M. Zhao & Z. Shi (Eds.), Proceedings of the 2005 IEEE International Conference on Neural Networks and Brain, 3. Beijing: IEEE Press, 2005, pp. 1564-1567.

Tikhonov-based Regularization of a Global Optimum Approach of One-layer Neural Networks with Fixed Transfer Function by Convex Optimization

Dik Kin Wong
CSLI, Stanford University

Marcos Perreau Guimaraes
CSLI, Stanford University

E. Timothy Uy
CSLI, Stanford University

Patrick Suppes
CSLI, Stanford University

***Abstract*— Regularization is useful for extending learning models to be effective for classifications. Given the success of regularized-perceptron-based (one-layer neural network) methods, we introduced a similar kind of regularization for two global-optimum approaches recently proposed by Castillo et al, which combined the degree of freedom of using nonlinear transfer functions with the computational efficiency of solving complex problems. We focused on the two approaches that used sigmoid transfer functions. The first linear approach involved solving a set of linear equations, while the second min-max approach was reduced to a linear programming problem. We introduced regularization in such a way that the first linear approach remained linear and had a close form solution, while the second min-max approach was converted from a linear programming into a quadratic programming problem. Electroencephalography recordings were used to show how classifications could be improved.**

I. Introduction

Castillo et al [1] defined the objective function of a single-layer neural network based on the input-space error $x - \hat{x}$ with some desired $\hat{x}$, instead of the classical approach which was based on the output-space error $y - \hat{y}$ with some desired $\hat{y}$. This reformulation allowed the use of convex optimizations. However, when we applied these optimization techniques to solve the weights of an multioutput perceptron (single-layer neural network) for electroencephalography (EEG) classification, the overfitting problem was readily observed. As we know, fitting parameters to training data by a complex learning model can induce an expense on the generalization results of the test set. This problem was especially serious when data were limited. The collection of EEG data is indeed expensive in the sense that making long and continuous recordings on conscious subjects is somewhat impossible, due to various problems such as scalp abrasion and human fatigue. These problems also restricted the frequency of how often an experimental session could be carried out. Consequently, a typical experiment would generally last for more than a month, but the number of trials collected would still be relatively small compared to many machine-learning problems. Under these circumstances, as we increased the complexity of our learning model, overfitting becomed a major problem (in the simple case of perceptrons, it is helpful to consider Cover's complexity theory on perceptron [2] to see how training rates can sometimes be meaningless).

A. Experiment and Methods

The data we studied here is EEG-recorded brainwave with the stimuli of 100 distinct geographical sentences in English. The sentences were presented both visually and auditorily. Typical sentences were "Berlin is not a city of Germany" and "Albuquerque is the largest city of New Mexico". Subjects were asked to indicate whether the sentence was true or false by pressing one of the two alternative keys. A total of 11 bipolar pairs were recorded by dry electrodes from Integrated Biosensing Technologies (Redwood City, CA). Data were sampled at 1kHz and hardware-filtered between 0.1Hz and 100Hz. The data were further downsampled 16 times before the classifiers were trained. More details of the experiment and data can be found in [3].

The downsampled data were scaled to a range between -1 and +1 and then used to train multioutput perceptrons, one for each channel. The term "training" simply refered to the optimization of the parameters of the learning model, which were the weight vectors in this case. In this paper, two approaches were regularized to obtain the "weight vectors" for better classifications. The number of outputs computed by each multioutput perceprton was equal to the number of classes, and a classification was made on the validation set by selecting the class which corresponded to the maximum output. To obtain the classification rates to take into multichannel in a simple way, channels were added in the order based on some rankings that was based on the individual performance of the channels. For the best subject, the ranking we deduced was C4-T6, Cz-T6, T5-P4, T4-T6, Cz-T5, T3-T5, C3-C4, C3-T5, C3-Cz, C3-T3, C4-Cz, C4-T4, F8-T4, Cz-F8, C4-F8. Data from these channels were concatenated, resulting in longer input vectors

0-7803-9422-4/05/$20.00 ©2005 IEEE

that became the input for the larger multioutput perceptrons. More discussion of this multichannel classifier can be found in [4].

There are two common ways to select a model that can generalize better: early stopping with validation and regularization. We chose to use regularization because it could be applied to enhance the formulation without the need to make changes in the actual singular-value-decomposition (SVD) computations for solving the set of linear equations (this was based on the assumption that the SVD was the only iterative process which could be stopped early and no other iterative steps were considered). Regularization was implemented to the two approaches with fixed (sigmoid) transfer functions [1].

