Inferring Meta-covariates Via an Integrated Generative and Discriminative Model

Keith J. Harris*
Dept. of Computing Sci.
University of Glasgow
keithh@dcs.gla.ac.uk

Lisa Hopcroft
Dept. of Computing Sci.
University of Glasgow
lisa@dcs.gla.ac.uk

Mark Girolami
Dept. of Computing Sci.
University of Glasgow
girolami@dcs.gla.ac.uk

Abstract

This paper develops an alternative method for analysing high dimensional data sets that combines model based clustering and multiclass classification. By averaging the covariates within the clusters obtained from model based clustering, we define "meta-covariates" and use them to build a multinomial probit regression model, thereby selecting clusters of similarly behaving covariates, aiding interpretation. This simultaneous learning task is accomplished by an EM algorithm that optimises a joint distribution which rewards good performance at both classification and clustering. We explore the performance of our methodology on a well known leukaemia data set.

1 Methodology

1.1 Model overview

In this paper, we present a recently developed procedure that potentially improves the interpretability and classification of high dimensional data sets, such as DNA microarray data, through the statistical coupling of a multinomial probit regression model with model based clustering. High dimensional data sets typically consist of several thousand covariates (genes in our microarray data example) and a much smaller number of samples. Analysing this data is statistically challenging, as the covariates are highly correlated, which results in unstable parameter estimates and inaccurate prediction. To alleviate this problem, we develop a statistical model which uses a small number of meta-covariates inferred from the data through a Gaussian mixture model, rather than all the original covariates, to classify samples. The advantage of this approach over using a sparse classification model [1, 2, 3] is that we can extract a much larger subset of covariates with essential predictive power and partition this subset into groups, within which the covariates are similar.

An overview of our procedure that combines model based clustering and multiclass classification is as follows. By averaging the features within the clusters obtained from a Gaussian mixture model [4, 5], we define "superfeatures" or "meta-covariates" and use them in a multinomial probit regression model, thereby attaining concise interpretation and accuracy. Similar ideas, from a non-Bayesian two-step perspective, have been looked at by [6, 7], and for binary classification by us previously [8]. With our simultaneous procedure, the clusters are formed considering the correlation of the predictors with the response in addition to the correlations among the predictors. This procedure was also partly inspired by recent empirical research that has shown that optimum predictive performance often corresponds to an intermediate trade-off between the purely generative and purely discriminative approaches to classification [9, 10]. We believe that the proposed methodology should be of general interest and have wide applicability outside of gene selection in areas such as proteomic biomarker selection, cognitive neuroscience and information retrieval.

*Inference Group website: http://www.dcs.gla.ac.uk/inference

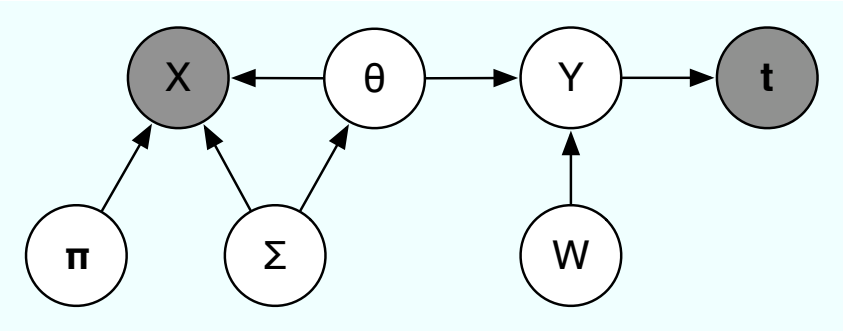

Figure 1: Graphical representation of the conditional dependencies within the meta-covariate multi-class classification model.

1.2 Model details

In the following discussion, we will denote the $N \times D$ design matrix as $X = [\mathbf{x}_1, \ldots, \mathbf{x}_D]$ and the $N \times 1$ vector of associated response values as $\mathbf{t}$ where each element $t_n \in \{1, \ldots, C\}$, where C denotes the number of classes. The $K \times N$ matrix of clustering mean parameters θ_{kn} is denoted by θ, the $K \times N$ matrix of clustering variance parameters σ^2_{kn} by Σ and the $K \times 1$ vector of mixing coefficients π_k by $\boldsymbol{\pi}$. We represent the $K \times 1$-dimensional columns of θ by $\boldsymbol{\theta}_n$ and the corresponding $N \times 1$-dimensional rows of θ by $\boldsymbol{\theta}_k$. The $D \times K$ matrix of clustering latent variables z_{dk} is represented as Z. The $K \times C$ matrix of regression coefficients is denoted by W with $\mathbf{w}_c$ denoting the $K \times 1$ vector of regression coefficients corresponding to class c and $\mathbf{w}_k$ denoting the $C \times 1$ vector of regression coefficients corresponding to cluster k. Finally, we denote the $N \times C$ matrix of auxiliary variables y_{nc} by Y, where the $N \times 1$-dimensional columns are denoted by $\mathbf{y}_c$ and the corresponding $C \times 1$-dimensional rows as $\mathbf{y}_n$.

