Vertical data integration for melanoma prognosis

Kaushala Jayawardana[1,4], Samuel Müller[1], Sarah-Jane Schramm[2,3], Graham J. Mann[2,3] and Jean Yang[1]

[1]School of Mathematics and Statistics, University of Sydney, NSW 2006 Australia

[2]The University of Sydney at Westmead Millennium Institute, Westmead, NSW 2145 Australia

[3]Melanoma Institute Australia, NSW 2060 Australia

[4]Corresponding author: Kaushala Jayawardana, e-mail: kaushala@maths.usyd.edu.au

Abstract

In this paper we outline the integration of clinical and omics data, to improve the prognosis capabilities of a predictive model. Traditionally, clinical data alone has been used to predict a disease outcome. However, with the generation of complex datasets from high-throughput biotechnologies, the interest of researchers has been focused on utilizing these data and the vast level of information they provide, to improve the prognosis of disease outcome. Integrating the components from different platforms has become a crucial step to better understand the relationships between clinical and omics data and the information they provide to explain/predict some response. It is an open question how to best combine different types of variables, as the large dimension of omics data can completely dominate the modelling procedure. We use clinical data from stage III melanoma patients, in a framework which combines bootstrap sampling to account for stability and multiple imputation to account for missingness in clinical data (B-MI), to produce a model with good predictive properties in unraveling the biomarkers in stage III melanoma. We exploit the availability of other high-throughput omics data on the same set of patients, more specifically, gene expression data, protein data and microRNA data, to explore methods in integrating omics data, with special focus on lasso based methods. Such an integration aims to add another dimension in understanding the predictive power of biomarkers in critical diseases like melanoma.

Keywords: omics data, prognostic model, lasso

1. Introduction

Outcome prediction is an important issue in biological studies as it aids to identify patients that could benefit from certain treatment therapies. Until recent years, phenotype data including clinical and pathological information of a patient has been used for this purpose. In recent years, with the development of high-throughput biotechnologies, interest has been increasingly focussed on using the vast level of genetic information these data provide to unveil more innate characteristics of patients to aid the prediction of disease outcome. An emerging area of research is when both types of data are used together in predicting the outcome in a vertical integration framework, as this contains the promise of improved prediction capability with more insight into a patient's characteristics. In this context, we explore the methods associated with the integration of clinical and omics data and its impact on prediction error.

Currently, there is a large number of studies in the literature on data integration; the majority holds a meta-analysis flavour where a set of statistical tools is used to combine multiple studies or data sources that answer related hypothesis (Tseng et al., 2012). This type of integration is commonly known as horizontal integration (Tseng et al., 2012), where the results from multiple studies on the same or similar data type are combined

for conclusive inferences. Tseng et al. (2012) provides a comprehensive review on horizontal genomic meta-analysis methods.

In contrast, the notion of "vertical data integration" refers to the integration of information for a common set of subjects measured from a number of distinct platforms or different molecular events such as DNA, mRNA and protein. This type of integration is often more challenging in the sense that data are obtained from very distinct platforms where the number of variables may differ extensively. For example clinical data typically has less than 100 variables and omics data has thousands of variables. Care is required such that variables from one platform do not overpower the other.

Over the past decade, some methods have been developed for vertical data integration in biological studies. However most of them focus on pairs of data sources such as clinical and expression data (Tibshirani and Efron, 2002), microarray and proteomics data (Daemen et al., 2009), gene expression and copy number data (Bergersen et al., 2011). In this study we focus on examining vertical integration in the context of outcome prediction using four distinct data sources. One of the earlier methods is "Prevalidation" (Tibshirani and Efron, 2002) where a microarray predictor is constructed and included as one extra variable alongside clinical variables. Subsequently this principle of pre-validation has been used by many studies (Boulesteix et al., 2008; van Vliet et al., 2012; Mann et al., 2013). In most of these studies the added predictive value of the additional data source is considered. Other methods used for vertical data integration include approaches based on Bayesian networks (Gevaert et al., 2006), tree-based methods (Boulesteix et al., 2008) and multivariate methods such as sparse canonical correlation analysis (Parkhomenko et al., 2007; Witten and Tibshirani, 2009).

