Optimal Linear Regression on Classifier Outputs

Yann Guermeur, Florence d'Alché-Buc and Patrick Gallinari

LIP6, Université Pierre et Marie Curie
Tours 46-00, Boîte 169
4, Place Jussieu, 75252 Paris cedex 05, France
{guermeur,dalche,gallinari}@laforia.ibp.fr

Abstract. We consider the combination of the outputs of several classifiers trained independently for the same discrimination task. We introduce new results which provide optimal solutions in the case of linear combinations. We compare our solutions to existing ensemble methods and characterize situations where our approach should be preferred.

1 Introduction

Statisticians pointed out long ago that combining predictive models led to better estimates and performance than simply selecting the best of them [1]. Model combination is thus a viable alternative to model selection and this area has been investigated by several researchers in the neural networks field [7]. Linear combinations - often coined linear ensemble methods - have proved empirically to be efficient for both regression and discrimination tasks. However, theoretical evidence has been mainly developed for regression. The case of discrimination is more complicated since combining estimates of class probabilities for computing better estimates - this is the framework considered in most approaches - introduces specific constraints which are absent in the case of regression. We consider here linear combinations of classifiers. The general framework is that of multivariate linear regression (MLR) where constraints ensure that combination outputs are probability estimates. This framework has already been considered in [5] where a suboptimal procedure was proposed. Linear opinion pool (see Sect. 2) is a particular case for which optimal solutions have been derived. We consider here the general constrained MLR model and show how optimal solutions can be obtained by solving nonlinear programming problems. We give some arguments characterizing cases where our approach should be preferred to existing alternatives. We first introduce in Sect. 2 the general multivariate linear regression model. Sect. 3 deals with the optimal solutions and their properties. Sect. 4 presents experimental results on a difficult problem. Sect. 5 is devoted to a comparison with nonlinear techniques.

2 Multivariate Linear Regression for Combination

Let us consider a Q-category discrimination problem and P classifiers whose outputs are estimated class posterior probabilities (i.e. they are positive and sum

to one). Let f_j denote the function computed by the j^{th} classifier and $f_{jk}(x)$ its k^{th} output which approximates $p(C_k|x)$. The general multivariate approach to classifier combination corresponds to the following problem:

Problem 1 *Given a convex cost function J, find the best regression function g, $g(x) = [g_1(x) \dots g_k(x) \dots g_Q(x)]^T$ with $g_k(x) = F(x)v_k$, $(1 \leq k \leq Q)$ $F(x) = [f_1(x)^T \dots f_j(x)^T \dots f_P(x)^T]$, $v_k = [v_{k11} \dots v_{klm} \dots v_{kPQ}]^T$, which takes its values in*

$$U = \left\{ u \in \mathbb{R}_+^Q / \sum_{k=1}^{k=Q} u_k = 1 \right\} \tag{1}$$

The outputs of the *combiner* g are linear combinations of all classifier outputs. They are constrained to be non-negative and sum to one. Linear opinion pool is a degenerate case for which $g(x) = \sum_{j=1}^{j=P} v_j f_j(x)$. Coefficients v_j are scalars. In this case, constraints are satisfied if and only if the combination is convex. In [5], the authors also consider the general MLR model. They propose to determine separately the optimal regression functions $\hat{g}_k$ for each class, by solving Q separate constrained quadratic programming problems. The outputs are then standardized so that they sum to unity: $\tilde{g}_k(x) = \hat{g}_k(x) / \sum_{l=1}^{l=Q} \hat{g}_l(x)$. However, this two-step procedure is suboptimal with respect to the optimization criterion.

