Preliminary Analysis of Multiple Kernel Learning: Flat Maxima, Diversity and Fisher Information

Theodoros Damoulas
Faculty of Computing and Information Science
Cornell University
Ithaca, 14850, NY, USA
damoulas@cs.cornell.edu

Mark Girolami
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, Scotland, UK
girolami@dcs.gla.ac.uk

Simon Rogers
Department of Computing Science
University of Glasgow
Glasgow, G12 8QQ, Scotland, UK
srogers@dcs.gla.ac.uk

Abstract

Gaining an insight into the potential effectiveness of Multiple Kernel Learning (MKL) methods requires analysis of the underlying ensemble process and resulting loss. An attempt to suggest such directions is described in this work which presents some preliminary results based on the "Flat Maximum Effect" and decompositions of the loss which leads to MKL problem being recast as one of optimal *Design of Experiments* (DoE).

1 Introduction

Multiple Kernel Learning methods aim at learning an optimal (in a predefined model-specific sense such as predictive likelihood or zero-one loss) combination of individual base kernels. Therefore such approaches follow the basic assumption that kernel combination parameter inference is crucial and beneficial over a fixed *a priori* combination. However, in a number of reported cases (Girolami and Zhong, 2007; Lewis et al., 2006) it has been observed that there is none or little such benefit. The opposite phenomenon of a significant performance improvement from an *a posteriori* combination rule has also been observed on other problems (Damoulas et al., 2008; Zien and Ong, 2007) indicating dataset dependent MKL behaviour. In this contribution we attempt to suggest preliminary ideas to provide a formal reasoning behind this phenomenon. Borrowing ideas from classifier construction analysis (Hand, 2006) and decomposition of the loss (Krogh and Vedelsby, 1995; Ueda and Nakano, 1996) we examine the conditions under which parameterised MKL methods are expected to improve over *a priori* fixed combinations. This leads to an optimal experiment design perspective on MKL.

2 Flat Maximum Effect

Hand (1997) first described the "Flat Maximum Effect" in the context of classifier performance as the phenomenon when "*often quite large deviations from the optimal set of weights will yield predictive performance not substantially worse than the optimal weights*". In Hand (2006) a set of regression coefficients are shown to be highly correlated with any other random set of regression coefficients *if* the predictor variables are correlated. This is generalised here for the kernel combination parameters and provides a starting argument for the need of *diversity* between individual information sources embedded as kernels. Consider S sources of information represented by S standardised kernels describing similarities between N training objects. For a test object $\mathbf{x}_i = \{x_1, \ldots, x_D\} \in \mathbb{R}^D$ the responses for two different convex linear combinations $\mathbf{b}$ and $\boldsymbol{\beta}$ are given by:

$$y_{\mathbf{b}} = \sum_{j=1}^{N} w_j \sum_{s=1}^{S} b_s k_s(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{S} b_s \theta_s \quad \text{and} \quad y_{\boldsymbol{\beta}} = \sum_{j=1}^{N} w_j \sum_{\tau=1}^{S} \beta_\tau k_\tau(\mathbf{x}_i, \mathbf{x}_j) = \sum_{\tau=1}^{S} \beta_\tau \theta_\tau \qquad (1)$$

The correlation between these two responses $y_{\mathbf{b}}$ and $y_{\boldsymbol{\beta}}$ is directly related and lower bounded by the correlation between the weighted base kernels:

$$\rho\left(y_{\mathbf{b}}, y_{\boldsymbol{\beta}}\right) = \frac{\sum\limits_{s,\tau=1}^{S} b_s \beta_\tau \rho\left(\theta_s, \theta_\tau\right)}{\sqrt{\sum\limits_{s,\tau,\lambda,k=1}^{S} b_s b_\tau \beta_\lambda \beta_k \Delta_{s\tau} \Delta_{\lambda k}}} \underset{\Delta_{s\tau} \leq 1 \forall s,\tau \in \mathbb{R}}{\Longrightarrow} \rho(y_{\mathbf{b}}, y_{\boldsymbol{\beta}}) \geq \sum_{s,\tau=1}^{S} b_s \beta_\tau \rho\left(\theta_s, \theta_\tau\right) \qquad (2)$$

where $\Delta_{s\tau} = \mathbb{E}\left(\theta_s \theta_\tau\right) - \mathbb{E}\left(\theta_s\right) \mathbb{E}\left(\theta_\tau\right)$ is the covariance. Hence, when the sources of information are highly correlated then any two responses y created by different sets of kernel combination weights $\mathbf{b}$ and $\boldsymbol{\beta}$ will be highly correlated, exhibiting the so called *Flat Maximum Effect* (FME) on MKL problems. FME therefore induces an insensitivity of the linear responses to tuning of the kernel combination weights, especially if sources have highly overlapping information content and are not diverse. In order to explicitly identify the role of diversity, we now examine two decompositions of the ensemble loss.