Recently, Bergersen et al. (2011) proposed a weighted Lasso approach where additional data enters the model indirectly by acting on the penalty parameter of each variable. This approach naturally assumes there exist a primary platform and this notion is consistent with many cancer prognosis studies. The weighted Lasso approach thus avoids a further increase in the number of variables and promises to be an innovative method of integrating omics data as it allows data dependent weights to be chosen. To date, a number of methods use the weighted Lasso to introduce variable specific penalization to guide the variable selection. This includes Zou (2006) in the adaptive Lasso; Shimamura et al. (2007) in graphical Gaussian models; Charbonnier et al. (2010) in time course expression data and Garcia et al. (2013) in structured variable selection.

In our study, we propose a novel two stage feature selection approach based on the weighted Lasso to incorporate multiple omics information such as gene expression, protein and microRNA with clinical data. In addition, we compare the predictive performance of these data individually and integratively. In particular, we focus on the balance between the number of variables included from different platforms such that equal credence is given to all data sources in the predictive models. We proceed in this paper with a brief description of the data used for this study. Then we give an overview of the methods used in the analysis including our classification framework and the weighted Lasso implementation in data integration framework. This is followed by the results and discussion.

2. Data

We use a data set from stage III melanoma patients with four different data sources and the outcome prognosis is binary. Good prognosis (GP) refers to survival greater than four years and poor prognosis (PP) to survival less than one year. These data has been processed carefully to ensure strong discrimination signal in individual platforms.

Clinical data comprise of important clinical, pathological and mutation information on 21 variables and 48 patients (Mann et al., 2013), gene expression data from Illumina beadarrays on 26085 variables and 47 patient samples (Mann et al., 2013), protein data from iTRAQ platform on 897 variables and 33 samples (Mactier et al., 2013) and microRNA data from Agilent one-color arrays on 390 variables and 45 samples after the pre-processing stage. Here, 24 patient samples are common between all four platforms.

3. Classification framework

We use a novel classification framework based on multiple imputation to address missing values in clinical data and bootstrapping to address stability with special focus on models with higher predictive capability. A brief summary of our framework is as follows: the data set undergoes *m* multiple imputations and from each data set a bootstrap sample is drawn. A logistic regression is fit to each of the imputed and bootstrapped data set, where variable selection is done using BIC criterion (Schwarz, 1978). Variables are aggregated by mean estimate (Graham et al., 2007; Rubin, 1987) across the *m* data sets where the variables that have an inclusion frequency greater than 50% are selected. This procedure is repeated for B bootstrap samples (B=500 here) and the variable estimates are retained for each of the B models. The cross validation (CV) error for all of the B models is calculated and the models with CV error less than the *CV cut off* are filtered out. Considering the selected models, the inclusion frequencies of the variables in those models are calculated, and the variables are ranked according to this inclusion frequency. Variables are added to a logistic regression model in a forward algorithm, where the variable with the highest inclusion frequency enters the model first and so on. The final model is the model with the lowest prediction error rate among all models in this forward path.
Our approach is a modified version of the BMI approach (Campaign, 2011, chap. 3), where we modify the approach to select models with higher predictive power. Another recent study that focuses on bootstrap samples and multiple imputation in variable selection is Schomaker and Christian (2013).

4. Vertical data integration

In the vertical data integration framework, it is crucial that equal credence is given to all data sources, such that the modelling procedure is not dominated by the platform size. Pre-validation (Tibshirani and Efron, 2002) is an example of extreme variable reduction where a single microarray predictor is derived to be modelled with clinical variables. Here, thousands of microarray variables are reduced to a single variable. On the other extreme, the integration of clinical and omics data could also be done in a single stage with no prior variable reduction where the weighted Lasso is performed on all clinical and gene expression variables (Jayawardana et al., 2013). Here weights based on protein/microRNA data are given to gene expression variables and suitable weights are given to clinical variables so that the large dimension of gene expression data do not dominate the modelling procedure.

