

Bayesian Data Fusion with Gaussian Process Priors : An Application to Protein Fold Recognition*

Mark Girolami¹

Department of Computing Science
University of Glasgow
girolami@dcs.gla.ac.uk

1 Introduction

Various emerging quantitative measurement technologies are producing genome, transcriptome and proteome-wide data collections which has motivated the development of data integration methods within an inferential framework. It has been demonstrated that for certain prediction tasks within computational biology synergistic improvements in performance can be obtained via integration of a number of (possibly heterogeneous) data sources. In [1] six different parameter representations of proteins were employed for fold recognition of proteins using Support Vector Machines (SVM). It was observed that certain dataset combinations provided increased accuracy over the use of any single dataset. Likewise in [2] a comprehensive experimental study observed improvements in SVM based gene function prediction when data from both microarray expression and phylogenetic profiles were combined. More recently protein network inference was shown to be improved when various genomic data sources were integrated [3]. In [4] it was shown that superior prediction accuracy of protein-protein interactions was obtainable when a number of diverse data types were combined in an SVM.

Whilst all of these papers exploited the kernel method [5] in providing a means of data fusion within SVM based classifiers it was only in [6] that a means of estimating an optimal linear combination of the kernel functions was presented using semi-definite programming. However, the methods developed in [6] are based on binary SVM's, whilst arguably the majority of classification problems within computational biology are inherently multiclass. It is unclear how this approach could be extended to discrimination over multiple-classes. In addition the SVM is non-probabilistic and whilst *post hoc* methods for obtaining predictive probabilities are available [7] these are not without problems such as overfitting. On the other hand Gaussian Process (GP) methods [8] for classification provide a very natural way to both integrate and infer optimal combinations of multiple heterogeneous datasets via composite covariance functions within the Bayesian framework. In this paper it is shown that GP's can be employed on large scale bioinformatics problems where there are multiple data sources and an example of protein fold prediction [1] is provided.

* Supported by EPSRC Grant EP/C010620/1

2 Data Fusion with Gaussian Process Priors

Let us denote each of \mathcal{J} independent (possibly heterogeneous) feature representations, $\mathcal{F}_j(X)$, of an object X by $\mathbf{x}_j \forall j = 1 \cdots \mathcal{J}$. For each object there is a corresponding polychotomous response target variable, t , so to model this response we assume an additive generalised multinomial probit regression model. Each distinct, and possibly heterogeneous, feature representation of X , $\mathcal{F}_j(X) = \mathbf{x}_j$, is nonlinearly transformed such that $f_j(\mathbf{x}_j) : \mathcal{F}_j \mapsto \mathbb{R}$ and a linear model is employed in this new space such that the overall nonlinear transformation is $f(X) = \sum_{j=1}^{\mathcal{J}} \beta_j f_j(\mathbf{x}_j)$.

2.1 Composite Covariance Functions

Rather than specifying a functional form for each of the functions $f_j(\mathbf{x}_j)$ we assume that each nonlinear function corresponds to a Gaussian process (GP) such that $f_j(\mathbf{x}_j) \sim GP(\boldsymbol{\theta}_j)$ where $GP(\boldsymbol{\theta}_j)$ corresponds to a Gaussian process with mean and covariance functions $m_j(\mathbf{x}_j)$ and $C_j(\mathbf{x}_j, \mathbf{x}'_j; \boldsymbol{\theta}_j)$ where $\boldsymbol{\theta}_j$ denotes a set of hyperparameters associated with the covariance function. Due to the assumed independence of the feature representations the overall nonlinear function will also be a realisation of a Gaussian process defined as $f(X) \sim GP(\boldsymbol{\theta}_1 \cdots \boldsymbol{\theta}_{\mathcal{J}}, \beta_1 \cdots \beta_{\mathcal{J}})$ where now the overall mean and covariance functions follow as $\sum_{j=1}^{\mathcal{J}} \beta_j m_j(\mathbf{x}_j)$ and $\sum_{j=1}^{\mathcal{J}} \beta_j^2 C_j(\mathbf{x}_j, \mathbf{x}'_j; \boldsymbol{\theta}_j)$.