3 Optimal Solutions

Let us express formally the constraints. Let $v = [v_1^T \dots v_k^T \dots v_Q^T]^T$ denote the vector of parameters, and $v_{kl}^* = \min_m v_{klm}$, $(1 \leq k \leq Q)$, $(1 \leq l \leq P)$. There are two constraints: outputs must be non-negative and sum to one. Non-negativity is expressed as:

$$(Ct_1) \sum_{l=1}^{l=P} v_{kl}^* \geq 0, (1 \leq k \leq Q)$$

These constraints correspond to Q^{P+1} linear inequations.
Summation to unity is equivalent to:

$$(Ct_2) \begin{cases} \sum_{k=1}^{k=Q} (v_{klm} - v_{klQ}) = 0, (1 \leq l \leq P), (1 \leq m < Q) \\ \sum_{k=1}^{k=Q} \sum_{l=1}^{l=P} v_{klQ} = 1 \end{cases}$$

The number of inequality constraints in Ct_1 makes the resolution of the convex programming problem prohibitive. However, we established in [3] the following result, which shows that inequalities in Ct_1 may be replaced by more restrictive but simpler constraints.

Proposition 1 *An optimal solution to Problem 1 is obtained by solving the following problem:*

Problem 2

$$\min_{v} J(v)$$
$$subject\ to \begin{cases} v \in \mathbb{R}_{+}^{PQ^2} \\ (Ct_2) \end{cases}$$

This simplification allows to handle the general optimization problem using classical algorithms such as the *gradient projection method* [6].

Either quadratic loss or entropic cost function may be used as training criterion J. The following result holds: every local solution to a convex programming problem is a global solution. Although cross-entropy should be preferred for classification, there is practically no difference for performance. The quadratic programming problem is easier to handle analytically and several properties can be derived which do not hold anymore for the entropic cost. An interesting property is the characterization of conditions for unicity of the solution. In the general case, there is a convex set of optimal solutions to Problem 2. They may be not equivalent for generalization. It is thus important to know whether the solution which has been obtained is unique or not. The following proposition allows to characterize this unicity *a posteriori*, i.e. when a solution has been obtained.

Proposition 2 *A necessary and sufficient condition for a solution $\hat{v}$ to Problem 2 with a quadratic cost function to be unique is:* $\forall (k,l) \hat{v}_{kl}^{*} = \min_m \hat{v}_{klm} = 0$.

A sketch of the proof is given in Appendix 1. Conditions in proposition 2 are often met in practice. The Kuhn-Tucker conditions may be used to characterize unicity *a priori*, before any solution has been found. For lack of place, we will not develop this topic further here.

4 Experiments

To assess our combiner, we have chosen the open problem of protein secondary structure prediction. This is a 3-class classification task which consists in assigning a conformation α-helix, β-strand or coil, to each residue of a sequence. The classifiers used are the neural architecture and statistical model in [4], with the nearest-neighbours algorithm of [10]. We have compared our optimal solution to other combiners: a single hidden layer neural network (MLP), a logistic regression model and an optimal convex combination. The training of the MLR model, both for the quadratic and the cross-entropic cost functions, was performed with the gradient projection method. We chose the same set of 126 protein chains which was selected to assess the system PHD [8]. However, base sequences were substituted to the profiles of the multiple alignments. The base is divided into seven subsets. This splitting was retained to implement a two-stage cross-validation procedure. A variation of *Stacked Generalization* [9] is used to avoid the generation of biased estimates, and every subset constitutes iteratively the test set. Table 1 summarizes the observed performance.

MLR compares favourably with existing methods. The subsequent 0.7% increase in recognition rate is significant for this task.

Table 1. Relative average performance of ensemble methods

Combiner	Recognition rate
MLR cross-entropy	66.5
MLR quadratic loss	66.3
convex average	65.7
MLP	65.8
Logistic regression	65.7

