with the tests and conclusions in [5]. The best feature selection suggested by the minimal TsE and TsFCP is to keep all the features. However, for a reduced model, the second best TsFCP (3.3) is already obtained with two or three variables including feature $\{1\}$ which is not highly ranked both by the AMAS-algorithm or the tests in [5] (also visually it does not appear relevant). ii) For “Seeds”, we have in all aspects an excellent result by the proposed technique: The most common variables in the best models encountered during the full exploration are correctly identified by the AMAS-algorithm. The best model for both TsE and TsFCP is obtained with the three variables $\{1, 4, 7\}$, whose AMAS-values for the full set are also clearly higher than for the rest of the original features. Moreover, for the reduced three feature model, the AMAS-ranking with the corresponding relative saliencies is $\{7\}/0.35$, $\{4\}/0.35$, and $\{1\}/0.29$, which again coincides with the overall importance of the individual features. iii) For “Glass”, with imbalanced class frequencies, the results appear more complex. There is a significant difference between the models of best TsE and best TsFCP, with low correlation of the two quantities in the reduced models in general. The best TsE with six variables is well captured in the order of the AMAS-ranking

Name of data	Abbr.	N	n_0	n_2	Class frequencies
Iris	Iris	150	4	3	[50 50 50]
Seeds	Seeds	210	7	3	[70 70 70]
Glass Identification	Glass	214	9	6	[70 76 17 13 9 29]

Table 1: Description of UCI classification data sets for the experiments.

Inds	TsE(Std)/TsFCP	Inds	TsE(Std)/TsFCP
$\{4\}$	1.08e-1(5.3e-2)/4.0	$\{1,3\}$	8.74e-2(7.2e-2)/3.3
$\{2,3,4\}$	6.95e-2(6.6e-2)/4.0	$\{1,3,4\}$	6.96e-2(6.1e-2)/3.3
$\{1,2,3,4\}$	**6.46e-2**(7.1e-2)/**2.7**	$\{3\}$/0.37, $\{4\}$/0.35, $\{2\}$/0.20, $\{1\}$/0.08	

Table 2: Results for Iris.

Inds	TsE(Std)/TsFCP	Inds	TsE(Std)/TsFCP
$\{1\}$	2.67e-1(4.1e-2)/13.8	$\{2\}$	2.68e-1(4.2e-2)/13.8
$\{2,7\}$	1.05e-2(3.9e-2)/5.2		
$\{1,4,7\}$	**8.90e-2**(5.3e-2)/**3.8**	$\{2,4,7\}$	9.04e-2(4.5e-2)/**3.8**
$\{1,2,4,7\}$	9.33e-2(6.1e-2)/4.8	$\{2,4,5,7\}$	9.98e-2(5.7e-2)/4.3
$\{1,2,3,4,7\}$	9.81e-2(6.3e-2)/4.8	$\{1,2,4,5,7\}$	1.05e-1(6.1e-2)/4.3
$\{1,2,4,5,6,7\}$	1.05e-1(6.6e-2)/5.2		
$\{7\}$/0.26, $\{4\}$/0.24, $\{1\}$/0.17, $\{2\}$/0.11, $\{5\}$/0.08, $\{3\}$/0.07, $\{6\}$/0.07			

Table 3: Results for Seeds.

Inds	TsE(Std)/TsFCP	Inds	TsE(Std)/TsFCP
$\{1\}$	7.08e-1(4.9e-2)/43.8	$\{4\}$	7.13e-1(6.2e-2)/41.2
$\{1,3\}$	5.61e-1(6.5e-2)/27.3	$\{1,2,3\}$	5.74e-1(5.1e-2)/**26.0**
$\{1,3,6,8\}$	5.68e-1(7.2e-2)/26.8		
$\{2,3,5,7,8\}$	5.57e-1(9.1e-2)/30.0	$\{1,3,5,6,8\}$	5.83e-1(9.0e-2)/28.1
$\{2,3,5,6,7,8\}$	**5.36e-1**(8.5e-2)/29.0	$\{2,3,4,6,8,9\}$	5.66e-1(9.6e-2)/28.8
$\{2,3,5,6,7,8,9\}$	5.82e-1(1.1e-1)/31.9	$\{1,2,3,4,5,7,8\}$	5.98e-1(7.4e-2)/26.1
$\{1,3,4,5,6,7,8,9\}$	5.72e-1(5.4e-2)/31.4	$\{2,3,4,5,6,7,8,9\}$	5.92e-1(9.4e-2)/29.4
$\{8\}$/0.36, $\{7\}$/0.22, $\{3\}$/0.10, $\{6\}$/0.09, $\{1\}$/0.07, $\{5\}$/0.05, $\{2\}$/0.05, $\{4\}$/0.04, $\{9\}$/0.03			

Table 4: Results for Glass.

(except feature $\{1\}$), which also identifies the individually mostly occuring feature in the best models of different size, namely $\{8\}$. However, an important feature for the best 10-Dob-SCV based MLP models with respect to TsE and TsFCP, namely $\{2\}$, does not have a high relative saliency. Altogether, the results in Table 4 show the difficulty in defining an appropriate reference in FS when the correlation between the model's generalization measure TsE and it's actual quality, TsFCP, is jeopardized.

4 Conclusions

We witnessed many potential advantages of feature selection according to [3], as stated in Section 1, with the AMAS-algorithm. The full feature set was the best overall choice for "Iris", even if only half of the original variables yielded to a high generalization accuracy in classification. For "Seeds", the best model, reliably determined by AMAS, contained only three of the original seven features. For

"Glass" with imbalanced class frequencies, the best models with respect to the two measures considered were different, but still with reduced subsets of features. The most important features for this data set were not identified perfectly when analyzing the full feature set "surface" using the AMAS-algorithm. To conclude, in two of the three cases, the AMAS-algorithm 1 applied to the full problem was able to detect and identify the most salient features of the best-generalization-error MLP model according to 10-Dob-SCV.

References

[1] T. Kärkkäinen, A. Maslov, and P. Wartiainen. Region of interest detection using MLP. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2014*, pages 213–218, 2014.

[2] T. Kärkkäinen. On cross-validation for MLP model evaluation. In *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science (8621), pages 291–300. Springer-Verlag, 2014.

[3] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[4] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, pages 121–129, 1994.

[5] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining.* Norwell, MA: Kluwer, 1998.

[6] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[7] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, 1990.

[8] M. Gevrey, I. Dimopaulos, and S. Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological Modelling*, 160:249–264, 2003.

[9] Y. Dimopoulos, P. Bourret, and S. Lek. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters*, 2(6):1–4, 1995.

[10] M. H. Shojaeefard, M. Akbari, M. Tahani, and F. Farhani. Sensitivity analysis of the artificial neural network outputs in friction stir lap joining of aluminum to brass. *Advances in Material Science and Engineering*, 2013:1–7, 2013.

[11] E. Romero and J. M. Sopena. Performing feature selection with multilayer perceptrons. *IEEE Transactions on Neural Networks*, 19(3):431–441, 2008.

[12] T. Kärkkäinen. MLP in layer-wise form with applications to weight decay. *Neural Computation*, 14:1451–1480, 2002.

[13] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[14] E. Eirola, E. Liitiäinen, A. Lendasse, F. Corona, and M. Verleysen. Using the Delta test for variable selection. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2008*, 2008.

[15] M. Saarela and T. Kärkkäinen. Analyzing student performance using sparse data. to appear in *Journal of Educational Data Mining*, 2015.