For target response values, $t \in \{1 \cdots K\}$ (i.e. a multiclass setting) and further assuming zero-mean GP functions then for N object samples, $X_1 \cdots X_N$, each defined by the \mathcal{J} feature representations, $\mathbf{x}_j^1 \cdots \mathbf{x}_j^N$, denoted by \mathbf{X}_j , and associated class specific response $\mathbf{f}_k = [f_k(X_1) \cdots f_k(X_N)]^T$ we have the overall GP prior as a multivariate Normal such that

$$\mathbf{f}_k \mid \mathbf{X}_{j=1 \cdots \mathcal{J}}, \boldsymbol{\theta}_{1k}, \cdots, \boldsymbol{\theta}_{\mathcal{J}k}, \alpha_{1k} \cdots \alpha_{\mathcal{J}k} \sim \mathcal{N}_{\mathbf{f}_k} \left(\mathbf{0}, \sum_j \alpha_{jk} \mathbf{C}_{jk}(\boldsymbol{\theta}_{jk}) \right)$$

where we employ α_{jk} to denote the positive random variables β_{jk}^2 and each $\mathbf{C}_{jk}(\boldsymbol{\theta}_{jk})$ is an $N \times N$ matrix with elements $C_j(\mathbf{x}_j^m, \mathbf{x}_j^n; \boldsymbol{\theta}_{jk})$.

A GP functional prior, over all possible responses (classes), is now available where possibly heterogeneous data sources are integrated via the composite covariance function. It is then, in principle, a straightforward matter to perform Bayesian inference with this model and no further recourse to *ad hoc* binary classifier combination methods or ancillary optimisations to obtain the data combination weights is required.

2.2 Bayesian Inference

The inference methods detailed in [9] are adopted where the auxiliary variables $y_{nk} = f_k(X_n) + \epsilon_{nk}$, $\epsilon_{nk} \sim \mathcal{N}(0, 1)$ are introduced. The $N \times 1$ dimensional vector of target class values associated with each X_n is given as \mathbf{t} where each element $t_n \in \{1, \cdots, K\}$. The $N \times K$ matrix of GP random variables $f_k(X_n)$ is

denoted by \mathbf{F} . We represent the $N \times 1$ dimensional columns of \mathbf{F} by $\mathbf{F}_{\cdot,k}$ and the corresponding $K \times 1$ dimensional vectors, $\mathbf{F}_{n,\cdot}$, which are formed by the indexed rows of \mathbf{F} . The $N \times K$ matrix of auxiliary variables y_{nk} is represented as \mathbf{Y} , where the $N \times 1$ dimensional columns are denoted by $\mathbf{Y}_{\cdot,k}$ and the corresponding $K \times 1$ dimensional rows as $\mathbf{Y}_{n,\cdot}$. The multinomial probit likelihood [9] is adopted which follows as

$$t_n = j \quad \text{if} \quad y_{nj} = \max_{1 \leq k \leq K} \{y_{nk}\}$$

and this has the effect of dividing \mathbb{R}^K into K non-overlapping K -dimensional cones $\mathcal{C}_k = \{\mathbf{y} : y_k > y_i, k \neq i\}$ where $\mathbb{R}^K = \cup_k \mathcal{C}_k$ and so each $P(t_n = i | \mathbf{Y}_{n,\cdot})$ can be represented as $\delta(y_{ni} > y_{nk} \forall k \neq i)$. Independent Gamma priors, with parameters φ_k , are placed on each α_{kj} and the individual components of $\boldsymbol{\theta}_{jk}$ (denote $\boldsymbol{\Theta}_k = \{\boldsymbol{\theta}_{jk}\}_{j=1 \dots \mathcal{J}}$), so this defines the full model likelihood and associated priors.