We propose a two stage feature selection approach which focuses on an intermediate variable reduction so that the gene expression data and clinical data has an equal standing. In the first stage we select the genes based on protein and microRNA information using the weighted Lasso and then integrate the selected genes and clinical variables to find the final predictive model. In the following section we give a brief overview of this methodology.

Weighted Lasso with data integration

In high dimensional omics data settings where we have larger number of variables than samples, many of the standard statistical methods fail to work. Lasso (Tibshirani, 1996) based methods, because of their penalization properties, became widely used in simultaneous estimation and variable selection. One such method is the weighted Lasso where data dependent weights are used in the penalization such that the penalty parameter varies for each covariate.

Suppose $\boldsymbol{y}$ is the response vector, $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p}]$ is the predictor matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, λ is a regularization parameter and w_j are data specific weights. Here the objective function to be minimized is:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|.$$

The adaptive Lasso (Zou (2006)) is a special case of the weighted Lasso where the weights $\hat{w}_j = 1/|\hat{\beta}_j|^{\gamma}$; $\gamma > 0$ and $\hat{\beta}_j$ denotes the ordinary least squares estimate after regressing $\boldsymbol{y}$ on $\boldsymbol{X}$. Furthermore, recent studies show that more stable results can be generated for the Lasso, when weights based on relevant external information or prior knowledge of the variables are used in penalty parameter (Charbonnier et al., 2010; Bergersen et al., 2011).

Two stage feature selection approach

Our implementation of the weighted Lasso is based on these studies that incorporate external information to guide the variable selection, so that our selected variables reflect information from multiple data sources and thus are more stable and accurate (Bergersen et al., 2011).

In the first stage of our integration framework we use the weighted Lasso to select genes that contain the information from protein and microRNA data. Here the weights used are $w_j = 1/|\eta_j|$ where η_j is the median Spearman correlation coefficient of gene_j and proteins or microRNAs. This choice was motivated by Bergersen et al. (2011), where they used the Spearman correlation coefficient to constitute weights in integrating gene expression and copy number data. In the case where we integrate all three omics platforms, $\eta_j = \frac{\eta_{j1}+\eta_{j2}}{2}$; the average of the correlations for each gene based on protein ($\text{cor}(gene_j, prot) = \eta_{j1}$) and microRNA ($\text{cor}(gene_j, miRNA) = \eta_{j2}$) data. It is conjectured that the genes whose expression values are highly correlated with the expression values of proteins are more relevant in predicting the outcome of patients.

In the second stage of our integration framework, we examine two ways of integrating clinical and gene expression data with and without variable selection. The first method involves further variable selection where we use the selected genes from the weighted Lasso on gene expression data and all the clinical variables as initial features in our modified BMI framework (Section 3). A final predictive model with a subset of genes and/or clinical variables will be determined. The second method does not involve variable selection where we use the pre-selected clinical variables from modified BMI and pre-selected genes from the weighted Lasso in a logistic regression model to get the final prediction error. This procedure is illustrated in Figure 1.

5. Results and Discussion

In this study we investigate the effect of using multiple data sources together to predict the outcome of melanoma patients. We use a data integration framework with the weighted Lasso using four relevant and distinct data sources of melanoma patients.

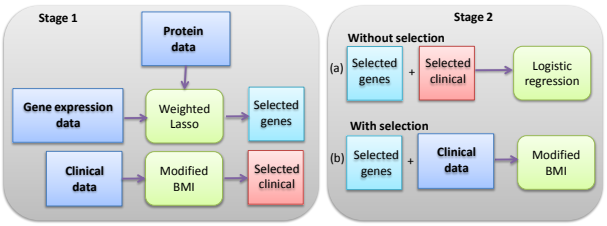

Figure 1: *Two stage weighted Lasso procedure:* A graphical illustration of our two stage feature selection procedure using the weighted Lasso

In our analysis, we computed the mean 5-fold cross validation (CV) error rate over 100 runs of the models in all settings to account for the variability common in most real data. In the individual platform analysis, we used the modified BMI framework for clinical data and for omics data we selected the best classifier and the molecular signature (set of features) based on the lowest prediction error rate. The mean 5-fold CV error rate for individual platform models ranged between 25% and 40% with gene expression data providing the lowest prediction error for our data set. For more details on individual platform analysis we refer to Jayawardana et al. (2013).