B. Training by weights computation

1) First approach: Linear Global Optimum: Castillo's first approach was to minimize the sum of errors, Q, in the input space for the j^{th} perceptron:

$$Q = \sum_{s=1}^{S} \left(w_0 + \sum_{i=1}^{I} w_i x_{is} - f^{-1}(y_s) \right)^2 .$$

Tikhonov regularization was used to obtain the solution of ill-posed linear systems and least squares problems [5]. For the Tikhonov regularization, the optimization objective of the linear model which consisted of the sum of least-squares errors, was augmented by a regularization term based on the norm of the solution scaled by the square of a regularization parameter λ. Similarly, such a regularization term could be added to Q. Its solution could then be obtained by solving the corresponding linear system of equations. When the transfer function $f(x)$ is linear, this formulation is exactly the same as that of the linear model, which solution can simply be computed via the pseudoinverse $(X^T X)^{-1}$. For example, when $f(x)$ is set to be the sigmoid function (with an inverse $f^{-1}(x)$), the only change being introduced was a scaling factor of the target function $\hat{y} = f(x)$. That is,

$$f(x) = \frac{1}{1 + e^{-x}}$$
$$f^{-1}(x) = -\log \frac{1}{x - 1}.$$

By setting $\hat{y} = [0.01, 0.99]$ (setting to $[0, 1]$ would pose numerical problem), we have $f^{-1}(y) = [-4.6, 4.6]$. The solution to the optimal weight, a column vector $\tilde{\mathbf{w}} = [w_0 ... w_I]^T$, for each perceptron can also be found by taking the derivative of the objective function and setting it to be zero as follow:

$$\min_{\tilde{\mathbf{w}}} \; Q$$
$$\text{where } Q = \sum_{s=1}^{S} \left(w_0 + \sum_{i=1}^{I} w_i x_{is} - f^{-1}(y_s) \right)^2 + \lambda^2 \sum_{i=0}^{I} w_i^2$$
$$\text{Set } \frac{\partial Q}{\partial w_i} = 0,$$
$$\sum_{i=1}^{I} \left(\sum_{s=1}^{S} x_{is} x_{ps} \right) w_i + \lambda^2 w_p = \sum_{s=1}^{S} \left(f^{-1}(y_s) - w_0 \right) x_{ps}$$
$$\text{where } p = 1, 2, ..., I.$$
$$\sum_{i=1}^{I} \left(\sum_{s=1}^{S} x_{is} \right) w_i + \lambda^2 w_0 = \sum_{s=1}^{S} \left(f^{-1}(y_s) - w_0 \right).$$

With a standard trick of setting all $x_{ij} = 1$ for $i = 0$, these derivative constraints simply derive the simple Tikhonov-regularized least-square solution. Defining $\mathbf{X}$ as a matrix of size $S \times (I+1)$ with entries x_{is} and $\hat{\mathbf{y}}$ with each component $\hat{y}_s = f^{-1}(y_s) - w_0$:

$$\tilde{\mathbf{w}} = (\mathbf{X}^T \mathbf{X} + \lambda^2 \mathbf{I})^{-1} \mathbf{X}^T \hat{\mathbf{y}}.$$

The classification results were preserved under the affine transformation $\hat{\mathbf{y}}' = \alpha \hat{\mathbf{y}} + \boldsymbol{\beta}$, with a constant vector $\boldsymbol{\beta}$ and non-zero constant α. For all $\tilde{\mathbf{w}}$ of the different perceptrons corresponded to the different channels:

$$\tilde{\mathbf{w}} = (\mathbf{X}^T \mathbf{X} + \lambda^2 \mathbf{I})^{-1} \mathbf{X}^T (\alpha \hat{\mathbf{y}} + \boldsymbol{\beta}).$$

This established the equivalence of classifications between regularizing the first linear approach and that of the least-squares approach (Tikhonov regularization has been shown to be effective in improving the classification rates for the latter case [4]).

2) Second approach: Min-Max Global Optimum: The second approach was set up to take into account the worse-case scenario. This was done by minimizing the maximum error, again in the input space. A penalty term could be introduced in a way similar to that of the first approach. By considering the worse case, unlike the sum of errors that was used in the typical linear model, an implicit form of regularization was already present. There are many relations that could be drawn to relate this min-max approach with other robust techniques, such as the widely-used support-vector machine (in which training could be done either by maximizing the error margin of the hyperplane with many classical optimization techniques, or by updating the weights using only the worst-classified trials in a neural network setting). Castillo et al set up the objective function as follow:

$$\varepsilon = \min_{\tilde{\mathbf{w}}} \left\{ \max_{s} \left| w_0 + \sum_{i=1}^{I} w_i x_{is} - f^{-1}(y_s) \right| \right\}. \quad (1)$$