2.3 MCMC Procedure

Samples from the full posterior $P(\mathbf{Y}, \mathbf{F}, \boldsymbol{\Theta}_{1 \dots K}, \boldsymbol{\alpha}_{1 \dots K}, \boldsymbol{\varphi}_{1 \dots K} | X_{1 \dots N}, \mathbf{t}, \mathbf{a}, \mathbf{b})$ (where \mathbf{a} & \mathbf{b} are hyper-parameters associated with the gamma priors) can be obtained from the following Metropolis-within-Blocked-Gibbs Sampling scheme indexing over all $n = 1 \dots N$ and $k = 1 \dots K$.

$$\begin{aligned} \mathbf{Y}_{n,\cdot}^{(i+1)} | \mathbf{F}_{n,\cdot}^{(i)}, t_n &\sim \mathcal{TN}(\mathbf{F}_{n,\cdot}^{(i)}, \mathbf{I}, t_n) \\ \mathbf{F}_{\cdot,k}^{(i+1)} | \mathbf{Y}_{\cdot,k}^{(i+1)}, \boldsymbol{\Theta}_k^{(i)}, \boldsymbol{\alpha}_k^{(i)}, X_{1,\dots,N} &\sim \mathcal{N}(\boldsymbol{\Sigma}_k^{(i)} \mathbf{Y}_{\cdot,k}^{(i+1)}, \boldsymbol{\Sigma}_k^{(i)}) \\ \boldsymbol{\Theta}_1^{(i+1)}, \boldsymbol{\alpha}_1^{(i+1)} | \mathbf{F}_{\cdot,1}^{(i+1)}, \boldsymbol{\varphi}_1^{(i)}, X_{1,\dots,N} &\sim P(\boldsymbol{\Theta}_k^{(i+1)}, \boldsymbol{\alpha}_k^{(i+1)}) \\ \boldsymbol{\varphi}_k^{(i+1)} | \boldsymbol{\Theta}_k^{(i+1)}, \boldsymbol{\alpha}_k^{(i+1)}, a_k, b_k &\sim P(\boldsymbol{\varphi}_k^{(i+1)}) \end{aligned}$$

where $\mathcal{TN}(\mathbf{F}_{n,\cdot}, \mathbf{I}, t_n)$ denotes a conic truncation of a multivariate Gaussian. An accept-reject strategy can be employed in sampling from the conic truncated Gaussian however this will very quickly become inefficient for problems with moderately large numbers of classes and as such a further Gibbs sampling scheme may be required.

Each $\boldsymbol{\Sigma}_k^{(i)} = \mathbf{C}_k^{(i)} (\mathbf{I} + \mathbf{C}_k^{(i)})^{-1}$ and $\mathbf{C}_k^{(i)} = \sum_{j=1} \alpha_{jk}^{(i)} \mathbf{C}_{jk}(\boldsymbol{\theta}_{jk}^{(i)})$ with the elements of $\mathbf{C}_{jk}(\boldsymbol{\theta}_{jk}^{(i)})$ defined as $C_j(\mathbf{x}_j^m, \mathbf{x}_j^n; \boldsymbol{\theta}_{jk}^{(i)})$. A Metropolis sub-sampler is required to obtain samples for the conditional $P(\boldsymbol{\Theta}_k^{(i+1)}, \boldsymbol{\alpha}_k^{(i+1)})$. Finally $P(\boldsymbol{\varphi}_k^{(i+1)})$ is a simple product of Gamma distributions.

2.4 Obtaining Predictive Posteriors

The predictive likelihood of a test sample X_* is $P(t_* = k | X_*, X_{1 \dots N}, \mathbf{t}, \mathbf{a}, \mathbf{b})$ which can be obtained by integrating over the posterior and predictive prior such that

$$\int P(t_* = k | \mathbf{f}_*) p(\mathbf{f}_* | \boldsymbol{\Omega}, X_{1 \dots N}) p(\boldsymbol{\Omega} | X_{1 \dots N}, \mathbf{t}, \mathbf{a}, \mathbf{b}) d\mathbf{f}_* d\boldsymbol{\Omega}$$

