Multi-layer Perceptrons with Embedded Feature Selection with Application in Cancer Classification*

Liefeng Bo, Ling Wang and Licheng Jiao

Institute of Intelligent Information Processing, Xidian University, Shannxi 710071, China

Abstract — This paper proposed a novel neural network model, named multi-layer perceptrons with embedded feature selection (MLPs-EFS), where feature selection is incorporated into the training procedure. Compared with the classical MLPs, MLPs-EFS add a preprocessing step where each feature of the samples is multiplied by the corresponding scaling factor. By applying a truncated Laplace prior to the scaling factors, feature selection is integrated as a part of MLPs-EFS. Moreover, a variant of MLPs-EFS, named EFS+MLPs is also given, which perform feature selection more flexibly. Application in cancer classification validates the effectiveness of the proposed algorithms.

Key words — Multi-layer perceptrons (MLPs), neural networks, Laplace prior, feature selection, cancer classification.

I. Introduction

With the wide application of information technology, one has a growing chance to confront high dimensional data sets with hundreds or thousands of features. Typical examples include text processing of internet documents, gene expression array analysis, and combinatorial chemistry. These data sets often contain many irrelevant and redundant features, hence an effective feature selection scheme becomes a key factor in determining the performance of a classifier. The purpose of feature selection is to find the smallest subset of features that result in satisfactory generalization performance. The potential benefits of feature selection are at least three-fold [1-2]:

① Improving the generalization performance of learning algorithm by eliminating the irrelevant and redundant features;

② Enhancing the comprehensibility of learning algorithm by identifying the most relevant feature subset;

③ Reducing the cost of future data collection and boosting the test speed of learning algorithm by only utilizing a small fraction of all features.

In MLPs, the input of hidden unit is the linear combination of all the components of samples. This mechanism is highly effective for feature extraction, but not for feature selection since the weight of each component is often non-zero. Therefore, various new methods have been proposed to select a good feature subset for MLPs. In 1997, Setiono and Liu [3] developed a backward elimination method, named neural network feature selector (NNFS) for feature selection. NNFS encourages the small weights to converge to zeros by adding a penalty term to the error function. In 2002, Hsu et al. [4] proposed a wrapper method, named ANNIGMA-wrapper for fast feature selection. In 2004, Sindhwani et al. [5] presented a maximum output information (MOI-MLP) algorithm for feature selection. Due to the vast and extensive literatures in feature selection for MLPs, we only mentioned a small fraction of them and the reader can refer to [6] for more information.

In this paper, we present a neural networks model, named MLPs-EFS that incorporates feature selection into the training procedure. Compared with the classical MLPs, MLPs-EFS add a preprocessing step where each feature of the samples is multiplied by the corresponding scaling factor. In order to achieve feature selection, a truncated Laplace prior is applied to the scaling factors. Some researchers [7] have shown that the Laplace prior promotes sparsity and leads to a natural feature selection. The sparsity-promoting nature of Laplace prior is discussed in the literature [8]. Moreover, we also derive a variant of MLPs-EFS, named EFS+MLPs that can perform feature selection more flexibly. Application in cancer classification validates the effectiveness of the proposed algorithms

The rest of this paper is organized as follows. Section II presents the MLPs-EFS and EFS+MLPs algorithms. Section III reports the experimental results of proposed algorithms and compares them with the results obtained by several existing algorithms. Section IV discusses the contributions of this paper

II. Multi-layer Perceptrons with Embedded Feature Selection

Consider the three-layer perceptrons with embedded feature selection shown in Fig. 1. The error measure we optimize is the sum of the squared difference between the desired output and the actual output,

$$E(\mathbf{w},\mathbf{v},\boldsymbol{\alpha}) = \frac{1}{lc}\sum_{m=1}^{c}\left(\mathbf{y}^{(m)} - \phi_o\left(\mathbf{H}\mathbf{v}^{(m)}\right)\right)^T\left(\mathbf{y}^{(m)} - \phi_o\left(\mathbf{H}\mathbf{v}^{(m)}\right)\right), \quad (1)$$

where we have the following:

- l is the number of training samples.