3 Ambiguity Decomposition

The first decomposition follows the ensemble regression analysis introduced by Brown and Wyatt (2003) which is an extension of the well-known bias-variance decomposition by Krogh and Vedelsby (1995). This so called *ambiguity decomposition* analysis is adopted and applied on the MKL scenario to identify the effect of diversity. Consider the (ensemble) response of the model for sample i under the standard convex linear combination of base kernels:

$$\mathbf{y}_e = \sum_{j=1}^{N} w_j \sum_{s=1}^{S} \beta_s k_s(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{S} \beta_s \mathbf{y}_s \qquad (3)$$

with the individual base kernel response defined as $\mathbf{y}_s = \sum_{j=1}^{N} w_j k_s(\mathbf{x}_i, \mathbf{x}_j)$ and the typical convex combination constraint $\sum_{s=1}^{S} \beta_s = 1$. Defining $\hat{\mathbf{y}}$ as the target regression variable and analysing the expression $\sum_{s=1}^{S} \beta_s \left(\mathbf{y}_s - \hat{\mathbf{y}}\right)^T \left(\mathbf{y}_s - \hat{\mathbf{y}}\right)$ leads to:

$$\boxed{\underbrace{\left(\mathbf{y}_e - \hat{\mathbf{y}}\right)^T \left(\mathbf{y}_e - \hat{\mathbf{y}}\right)}_{\text{Composite Error}} = \underbrace{\sum_{s=1}^{S} \beta_s \left(\mathbf{y}_s - \hat{\mathbf{y}}\right)^T \left(\mathbf{y}_s - \hat{\mathbf{y}}\right)}_{\text{Weighted Ind. Error}} - \underbrace{\sum_{s=1}^{S} \beta_s \left(\mathbf{y}_s - \mathbf{y}_e\right)^T \left(\mathbf{y}_s - \mathbf{y}_e\right)}_{\text{Ambiguity}}} \qquad (4)$$

where the first term of the right hand side is the weighted average error of individual base kernel responses and the second term is the *Ambiguity* term which describes the variability or diversity of the individual base kernels with respect to the ensemble response.

We can now see that in order to achieve a low overall composite error we need *accurate* (low weighted individual error) but *diverse* (high "ambiguity") individual base kernels with respect to the resulting ensemble.

3.1 Bias-Variance-Covariance Decomposition

The diversity requirement identified above is only with respect to any base kernel response and the overall ensemble, with no insight regarding the interaction of the base kernels. Towards that, we propose an analysis following the Bias-Variance-Covariance loss decomposition of Ueda and Nakano (1996) which is here extended to the case of the MKL ensemble output $\mathbf{y}_e$:

$$\mathbb{E}\left\{(\mathbf{y}_e - \hat{\mathbf{y}})^{\mathsf{T}} ((\mathbf{y}_e - \hat{\mathbf{y}})\right\} = \sum_{s,\sigma=1}^{S} \beta_s^{\mathsf{T}}\beta_\sigma \underbrace{\left(\mathbb{E}\{\mathbf{y}_s\} - \hat{\mathbf{y}}\right)^{\mathsf{T}} \left(\mathbb{E}\{\mathbf{y}_\sigma\} - \hat{\mathbf{y}}\right)}_{\text{Bias}^{\mathsf{T}}\text{Bias}}$$

$$+\sum_{s=1}^{S} \beta_s^{\mathsf{T}}\beta_s \underbrace{\mathbb{E}\left\{(\mathbf{y}_s - \mathbb{E}\{\mathbf{y}_s\})^{\mathsf{T}}(\mathbf{y}_s - \mathbb{E}\{\mathbf{y}_s\})\right\}}_{\text{Variance}} + \sum_{s=1,\sigma\neq s}^{S} \beta_s\beta_\sigma \underbrace{\mathbb{E}\left\{(\mathbf{y}_s - \mathbb{E}\{\mathbf{y}_s\})^{\mathsf{T}}(\mathbf{y}_\sigma - \mathbb{E}\{\mathbf{y}_\sigma\})\right\}}_{\text{Covariance}} \quad (5)$$

The above decomposition of the MKL regression loss follows other ensemble learning methods (Kittler et al., 1998; Tax et al., 2000; Ueda and Nakano, 1996) in including an additional *Covariance* term between the ensemble members (base kernels in our case) which offers an alternative description for the effect of diversity and its contribution on the overall loss. Hence now there is a trade-off not only between bias and variance of our estimates (fitting versus generalisation) but also between the latter and the covariance across base kernels (fitting versus generalisation versus diversity).