where $\Omega = \mathbf{Y}, \Theta_{1 \dots K}, \alpha_{1 \dots K}$. A Monte-Carlo estimate is obtained by using samples drawn from the full posterior $\frac{1}{S} \sum_{s=1}^S \int P(t_* = k | \mathbf{f}_*) p(\mathbf{f}_* | \Omega^{(s)}, X_{1 \dots N}) d\mathbf{f}_*$ and the integral over the predictive prior requires further conditional samples to be drawn from each $p(\mathbf{f}_* | \Omega^{(s)}, X_{1 \dots N})$ finally yielding an estimate of $P(t_* = k | X_*, X_{1 \dots N}, \mathbf{t}, \mathbf{a}, \mathbf{b})$

$$\frac{1}{LS} \sum_{l=1}^L \sum_{s=1}^S P(t_* = k | \mathbf{f}_*^{(l,s)}) = \frac{1}{LS} \sum_{l=1}^L \sum_{s=1}^S E_{p(u)} \left\{ \prod_{j \neq k} \Phi(u + f_{*,k}^{(l,s)} - f_{*,j}^{(l,s)}) \right\}$$

2.5 Variational Approximation

From the above conditionals which appear in the Gibbs sampler it can be seen that a mean field approximation gives a simple iterative scheme which provides a computationally efficient alternative to the full sampler, details of which are given in [9]. Consider a toy dataset consisting of three classes having ten features only two of which are predictive of the class labels [9]. We can compare the time taken to obtain reasonable predictions from the MCMC and the Variational schemes. Figure 1 (a) shows the samples of the covariance function parameters Θ drawn from the Metropolis subsampler and overlaid in black the corresponding approximate posterior mean estimates obtained from the variational scheme [9]. It is clear that after 100 calls to the sub-sampler the samples obtained reflect the relevance of the features, however the deterministic steps taken in the variational routine achieve this in just over ten computational steps of equal cost to the Metropolis scheme. Figure 1 (b) shows the predictive error incurred by the classifier and under the MCMC scheme 30,000 CPU seconds are required to achieve the same level of predictive accuracy under the variational approximation obtained in 200 seconds (a factor of 150 times faster). This is due, in part, to the additional level of sampling from the predictive prior which is required when using MCMC to obtain predictive posteriors. Because of this we adopt the variational approximation for the following large scale experiment.

3 Protein Fold Prediction with GP Based Data Fusion

To illustrate the proposed GP based method of data integration a protein fold classification problem originally studied in [1] is considered. The task is to devise a predictor of 27 SCOP classes from a set of low homology protein sequences. Six different feature sets are available characterizing (1) Amino Acid composition (AA); (2) Hydrophobicity profile (HP); (3) Polarity (PT); (4) Polarizability (PY); (5) Secondary Structure (SS); (6) Van der Waals volume. In [1] a number of combination strategies were employed in devising a multiway classifier from a series of binary SVM's. The best predictive accuracy obtained on an independent set of low sequence similarity proteins was 53%. It was noted after extensive careful experimentation by the authors that a combination of Gaussian kernels each composed of the (AA), (SS) and (HP) datasets improved predictive accuracy. We employ the proposed GP based method (mean field approximation) in

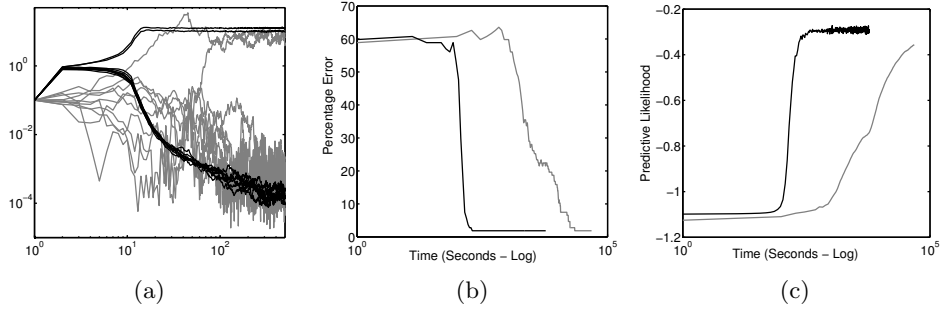


Fig. 1. (a) Progression of MCMC and Variational methods in estimating covariance function parameters, (b) the development of percentage error under the MCMC (gray) and Variational (black) schemes, (c) the development of predictive likelihood under both schemes.