* This research was supported by the National Natural Science Foundation of China under Grant No. 60372050.

- c is the number of output units that is equal to the number of category.
- $\mathbf{y}^{(m)}$ is the output vector of the m-th class using the "one-vs-all" encoding scheme, which satisfies $y_i^{(m)}=1$ if the *i*-th sample belongs to the *m*-th class, and $y_i^{(m)}=0$ otherwise.
- $\mathbf{v}^{(m)}$ is the weight vector from hidden units to the m-th output unit.
- $\mathbf{H}_{ij}$ is the output of the *i*-th sample at the *j*-th hidden unit,

$$\mathbf{H}_{ij}=\phi_h\left(\left(\mathbf{w}^{(j)}\right)^T\left(\boldsymbol{\alpha}^2\otimes\mathbf{x}^{(i)}\right)+b^{(j)}\right). \tag{2}$$

- $\otimes$ denotes the elementwise multiplication.
- $\mathbf{x}^{(i)}$ is the i-th input vector.
- $\boldsymbol{\alpha}^2$ is the scaling factors of input vector. Note that $\boldsymbol{\alpha}^2$ denotes $\boldsymbol{\alpha}\otimes\boldsymbol{\alpha}$.
- $\mathbf{w}^{(j)}$ is the weight vector from the input layer to the *j*-th hidden unit.
- $b^{(j)}$ is the bias term of the *j*-th hidden unit.
- ϕ_o and $\phi_h(\bullet)$ are the activation function that often is the logistic function, hyperbolic tangent function, or linear function.

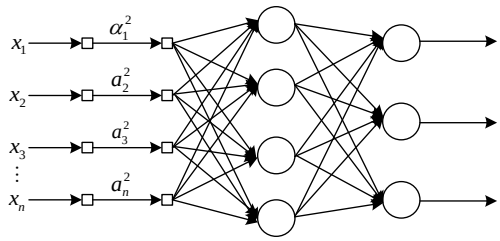

Fig. 1. Three-layer perceptrons with embedded feature selection

Compared with the classical MLPs, MLPs-EFS add a preprocessing step where each feature of the samples is multiplied by the corresponding scaling factor α_i^2. After MLPs-EFS are trained, the importance of features can be identified by the magnitude of the scaling factors. That is, the irrelevant and the redundant features are associated with the small scaling factors and the relevant features with the large scaling factors. Since we hope that the irrelevant and the redundant features affect learning procedure as little as possible, it is necessary to make the small scaling factors close to zeros. To achieve this goal, we adopt a truncated Laplace prior as shown in Fig. 2. over the scaling factors

$$P_1\left(\boldsymbol{\alpha}^2\right)\propto\begin{cases}\exp\left(-\dfrac{\lambda_1}{n}\left\|\boldsymbol{\alpha}^2\right\|_1\right) & \text{if } \boldsymbol{\alpha}^2\geq 0\\ 0 & \text{if } \boldsymbol{\alpha}^2<0\end{cases}, \tag{3}$$

where n is the feature dimensionality of the samples, $\left\|\boldsymbol{\alpha}^2\right\|_1=\sum_{i=1}^{n}\alpha_i^2$ denotes the 1-norm and λ_1 controls the intensity of the prior.

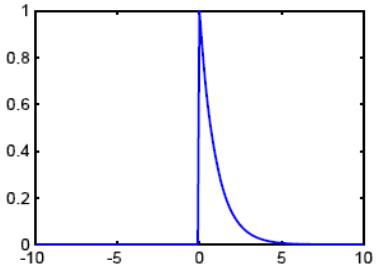

Fig. 2. Prior over the scaling factor α_i^2

If we do not restrict the weights from input layer to hidden layer, it may become very large. To avoid this case, we adopt a Gaussian prior over those weights

$$P_2\left(\mathbf{w}\right)\propto\exp\left(-\frac{\lambda_2}{n\times h}\sum_{j=1}^{h}\left\|\mathbf{w}^{(j)}\right\|_2^2\right), \tag{4}$$

where h is the number of hidden units, $\left\|\mathbf{w}^{(j)}\right\|_2^2=\sum_{i=1}^{n}\left(w_i^{(j)}\right)^2$ denotes the 2-norm and λ_2 controls the intensity of the prior.