5 Comparison with Nonlinear Methods

The superiority of linear combiners over nonlinear ones, a phenomenon often observed in practice, has two explanations. Data set sizes are frequently too small with respect to the number of parameters of ordinary nonlinear models. Moreover, these models are less suited than linear ones in many cases. We illustrate this point by means of a particular example. We consider the logistic regression model [2]. It is identical to the single layer perceptron with *softmax* nonlinearity:

$$g_k(z) = \frac{e^{h_k(z)}}{\sum_l e^{h_l(z)}} \tag{2}$$

where the h_l are affine combinations of the predictors. This model, which computes a linear discriminant function, performs worse than a simple optimal convex conbination for the problem described below.
Let us consider a classification problem with two classes and two classifiers. We assume that the true posterior probabilities are given by:

$$p(C_1|x) = \theta f_{11}(x) + (1-\theta) f_{21}(x) \tag{3}$$

with $\theta \in]0,1[$. $t = f_{11}(x)$ and $u = f_{21}(x)$ are supposed to be independently uniformly distributed on $[0,1]$. Let $\hat{g}$ be the logistic regression function which maximizes the expectation of the log-likelihood function. Founding $\hat{g}$ is equivalent to minimizing with respect to v the functional:

$$J(v) = -\int_0^1 \int_0^1 p(C_1|t,u) ln(g(t,u)) + (1 - p(C_1|t,u)) ln(1 - g(t,u)) dt du \tag{4}$$

We will show that $\hat{g}$ does not allow to determine the true bayesian decision boundary. With no loss of generality, we can restrict the family of functions considered to:

$$g(t,u) = \frac{1}{1 + exp(-k(\hat{\theta}t + (1-\hat{\theta})u - \frac{1}{2}))} \tag{5}$$

$\hat{\theta} \in]0,1[$, $k > 0$. The true boundary will be found if and only if $\hat{\theta} = \theta$. We demonstrate that the resulting value of $\hat{\theta}$ is actually different from θ, provided $\theta \neq \frac{1}{2}$. *Proof*: see Appendix 2. Simulation results are displayed in Figure 1. For symmetry reasons, the study has been restricted to values of θ superior to 0.5.

The slight discrepancy between θ and the estimate of $\hat{\theta}$ can be easily observed.

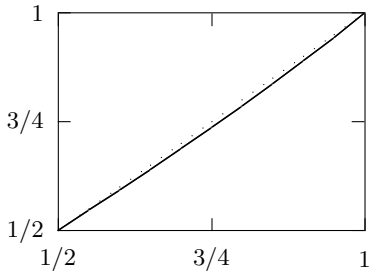

Fig. 1. Empirical estimate of $\hat{\theta}$ as a function of θ

6 Discussion

We have established how the standard multivariate linear regression model could be constrained in order to improve the estimates of the posterior probabilities of classes generated by a set of experts. The problem has been solved as a nonlinear programming problem. The linear regression approach presents several advantages. Training can be adapted to take into account complexity control. From this viewpoint, optimization methods that use the *active set method* and produce, at each step of the training phase, a *feasible point*, such as the gradient projection method, are of particular interest. We are currently studying these properties with the objective to improve performance in generalization.

References

1. Bates, J.M. and Granger, C.W.J. (1969): The Combination of Forecasts. *Operational Research Quaterly*, **Vol. 20**, 451-468.
2. Bishop, C.M. (1995): *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford.
3. Guermeur, Y., d'Alché-Buc, F. and Gallinari, P. (1997): Combinaison Linéaire Optimale de Classifieurs. *XXIX-ièmes Journées de Statistique*, 425-428.
4. Guermeur, Y. and Gallinari, P. (1996): Combining Statistical Models for Protein Secondary Structure Prediction. *ICANN'96*, Bochum, 599-604.
5. LeBlanc, M. and Tibshirani, R. (1993): Combining estimates in regression and classification. *Technical Report 9318*, Department of Preventive Medicine and Biostatistics and Department of Statistics, University of Toronto.
6. Rosen, J.B. (1960): The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints. *J. Soc. Indust. Appl. Math.*, **Vol. 8**, N^o 1, 181-217.
7. Perrone, M.P. and Cooper, L.N. (1993): When Networks Disagree: Ensemble Methods for Hybrid Neural Networks. *Technical Report*, Institute for Brain and Neural Systems, Brown University, Providence, Rhode Island.
8. Rost, B. and Sander, C. (1993): Prediction of Protein Secondary Structure at Better than 70% Accuracy. *J. Mol. Biol.*, **232**, 584-599.
9. Wolpert, D.H. (1992): Stacked Generalization. *Neural Networks*, **Vol. 5**, 241-259.