The conclusions from the FME and the loss decompositions identify the key role of diversity between base kernels but also the significance of individual accuracy (bias and weighted individual error terms). This results in the need to assess the information content and the (co)variance of the base kernel estimates which directly leads to Fisher Information and optimality criteria based on the Design of Experiments (DoE).

4 Optimal Experiment Design and Fisher Information

MKL naturally fits into the experiment design framework in that a Design Matrix has to be optimally constructed where in this case the design points are all fixed across differing data representations and the overall weighting has then to be selected. Following the *optimal experiment design* principles we seek to maximise the information content for the model parameters with respect to the evidence observed. This leads to examination of the log-likelihood curvature which expresses the variance of the latter with respect to small parameter permutations and hence acts as a measure of the information density in specific regions of the parameter space. The Fisher Information for parameters $\boldsymbol{\theta}$, evidence y and log-likelihood $\mathcal{L} = \log p(y|\boldsymbol{\theta})$ follows in standard form as:

$$\mathbf{F}(\boldsymbol{\theta}) = -\mathbb{E}\left\{\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}}\right\} \quad (6)$$

where expectation is taken with respect to the likelihood. In this section, a linear regression example for MKL is considered $\mathbf{y} = \mathbf{K}_{\boldsymbol{\beta}}\mathbf{w} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ where $\mathbf{K}_s$ is the s^{th} base kernel $\in \mathbb{R}^{N\times N}$ and $\mathbf{K}_{\boldsymbol{\beta}}$ is the composite kernel $\in \mathbb{R}^{N\times N}$. The Fisher information matrix for the parameters of the linear model is then given by $\mathbf{F}(\mathbf{w}) = \mathbf{K}_{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{K}_{\boldsymbol{\beta}}$. We now seek to find the optimal design parameters $\hat{\boldsymbol{\beta}}$ that maximize an optimal design criterion (Fisher, 1935). Following the *A-optimality* criterion of $\mathrm{Trace}(\mathbf{F}(\mathbf{w}))$ results in a constrained quadratic optimisation, with respect to $\boldsymbol{\beta}$, of

$$\texttt{argmax} \;\; \boldsymbol{\beta}^{\mathsf{T}}\Omega\boldsymbol{\beta} \;\; \texttt{s.t.} \;\; \boldsymbol{\beta}^{\mathsf{T}}\mathbf{1} = 1 \quad (7)$$

where $\Omega_{s,k} = \sum_{m,n} K^s_{m,n}K^k_{m,n}$ is an empirical estimate of the correlation between kernels (data sources). This can be employed in a probit regression setting although there will now be a conditioning on the N auxiliary variables. The formulation makes it amenable to some analysis which provides insights into the MKL problem and details of this along with examples will be presented at the workshop.

References

Brown, G. and Wyatt, J. (2003). The use of the Ambiguity Decomposition in neural network ensemble learning methods. In *International Conference on Machine Learning (ICML '03).*

Damoulas, T., Ying, Y., Girolami, M. A., and Campbel, C. (2008). Inferring sparse kernel combinations and relevant vectors: An application to subcellular localization of proteins. In *IEEE, International Conference on Machine Learning and Applications (ICMLA '08)*, pages 577–582.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

Girolami, M. and Zhong, M. (2007). Data integration for classification problems employing Gaussian process priors. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 465–472, Cambridge, MA. MIT Press.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. Wiley, Chichester.

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–15.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Krogh, A. and Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. In Tesauro, G., Touretzky, D. S., and Leen, T. K., editors, *Advances in Neural Information Processing Systems 7*, pages 231–238, Cambridge, MA. MIT Press.

Lewis, D. P., Jebara, T., and Noble, W. S. (2006). Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics*, 22(22):2753–2760.

Tax, D. M. J., van Breukelen, M., Duin, R. P. W., and Kittler, J. (2000). Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33:1475–1485.

Ueda, N. and Nakano, R. (1996). Generalization error of ensemble estimators. In *IEEE International Conference on Neural Networks*.

Zien, A. and Ong, C. S. (2007). Multiclass multiple kernel learning. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 1191–1198, New York, NY, USA. ACM.