Considering the error measure along with the prior, we obtain a penalized maximum likelihood estimate of the scaling factors and the weights by minimizing the following expression

$$\begin{aligned}f\left(\mathbf{w},\mathbf{v},\boldsymbol{\alpha}\right)&=E\left(\mathbf{w},\mathbf{v},\boldsymbol{\alpha}\right)+\log\left(P_1\left(\boldsymbol{\alpha}\right)\right)+\log\left(P_2\left(\mathbf{w}\right)\right)\\&=E\left(\mathbf{w},\mathbf{v},\boldsymbol{\alpha}\right)+\frac{\lambda_1}{n}\sum_{i=1}^{n}\alpha_i^2+\frac{\lambda_2}{n\times h}\sum_{j=1}^{h}\sum_{i=1}^{n}\left(w_i^{(j)}\right)^2.\end{aligned} \tag{5}$$

Like the original MLPs, the optimization problem that MLPs-EFS confront with is unconstrained. Hence, all the optimization algorithms applied to MLPs are also suitable for MLPs-EFS. In order to speedup the training procedure, instead of the standard backpropagation (BP) algorithm, a variant of the conjugate gradient, named scaled conjugate gradient (SCG) [9], is used to find a local minimum of the objective function $f\left(\mathbf{w},\mathbf{v},\boldsymbol{\alpha}\right)$. SCG is fully automated including no user dependent parameters and avoiding a time consuming line-search. Experimental study performed by Moller in 1993 showed that SCG yields a speedup of at least an order of magnitude relative to BP. Demuth's test report [10] also showed that SCG performs well over a wide variety of problems, particularly for networks with a large number of weights.

MLPs-EFS add the scaling factors; hence, the number of its free parameters is slightly larger than that of MLPs. However the training time of MLPs-EFS is comparable with that of MLPs because the number of free parameters added by MLPs-EFS is far smaller than the total number of the free parameters of MLPs. For example, for a network with ten hidden units and three output units, the number of free parameters of MLPs-EFS is only 1.08 times that of MLPs if the feature dimensionality of the samples is 10. Similar conclusion also holds true for the test time.

After MLPs-EFS are trained, there exist some scaling factors that are very close to but not exactly zeros. To remove the features with small scaling factors from network, we need

a variant of MLPs-EFS, named EFS+MLPs that can be described as the following:

1. Train MLPs-EFS with the full features;
2. Rank the importance of features according to the magnitude of the scaling factors;
3. Reserve the d most important features;
4. Train MLPs only with the reserved d features.

III. Comparisons with Existing Algorithms

In order to know how well MLPs-EFS and EFS+MLPs work, we compare them with FDR+MLPs and SVM RFE [11] on three data sets. XOR1 are artificially constructed problems which are variants of classical XOR problem. Leukemia and Lymphoma are cancer data sets, which are available at http://llmpp.nih.gov/lymphoma and http://www-genome.wi.-mit.edu/mpr. For cancer classification, one needs to determine the relevant genes in discrimination as well as discriminate accurately, so our algorithms are very suitable for these problems.

The regularization parameter λ_2 in MLPs-EFS is fixed to 0.00001, and λ_1 and h are chosen using the ten-fold cross validation on training samples. All the scaling factors are initialized to 1. The input and hidden-to-output weights are randomly initialized in the range $[1/\sqrt{h}, 1/\sqrt{h}]$.

FDR+MLPs: Fisher discriminant ratio (FDR) is a well known filter method that assigns the importance of each feature independently based on its ability to distinguish classes. FDR of the *i*-th feature is given by

$$FDR(i) = \frac{\sum_{k=1}^{C}\left(\mu^{(k)}(i) - \mu(i)\right)^2}{\sum_{k=1}^{C}\sum_{j=1}^{m_k}\left(x_j^{(k)}(i) - \mu^{(k)}(i)\right)^2} \qquad (6)$$

where $\mu(i)$ and $\mu^{(k)}(i)$ are the mean and the class-conditional mean of the *i*-th feature, respectively, and C and m_k are the number of classes and of samples that belong to the *k*-th class. A larger FDR value suggests that the corresponding feature is better able to distinguish classes. Hence, feature selection can be implemented by removing the features with smaller FDR values. In this paper, FDR is regarded as a baseline method.