In pre-validation, which is one of the extreme variable reduction approaches, the integration yielded 25% to 35% mean 5-fold CV errors with improvement in prediction accuracy over the individual platform analysis in most cases. In our two stage approach, in the first stage we repeat the weighted Lasso on genes 100 times and select the genes with over 50% inclusion frequency in all models to ensure stability. In the second stage we modelled the selected genes with clinical variables in two settings. When we perform the procedure with variable selection in our modified BMI framework, the mean 5-fold CV error ranged between 10% to 25%. When the procedure is repeated without variable selection, the mean 5-fold CV error ranged between 10% to 20%.

In both settings the prediction error was significantly lower than that of the models from the individual analysis and from the pre-validation method. This shows that the weighted Lasso is a very useful approach that allows to maintain the balance between the number of variables in multiple data sources in an integrative framework, without increasing the number of variables excessively. Hence, this approach is very constructive in addressing the challenge of integrating distinct data sources.

It should be noted that although some variables are not included in the final model this does not give a clear implication of their importance in predicting melanoma. There could exist equivalent models with respect to their predictive accuracy and the effect of important variables could be masked by another group of variables because of the collinearity in the variables and is an interesting topic of further research.

Our data integration framework using the weighted Lasso greatly improves the prediction accuracy of stage III melanoma patients. We conclude our study bringing to attention that the usage of information from different stages in the flow of genetic information is very beneficial as it reflects the complete biology underlying a disease. It also accounts for the error introduced by missing information in one data source and might reduce false positives than when using the information from a single data source.

References

Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Stat Appl Genet Mol Biol*, 10(1).

Boulesteix, A. L., Porzelius, C., and Daumer, M. (2008). Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698–1706.

Campaign, A. (2011). *Challenges associated with clinical studies and the integration of gene expression data*. PhD thesis, School of Maths and Stats, University of Sydney.

Charbonnier, C., Chiquet, J., and Ambroise, C. (2010). Weighted-LASSO for structured network inference from time course data. *Stat Appl Genet Mol Biol*, 9:Article 15.

Daemen, A., Gevaert, O., Ojeda, F., et al. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome Med*, 1(4):39.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, 33(1):1–22.

Garcia, T. P., Müller, S., Carroll, R. J., et al. (2013). Structured variable selection with q-values. *Biostatistics*. doi:10.1093/biostatistics/kxt012.

Gevaert, O., De Smet, F., Timmerman, D., et al. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14):184–190.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev Sci*, 8(3):206–213.

Jayawardana, K., Müller, S., Schramm, S., et al. (2013). Data integration and model selection for melanoma prognosis. *Preprint*.

Mactier, S., Kaufman, K. L., Wang, P., et al. (2013). Identification of prognostic markers for survival in lymph node metastases from Stage III melanoma patients. *J. Proteome Res. Submitted*.

Mann, G. J., Pupo, G. M., Campain, A. E., et al. (2013). BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. *J. Invest. Dermatol.*, 133(2):509–517.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc*, 1 Suppl 1:S119.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Chapman & Hall/CRC, New York.

Schomaker, M. and Christian, H. (2013). Model selection and model averaging after multiple imputation. *Comput Stat & Data An*. doi:10.1016/j.csda.2013.02.017.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Ann Stat*, 6(2):461–464.

Shimamura, T., Imoto, S., Yamaguchi, R., et al. (2007). Weighted lasso in graphical Gaussian modeling for large gene network estimation based on microarray data. *Genome Inform*, 19:142–153.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288.

Tibshirani, R. J. and Efron, B. (2002). Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol*, 1:Article1.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.*, 40(9):3785–3799.

van Vliet, M. H., Horlings, H. M., van de Vijver, M. J., et al. (2012). Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS One*, 7(7). doi:10.1371/journal.pone.0040358.

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol*, 8:Article28.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.