The graphical representation of the conditional dependency structure in the meta-covariate classification model is shown in Fig. 1. From Fig. 1 we see that the joint distribution of the meta-covariate classification model is given by

$$p(\mathbf{t}, Y, X, \boldsymbol{\pi}, \theta, \Sigma, W) = p(\mathbf{t}, Y|\theta, W)p(X|\boldsymbol{\pi}, \theta, \Sigma)p(\boldsymbol{\pi})p(\theta|\Sigma)p(\Sigma)p(W). \tag{1}$$

The distribution $p(X|\boldsymbol{\pi}, \theta, \Sigma)$ is the likelihood contribution from our clustering model, which we chose to be a normal mixture model with unequal weights and diagonal covariance matrices, that is,

$$p(X|\boldsymbol{\pi}, \theta, \Sigma) = \prod_{d=1}^{D} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_d|\boldsymbol{\theta}_k, \Sigma_k), \tag{2}$$

where $\Sigma_k = \text{diag}\left(\sigma^2_{k1}, \ldots, \sigma^2_{kN}\right)$. Similarly, $p(\mathbf{t}, Y|\theta, W)$ is the likelihood contribution from our classification model, which we chose to be a multinomial probit regression model whose covariates are the means of each cluster, that is, $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$. Thus,

$$p(\mathbf{t}, Y|\theta, W) = \prod_{n=1}^{N} p(t_n|\mathbf{y}_n)p(\mathbf{y}_n|\boldsymbol{\theta}_n, W), \tag{3}$$

where

$$p(t_n|\mathbf{y}_n) = \delta(y_{ni} > y_{nc} \,\forall\, c \neq i) \text{ if } t_n = i \tag{4}$$

and

$$p(\mathbf{y}_n|\boldsymbol{\theta}_n, W) = \mathcal{N}\left(\mathbf{y}_n|W^T\boldsymbol{\theta}_n, I\right). \tag{5}$$

Finally, the model was completed by specifying a uniform prior for $\boldsymbol{\pi}$, vague inverse Gamma priors for σ^2_{kn}, and vague normal priors for θ and W. Thus,

$$p(\theta|\Sigma) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_0, h\Sigma_k) \tag{6}$$

$$p(\Sigma) = \prod_{k=1}^{K} \prod_{n=1}^{N} \frac{\xi^{\nu}}{\Gamma(\nu)} \left(\sigma_{kn}^2\right)^{-\nu-1} \exp\left(\frac{-\xi}{\sigma_{kn}^2}\right) \tag{7}$$

and

$$p(W) = \prod_{c=1}^{C} \mathcal{N}(\mathbf{w}_c|\mathbf{0}, l_c I), \tag{8}$$

where the hyperparameters $\boldsymbol{\theta}_0$, h, ν, ξ and l_c, $c = 1, \ldots C$, are chosen such that weak prior information is specified.

1.3 Summary of our inference approach

Given the number of clusters K, we would like to infer the full posterior distribution of the parameters. At the workshop we will present details of a population Markov chain Monte Carlo algorithm that allows us to sample from our inherently multimodal and high dimensional posterior distribution. However, here we present details of an EM algorithm that allows us to maximise the joint distribution with respect to the parameters (comprising the mixing coefficients, the means and variances of the clusters, and the regression coefficients):
1. Initialise $\boldsymbol{\pi}$, θ, Σ, W, the responsibilities $\gamma(z_{dk})$ and $E(Y)$, and evaluate the initial value of the log joint distribution.
2. E-step. Evaluate:

$$\gamma(z_{dk}) = \frac{\pi_k \left(\prod_n \sigma_{kn}^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_n \frac{(x_{nd}-\theta_{kn})^2}{\sigma_{kn}^2}\right\}}{\sum_j \pi_j \left(\prod_n \sigma_{jn}^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_n \frac{(x_{nd}-\theta_{jn})^2}{\sigma_{jn}^2}\right\}} \tag{9}$$

and

$$E(y_{nc}) = \begin{cases} \mathbf{w}_c^T \boldsymbol{\theta}_n - \frac{E_{p(u)}\{\mathcal{N}(u|(\mathbf{w}_c-\mathbf{w}_i)^T\boldsymbol{\theta}_n,1)\Phi_u^{n,i,c}\}}{E_{p(u)}\{\Phi(u+(\mathbf{w}_i-\mathbf{w}_c)^T\boldsymbol{\theta}_n)\Phi_u^{n,i,c}\}} & \text{if } c \neq i \\ \mathbf{w}_i^T \boldsymbol{\theta}_n - \left(\sum_{j\neq i}\left(E(y_{nj}) - \mathbf{w}_j^T\boldsymbol{\theta}_n\right)\right) & \text{if } c = i. \end{cases} \tag{10}$$