SVM RFE: Recursive feature elimination (RFE) is a recently proposed feature selection algorithm described by Guyon et al. This method tries to find the best feature subset which leads to the largest margin of class separation for support vector machine (SVM) classifier. RFE approximates the solution of this combinatorial problem by a greedy algorithm, which removes the features that decrease the margin step by step.

XOR problem is very popular for testing the performance of neural networks. Here, we restrict each feature to be drawn from the uniform distribution between -1 and 1. The label is defined as

$$y = \begin{cases} +1 & if\ x_1 x_2 \geq 0 \\ -1 & otherwise \end{cases} \quad x_1, x_2 \in U(-1, +1). \qquad (7)$$

The points with $y = 1$ lie in the first or third quadrant and those with $y = -1$ in the second or fourth quadrant. The optimal decision function of this problem is non-linear and the highest recognition rate of linear classifiers is only 66.67%.

XOR1: XOR1 is a variant of XOR problem (7) and derived by adding 18 redundant features which are copies of the first two features and 20 irrelevant features which are drawn from the uniform distribution between -1 and 1, simultaneously. The full data set consists of 1000 training samples and 1000 test samples.

From Fig. 3, we observe that EFS+MLPs successfully identify the relevant features on XOR1 problems; however FDR+MLPs fail to find the exact relevant features. This group of experiments suggests that EFS+MLPs can effectively handle both irrelevant and redundant features and however FDR+MLPs have some difficulties in removing the relevant and redundant features.

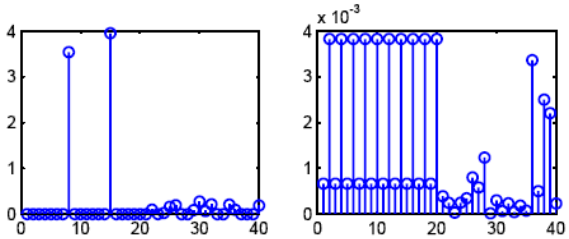

Fig. 3. Scaling factor and FDR of each feature on XOR1

Leukemia: This problem is to distinguish two variants of leukemia (ALL and AML). The data set consists of 38 training samples (27 ALL and 11 AML) and 34 test samples (20 ALL and 14 AML). The number of features for this problem is 7129 and much higher than that of the samples. Before experiments, all the features are scaled to have zero mean and unit variance. The final accuracies are averaged over 30 random splits. Table 1. lists the performance of MLPs-EFS and EFS+MLPs on feature subsets of size 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 7129, and compares them with the results obtained by FDR+ MLPs and SVM RFE. For this problem, EFS+MLPs and SVM RFE obtain the highest accuracy. MLPs-EFS significantly outperform MLPs without feature selection.

Lymphoma: This problem is to distinguish the malignant and normal lymphoma. 61 of the samples are in classes "DLCL", "FL" or "CLL" (malignant) and 35 are labeled "otherwise" (normal). The number of features for this problem is 4026 and much higher than that of the samples. Before experiments, all the features are scaled to have zero mean and unit variance. Most authors have randomly split the 96 samples set into a training set of size 60 and a test set of size 36. For the sake of comparison, we adopt the same scheme. The final accuracies are averaged over 30 random splits. Table 2. lists the performance of MLPs-EFS and EFS+MLPs on feature subsets of size 20, 50 100, 250, 500 1000, 2000, 3000 and 4206, and compares them with the results obtained by FDR+MLPs and SVM RFE. We observe that EFS+MLPs obtain the highest accuracy on feature subset of size 250. The accuracy of MLPs-EFS is higher than that of MLPs without feature selection.

Table 1. Accuracies of four classifiers on Leukemia data set and F. denotes the number of the features.