They obtained the solution by explicitly formulating (1) as a linear programming (LP) problem. The problem was then to minimize ε subject to two inequalities for all the trials x_{is} with output y_s indexed by s:

$$w_0 + \sum_{i=1}^{I} w_i x_{is} - \varepsilon \leq f^{-1}(y_s)$$
$$-w_0 - \sum_{i=1}^{I} w_i x_{is} - \varepsilon \leq -f^{-1}(y_s).$$

It turned out that this robust generalization technique was not enough to overcome the overfitting problem for our EEG classification task. Regularization was still needed. One initial thought was to solve the LP problem first, and then figure out all the subset of constraints that still remained active. Based on these active inequality constraints, another linear problem could be defined and the solution corresponding to the smallest L_2 norm can be computed as our most preferred regularized solution in the feasible set. However, we found that the feasible set of the linear problem was often close to singleton, therefore making the second optimization step trivial but useless. The final solutions and results were not much different from the ones obtained originally. This motivated us to explore the second method, in which we replaced the absolute error with a squared one and added a regularization term based on the L_2 norm of the weight vector, in such a way that was very much like what we had done for the first approach, but in the form of a minimax problem here. For each single-output perceptron,

$$\varepsilon_{reg} = \min_{w} \left[\max_{s} \left(w_0 + \sum_{i=1}^{I} w_i x_{is} - f^{-1}(y_s) \right)^2 + \lambda^2 \sum_{i=0}^{I} w_i^2 \right]. \quad (2)$$

We solved the problem by mapping it into a quadratic programming problem, in the form of:

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{P}\mathbf{w} + \mathbf{c}^T\mathbf{w} + \mathbf{d} \quad (3)$$
$$\text{subj to } \mathbf{A}_{ineq}\mathbf{w} \leq \mathbf{b}_{ineq}$$
$$\mathbf{A}_{eq}\mathbf{w} = \mathbf{b}_{eq}.$$

With column vector $\tilde{\mathbf{w}} = [w_0 ... w_I]^T$, (2) could be rewritten in quadratic form as:

$$\min_{\tilde{\mathbf{w}},\varepsilon} [\ \tilde{\mathbf{w}}^T \quad \epsilon\ ] \begin{bmatrix} \lambda^2 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \lambda^2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{w}} \\ \epsilon \end{bmatrix},$$

subjects to the following sets of inequalities:

$$w_0 + \sum_{i=1}^{I} w_i x_{is} - \varepsilon \leq f^{-1}(y_s)$$
$$-w_0 - \sum_{i=1}^{I} w_i x_{is} - \varepsilon \leq -f^{-1}(y_s).$$

Comparing that with the general quadratic form of (3), we had:

$$\mathbf{x} = \begin{bmatrix} \tilde{\mathbf{w}} \\ \epsilon \end{bmatrix}, \ \mathbf{P} = \begin{bmatrix} \lambda^2 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & \lambda^2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$
$$\mathbf{A}_{ineq} = \begin{bmatrix} \mathbf{A}^+_{ineq} \\ \mathbf{A}^-_{ineq} \end{bmatrix},$$
$$\mathbf{b}_{ineq} = \begin{bmatrix} f^{-1}(\mathbf{y}) \\ -f^{-1}(\mathbf{y}) \end{bmatrix}$$
$$\mathbf{A}^+_{ineq} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{I1} & -1 \\ 1 & \vdots & \vdots & \vdots & -1 \\ 1 & x_{1s} & \cdots & x_{Is} & -1 \end{bmatrix},$$
$$\mathbf{b}^+_{ineq} = f^{-1}(\mathbf{y})$$
$$\mathbf{A}^-_{ineq} = \begin{bmatrix} -1 & -x_{11} & \cdots & -x_{I1} & -1 \\ -1 & \vdots & \vdots & \vdots & -1 \\ -1 & -x_{1s} & \cdots & -x_{Is} & -1 \end{bmatrix},$$
$$\mathbf{b}^-_{ineq} = -f^{-1}(\mathbf{y})$$

II. Main Results

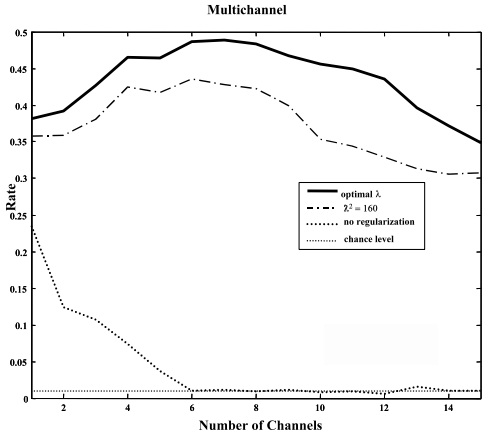

Fig. 1. Average classification rates of the first approach against number of channels used with different λs. The optimal λ varies with the number of channels.