devising a classifier for this task where now we employ a composite covariance function, a linear combination of RBF functions for each data set. Figure (2) shows the predictive performance of the GP classifier in terms of percentage prediction accuracy (a) and predictive likelihood on the test set (b). We note a significant synergistic increase in performance when all data sets are combined and weighted (MA). Although the test error is the same for an equal weighting of the data sets (MF) and that obtained using the proposed inference procedure (MA) for (MA) there is a small increase in predictive likelihood i.e. more confident correct predictions being made. It is interesting to note that the weighting obtained (posterior mean for α) Figure (2.c) weights the (AA) & (SS) with equal importance whilst other data sets play less of a role in performance improvement. The overall performance accuracy achieved is 62%.

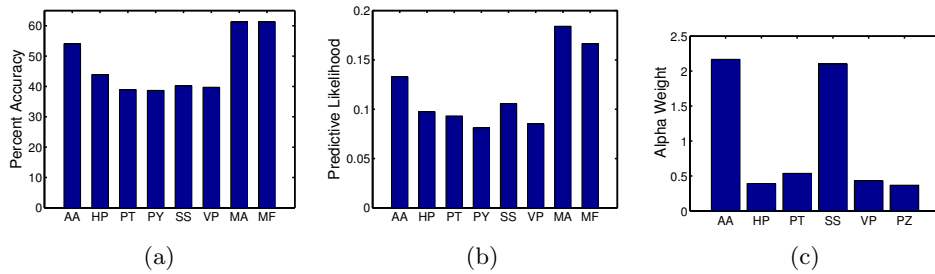


Fig. 2. (a) The prediction accuracy for each individual data set and the corresponding combinations, (MA) employing inferred weights and (MF) employing a fixed weighting scheme (b) The predictive likelihood achieved for each individual data set and with the integrated data (c) The posterior mean values of the covariance function weights $\alpha_1 \cdots \alpha_6$.

Bibliography

- [1] Ding, C., Dubchak, I.: Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **17** (2001) 349–358
- [2] Pavlidis, P., Weston, J., Cai, J., Noble, W.S.: Learning gene functional classifications from multiple data types. *Journal of Computational Biology* **9**(2) (2002) 401–411
- [3] Yamanishi, Y., Vert, J.P., Kanehisa, M.: Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* **20**, **Suppl. 1** (2004) 363–370
- [4] Ben-Hur, A., Noble, W.: Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21**, **Suppl. 1** (2005) 38–46
- [5] Shawe-Taylor, J., Cristianini, N.: Kernel methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
- [6] Lanckriet, G.R.G., Bie, T.D., Cristianini, N., Jordan, M.I., Noble, W.S.: A statistical framework for genomic data fusion. *Bioinformatics* **20** (2004) 2626–2635
- [7] Platt, J.: Probabilities for support vector machines. In Smola, A., Bartlett, P., Schlkopf, B., Schuurmans, D., eds.: *Advances in Large Margin Classifiers*, MIT Press (1999) 61–74
- [8] Rasmussen, C.E., Williams, C.K.I.: *Gaussian Processes for Machine Learning*. MIT Press (2006)
- [9] Girolami, M., Rogers, S.: Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation* **XX** (2006) XX

4 Conclusion

Kernel based methods for data integration have been previously proposed though restricted to SDP methods based on binary SVM's. In this contribution it has been shown that full Bayesian inference can be achieved for integrating multiple datasets in the multiway classification setting employing GP priors and this has been illustrated successfully with a protein fold prediction problem.