F.	FDR+MLPs	MLPs-EFS	EFS+MLPs	SVM RFE
4	88.24	/	91.18	91.18
8	91.18	/	97.06	100.00
16	97.06	/	100.00	100.00
32	97.06	/	100.00	97.06
64	97.06	/	97.06	94.12
128	94.12	/	97.06	97.06
256	91.18	/	94.12	94.12
512	94.12	/	91.18	88.24
1024	94.12	/	94.12	94.12
2048	91.18	/	91.18	85.29
4096	73.53	/	85.29	70.59
7129	82.35	94.12	82.35	85.29

Table 2. Accuracies of five classifiers on Lymphoma data set and F. denotes the number of the features.

F.	FDR+MLPs	MLPs-EFS	EFS+MLPs	SVM RFE
20	87.78	/	91.57	90.83
50	91.39	/	94.26	93.06
100	91.67	/	94.35	93.43
250	91.57	/	94.72	93.52
500	92.31	/	93.89	93.06
1000	92.22	/	92.87	92.04
2000	91.11	/	92.50	92.13
3000	90.00	/	92.31	92.13
4206	90.09	93.08	92.13	92.13

IV. Discussions

Due to space limitation, this paper only discusses the MLPs, but it is possible to apply our method to other neural network models since the proposed idea is very general. Also, comparing our algorithm with Biomimetic Pattern Recognition [12] is an interesting work. Further research will focus on these.

REFERENCES

[1] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, pp. 245-271, 1997.

[2] L. F. Bo, L. Wang and L. C. Jiao, "Feature scaling for kernel Fisher discriminant analysis using leave-one-out cross validation", Neural Computation, vol. 18, no. 4, pp. 961-978, 2006.

[3] R. Setiono and H. Liu, "Neural network feature selector," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp.645-662, 1997.

[4] C. N. HSU, H. J. Huang and D. Schuschel, "The ANNIGMA-wrapper approach to Fast Feature Selection for Neural Nets," *IEEE Transactions on Systems, Man, and Cybernetics Part B*, Vol. 32, pp. 207-212, 2002.

[5] V. Sindhwani, S. Rakshit, D. Deodhare, D. Erdogmus, J. Principe and P.Niyogi, "Feature Selection in MLPs and SVMs Based On Maximum Output Information," *IEEE Transactions on Neural Networks*, vol. 15, pp. 937-948, 2004.

[6] G. P. Zhang, "Neural Networks for Classification: A Survey," *IEEE Transactions on Systems, Man, and Cybernetics Part B*, vol. 30, pp. 451-462, 2000.

[7] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society (B)*, vol. 58, pp. 267–288, 1996.

[8] D. Donoho and M. Elad, "Optimally sparse representations in general (nonorthogonal) dictionaries by l_1 minimization," *Proceedings of the National Academy of Science*, vol. 100, pp. 2197–2202, 2003.

[9] M. F. Moller, "A scaled conjugate gradient for fast supervised learning," *Neural Network*, vol. 6, pp. 525-533, 1993.

[10] H. Demuth and M. Beale, *Neural network toolbox for use with MATLAB*, The MathWorks Inc., Natick, MA, 1998.

[11] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 45. pp. 389-422, 2002.

[12] S. J. Wang and J. L. Lai. "A More Complex Neuron in Biomimetic Pattern Recognition," *International Conference on Neural Networks and Brain*, Beijing, China, vol. 3, pp. 1487 – 1489, 2005.

Liefeng Bo is a Ph.D. candidate in Institute of Intelligent Information Processing, Xidian University. His current research interests include kernel-based learning, manifold learning, neural networks and computer version. He has published several papers in some leading journals such as Neural Computation and IEEE Transactions on Neural Networks. For more information, visit Bo's homepage http://see.xidian.edu.cn/graduate/lfbo/.

Ling Wang is a Ph.D. candidate in Institute of Intelligent Information Processing, Xidian University. Her current research interests include pattern recognition, statistical machine learning, and image processing.

Licheng Jiao is a professor and Ph.D. supervisor. He is the author or coauthor of more than 150 scientific papers. His current research interests include signal and image processing, nonlinear circuit and systems theory, learning theory and algorithms, optimization problems, wavelet theory, and data mining.