10. Zhang, X., Mesirov, J.P. and Waltz, D.L. (1992): Hybrid System for Protein Secondary Structure Prediction. *J. Mol. Biol.*, **225**, 1049-1063.

7 Appendix 1

Let $Z = \{(x_i, y_i)\}, (1 \leq i \leq N)$ be the training sample. $F(X)$ denotes the matrix of explanatory variables (its lines are the vectors $F(x_i)$). A is a block-diagonal matrix with Q identical blocks equal to $\frac{1}{N}F(X)^T F(X)$. Let $Y_k = [y_{ik}] \in \mathbb{R}^N$, $(1 \leq k \leq Q)$ and $b = \frac{1}{N}[Y_1^T F(X) \dots Y_k^T F(X) \dots Y_Q^T F(X)]^T$. The sample-based estimate of $J(v)$ is then:

$$\hat{J}(v) = \frac{1}{2}v^T Av - b^T v + \frac{1}{2} \tag{6}$$

A base of $F(X)^T F(X)$ kernel is given by the set of vectors $w_j, (1 \leq j \leq P-1)$:

$$w_j = [1_Q^T, -\delta_{j,1}1_Q^T, \dots, -\delta_{j,P-1}1_Q^T]^T \tag{7}$$

where 1_Q is a column vector of Q ones. $Ker(A) = (Ker(F(X)^T F(X)))^Q$. The following lemma holds:

Lemma 1. *If $\hat{v}$ and $\hat{v} + w$ are optimal solutions to Problem 2 with J being the mean squared error, then $w \in Ker(A)$.*

The proof relies on the following argument:

$$\hat{J}(\hat{v} + w) = \hat{J}(\hat{v}) \Longrightarrow \frac{1}{2}w^T Aw + (A\hat{v} - b)^T w = 0 \Longrightarrow w \in Ker(A) \tag{8}$$

From (7), it is clear that if condition in proposition 2 holds, any point $\hat{v} + w$ with $w \in Ker(A) \setminus \{0\}$ will have negative components and so will not be a feasible solution. This ensures the unicity of the optimal solution.

8 Appendix 2

The objective function is equal to:

$$J(\hat{\theta}, k) = -\int_0^1 \int_0^1 ln(g(t,u))dtdu - \frac{k}{12}((2\theta - 1)\hat{\theta} + (1 - \theta)) \tag{9}$$

Assuming the optimum is obtained for $\hat{\theta} = \theta$ is equivalent to solving for k:

$$\begin{cases} \frac{\partial J}{\partial \hat{\theta}}(\theta, k) = 0 \\ \frac{\partial J}{\partial k}(\theta, k) = 0 \end{cases} \tag{10}$$

After some algebra, system (10) is shown to be equivalent to:

$$\begin{cases} \int_0^1 (1-2z) ln(cosh(\frac{k}{2}(\theta z - \frac{1}{2})))dz = \frac{k\theta(1-\theta)}{12} \\ \int_0^1 (1-2z) ln(cosh(\frac{k}{2}((1-\theta)z - \frac{1}{2})))dz = \frac{k\theta(1-\theta)}{12} \end{cases} \tag{11}$$

Let $h_1(z) = (1-2z)ln(cosh(\frac{k}{2}(\theta z - \frac{1}{2})))$, $h_2(z) = (1-2z)ln(cosh(\frac{k}{2}((1-\theta)z - \frac{1}{2})))$.

$$\forall(\theta, k, z) \in]0.5, 1[\times]0, +\infty[\times]0.5, 1[, h_1(z) + h_1(1-z) > h_2(z) + h_2(1-z) \tag{12}$$

The integrals have different values for $\theta \neq \frac{1}{2} \Longrightarrow$ system (10) has no solution. □