As pointed out earlier, the first approach of presenting the error in the input space was equivalent to the linear model with a set of different target values. Figure 1 shows three curves of the classification rates versus the number of channels used, based on 10 permutations of 600 test trials. In Figure 1, the top three curves corresponded to different choices of λs. The top curve was plotted with the assumption that the optimal λ was accessible, which differed according to the number of channel being used. The middle curve corresponded to fixing $\lambda^2 = 160$, a good value based on previous findings. The bottom curve was

the case of no regularization, i.e,. $\lambda = 0$. The horizontal line at the bottom (rate = 0.01) was the chance level of the experiment (which is the mean of the binomial distribution associated with the experiment specified by the number of classes). From the plot, the optimal number of channel was 8. The importance of λ was especially indicated by the bottom curve of no regularization, or more precisely, the difference between the curve of good λs and that of $\lambda = 0$. It demonstrated how multichannel improvement relied on the proper use of regularization. For example, with 8 channels, the classification rate was close to 50% with regularization but only 1% without).

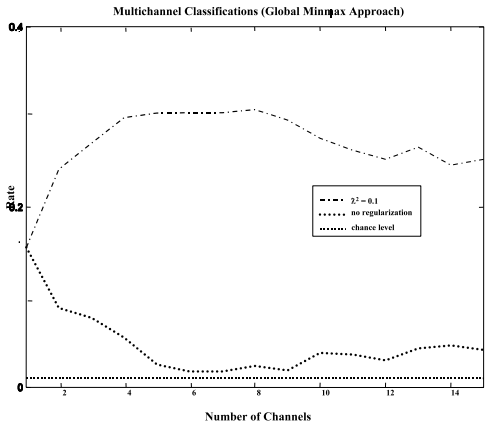

Fig. 2. Average classification rates of the second approach against number of channels used with $\lambda = 0$ (no regularization) and $\lambda^2 = 0.1$.

Figure 2 shows the results of the second approach, in which the maximum error was minimized by fitting the weight vectors. We solved the QP problem by an interior-point method: LOQO [6]. LOQO augments the objective function by adding an approximation of the logarithmic barrier function which is differentiable. The approximated logarithmic function is parametrized by t, and for each t, a convex problem with only equality constraints are formulated. The problem can then be solved by the Newton method. Results obtained from different t traces the path called central path, approaching the optimum solution.

The computation of the interior method was expensive compared to the first approach. For this reason, we computed classification rates only for $\lambda^2 = 0$ and 0.1, and we did not attempt to search over a wide enough range of λs for the optimal one. One point worth mentioning is that the good λ here had a very different range than that of the first approach. This was because the sum of errors here was based on a single trial, but was based on 1400 trials for the first approach. The range of λ was estimated by normalizing with a factor of $\sqrt{1400}$. Results showed that the overall classification rate was in fact worse than the first approach, including both the case for $\lambda^2 = 0.1$ and the no-regularization case, which we would have expected to be otherwise. However, with an approximation of λ and without a great attempt to search for optimal λ, this was not something to be totally surprised. More importantly, it was the similarity of the general shapes among the different approaches that was worth focusing on. Given the universal importance of regularization, it simplification and understanding thus became especially desirable.

III. Conclusions

The importance of generalization techniques to classification is well-known [7][8][9], while the actual mechanism of each classification method varies. Some examples were parametrizing the kernel size for Bayesian-based approach, introducing Tikhonov Regularization for linear-square problems and restricting the size of decision trees. We showed here some universal needs of generalization and formulated how the one-layer neural network could be regularized when some global optimum approaches were applied.

References

[1] E. Castillo, O. Fontenla-Romero, B. Guijarro-Berdinas, and A. Alonso-Betanzos, "A global optimum approach for one-layer neural networks," *Neural Comp.*, vol. 14, no. 6, pp. 1429–1449, 2002.

[2] T. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Ttransactions on Electronic Computers*, vol. 14, pp. 326–334, 1965.

[3] P. Suppes, *Representation and Invariance of Scientific Structures*. Stanford: CSLI Publications, 2003.

[4] D. K. Wong, M. Perreau Guimaraes, E. T. Uy, and P. Suppes, "Classification of individual trials based on the best independent component of eeg-recorded sentences," *Neurocomputing*, vol. 61, pp. 479–484, 2004.

[5] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov regularization and total least squares," 1997.

[6] R. J. Vanderbei, "LOQO: An interior point code for quadratic programming," *Optimization Methods and Software*, vol. 11, pp. 451–484, 1999.

[7] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Computation*, vol. 7, no. 2, pp. 219–269, 1995.

[8] D. J. C. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[9] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford, 1995.