3. M-step. Evaluate:

$$\theta_{kn} = \frac{\sum_{c=1}^{C}\left(E(y_{nc}) - \sum_{k'\neq k} w_{k'c}\theta_{k'n}\right) w_{kc} + \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D}\gamma(z_{dk})x_{nd} + \frac{\theta_{0n}}{h}\right)}{\sum_{c=1}^{C} w_{kc}^2 + \frac{1}{\sigma_{kn}^2}\left(D_k + \frac{1}{h}\right)}, \tag{11}$$

$$\sigma_{kn}^2 = \frac{\sum_{d=1}^{D}\gamma(z_{dk})(x_{nd}-\theta_{kn})^2 + \frac{1}{h}(\theta_{kn}-\theta_{0n})^2 + 2\xi}{D_k + 2\nu + 3}, \tag{12}$$

$$\pi_k = \frac{D_k}{D}, \tag{13}$$

and

$$\mathbf{w}_c = \left(\theta\theta^T + \frac{1}{l_c}I\right)^{-1} \theta E(\mathbf{y}_c), \tag{14}$$

where

$$D_k = \sum_{d=1}^{D} \gamma(z_{dk}). \tag{15}$$

After updating W in this manner, set the elements of the first row and the first column of the matrix to 1, so that the model is identifiable.
4. Evaluate the log joint distribution and check for convergence. If the convergence criterion is not satisfied return to step 2.

A visual summary of our meta-covariate classification method when there is only two classes is given in Fig. 2.

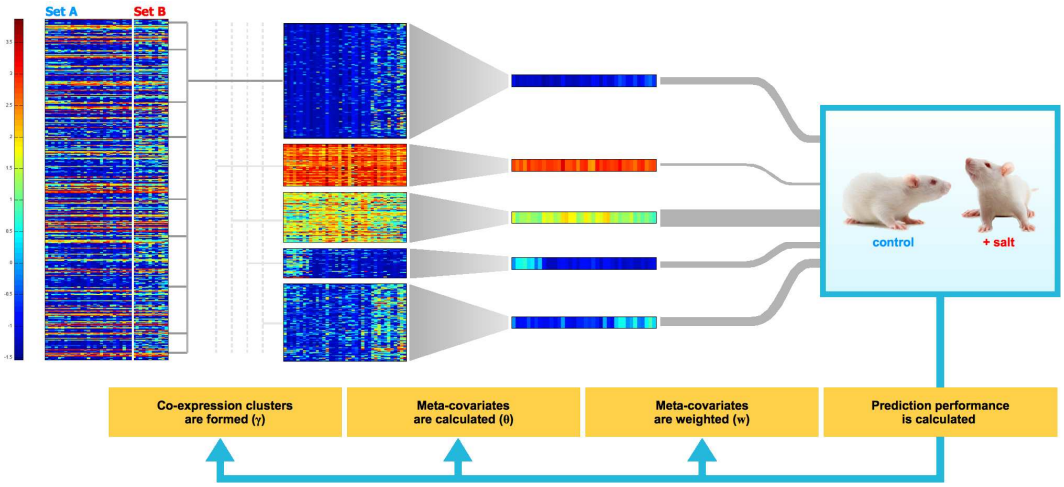

Figure 2: The meta-covariate method. Gene expression data are used to form clusters of probes (clustering is represented by the $D \times K$ matrix of responsibilities γ). N-dimensional meta-covariates ($\boldsymbol{\theta_k}$) are calculated from these clusters and used to make predictions in a probit regression model (with regression coefficients w_k). The novelty of our method is highlighted in turquoise: the predictive performance is used to update γ, $\boldsymbol{\theta_k}$ and w_k, thereby iteratively improving the clustering structure and the predictive performance.

2 A test case: leukaemia data

Leukaemia is a broad term to describe cancer of the blood or bone marrow, where blood cells are manufactured post-natally. There are many different kinds of blood cells, two of which are myeloid cells and lymphoid cells [11]. Acute myeloid leukaemia (AML) describes the rapid accumulation of abnormally proliferating myeloid cells and acute lymphoid leukaemia (ALL) describes the rapid accumulation of abnormally proliferating lymphoid cells [11]. In 1999, Golub *et al.* published work where AML and ALL samples could be classified according to their gene expression profiles [12]; using a 'weighted vote' of 50 probes, they successfully classified all but one of the samples in the test set. This data set has been subject to much analysis in the past decade and predictions made from these data are consistently good, regardless of the approach taken [1, 2, 3, 6].

Although AML and ALL are both forms of leukaemia, they occur in different types of cell [12]. As such, there will be many differences between the two samples in this data set that are attributable to cell type, rather than the molecular pathology of the two diseases. These cellular differences may be responsible for the ease with which AML and ALL are discriminated in the literature.

Nevertheless, given the many successful analyses of these data, there are clearly observable transcriptional differences between the AML and ALL samples in this data set. We can therefore employ the Golub leukaemia data set as a "proof of concept" to confirm that our meta-covariate method performs at least similarly to existing methods. And further, we might expect that myeloid or lymphoid-specific functions to dominate the functional characterisation of the most influential clusters. However, given the heterogeneity inherent in the experimental design [12] and the disease itself [13], biological interpretation will be somewhat limited.

Full details of our results for the Golub leukaemia data will be given in a future paper. Here, we will just give a summary.

One interesting finding was that there was a significant correlation between the mean variance of a cluster and its influence: perhaps contrary to expectation, the more influential clusters tend to be more variable ($\rho = 0.54, p = 0.01$). This can be explained by considering how $\boldsymbol{\theta}_k$ is calculated (see Equation 11). $\boldsymbol{\theta}_k$ is comprised of both a model mismatch component, which describes how well the current classification model matches the response data, and a standard clustering component [4]. As the size of a cluster decreases—that is, as $\boldsymbol{\gamma}_k$ becomes more sparse—the model mismatch term will dominate the calculation of $\boldsymbol{\theta}_k$, as the standard clustering component, dependent on $\boldsymbol{\gamma}_k$,

will diminish. Conversely, as the cluster becomes larger and $\boldsymbol{\gamma}_k$ becomes less sparse, the standard mixture modelling component will dominate the calculation. Furthermore, as the cluster becomes more influential and the values of $\mathbf{w}_k$ increases, the model mismatch term will dominate further. Therefore, the model will tend to form smaller, influential, more variable clusters and larger, less influential and less variable clusters, thereby automatically inducing sparsity in the model.

Our meta-covariate classification model discriminates perfectly between AML and ALL samples, in both the training and test sets. This provides good evidence that our meta-covariate model is able to predict successfully. In a future paper we will consider the most influential clusters in detail, which will illustrate the biological information our model has been able to capture.

3 Further work

Our novel statistical methodology has also been recently used to analyse data from stroke prone spontaneously hypertensive rats (SHRSP) that exhibit salt sensitivity. Our meta-covariate analysis was able to identify genes involved in transcriptional activation and circadian rhythm which may contribute to the enhanced salt sensitivity in the SHRSP compared to two other congenic strains of rats. This work will not be discussed in detail here, but will be discussed at the workshop and in a future paper.

References

[1] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, and B. K. Mallick. Gene selection: a Bayesian variable selection approach. *Bioinformatics*, 19(1):90–97, January 2003.

[2] K. Bae and B. K. Mallick. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–3430, July 2004.

[3] Y. Kim, S. Kwon, and S. Heun Song. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics and Data Analysis*, 51(3):1643–1655, December 2006.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, September 2007.

[6] B. Hanczar, M. Courtine, A. Benis, C. Henegar, K. Clément, and J. D. Zucker. Improving classification of microarray data using prototype-based feature selection. *SIGKDD Explorations*, 5(2):23–30, December 2003.

[7] M. Y. Park, T. Hastie, and R. Tibshirani. Averaged gene expressions for regression. *Biostatistics*, 8(2):212–227, April 2007.

[8] K. Harris, L. McMillan, and M. Girolami. Inferring meta-covariates in classification. In *Pattern Recognition in Bioinformatics*, volume 4, pages 150–161. Springer, September 2009.

[9] G. Bouchard and B. Triggs. The trade-off between generative and discriminative classifiers. In *IASC International Symposium on Computational Statistics (COMPSTAT)*, volume 16, pages 721–728. Physica-Verlag, August 2004.

[10] C.M. Bishop and J. Lasserre. Generative or discriminative? Getting the best of both worlds. In *Bayesian Statistics*, volume 8, pages 3–24. Oxford University Press, June 2007.

[11] A. V. Hoffbrand, P. A. H. Moss, and J. E. Pettit. *Essential Haematology*. Blackwell Publishing, 5th edition, 2006.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.

[13] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br J Haematol*, 33:451–458, 1976.