SAND2015-0149R

Sandia National Laboratories

Tissue Classification

David Robinson, 01461

12 January, 2015

1 Problem Statement

The project began as a effort to support InLight and Lumidigm. With the sale of the companies to a non-New Mexico entity, the project then focused on supporting a new company Medici Technologies.

The Small Business (SB) is attempting to quantify glucose in tissue using a series of short interferometer scans of the finger. Each scan is produced from a novel presentation of the finger to the device.

The intent of the project is to identify and, if possible, implement improved methods for classification, feature selection, and training to improve the performance of predictive algorithms used for tissue classification. The initial task involves investigation of alternative methods for direct prediction of the analyte or tissue component of interest. The second task will develop and demonstrate improved classification capabilities on actual tissue data sets.

Deliverables include knowledge transfer of the methods investigated, results generated, and the code used for processing of the data.

2 Background

2.1 Analysis Background

The analysis by the SB was based on a linear discriminate analysis (LDA) and ensembles of decision trees (like random forest) applied to the data for the SB [Description of Data for Classifier Training v2.docx provided by SB]. For each subject, repeated measurements of the glucose level and the spectra are available. Figure 1 depicts simulated spectra from a single measurement for 20 subjects.

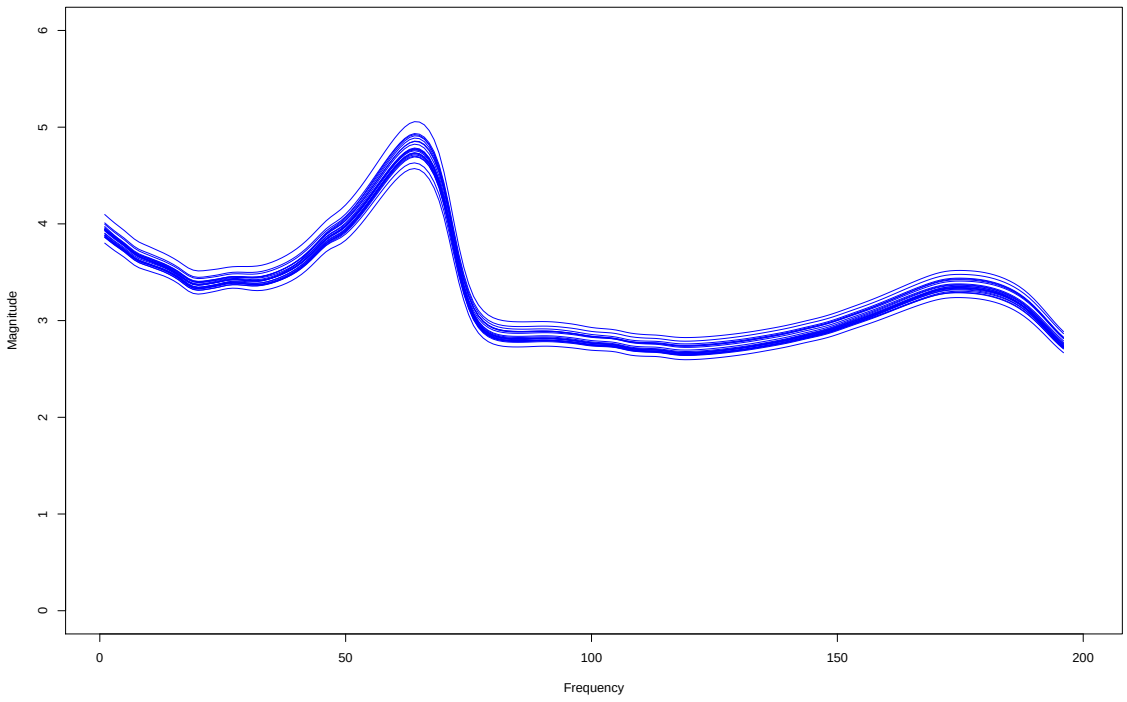

Figure 1: Simulated Spectra for Twenty Subjects

The SB has found that the use of these standard methods for multi-class classification and feature selection have proven inadequate.

Fundamental to this approach is that each class or cluster is associated with a unique linear model for glucose prediction. These classification methods

result in a 'crisp' classification of subjects in to clusters. Crisp classification indicates that the subjects are strictly in a single cluster; there is no overlap between the membership in clusters. Subjects with a feature set close to those already in a class are presumed to be characterized by the glucose prediction model for that class.

Based upon the experience of the SB with the data, they observed the following limitations of their approach. First, the use of linear models did not appear to capture the true nature of the class boundaries encountered in the data. Second, feature selection methods dont appropriately deal with the multi-class classification case as the features optimal for the discrimination between two classes, say A and B, are not necessarily the best features for the discrimination between any other classes. Third, classes are often not well balanced and due to the error structure one or more solutions may be acceptable but from a training perspective this introduces noise.[Description of Data for Classifier Training v2.docx provided by SB]

3 Alternative Approaches

As noted previously, the second task of the effort is to develop and demonstrate improved classification capabilities on actual tissue data sets. Data sets were provided too late in the project to implement the methods described below. However sufficient detail is provided that will permit implementation of the methods in either the JAGS or Stan open source programming languages.

The methods employed by the SB are common, popular approaches for classification. However, they are limited in many respects. In particular, they require the possibility of a linear boundary between classes; that is, the classes are distinct and linearly separable.

Modern alternative methods allow for feature sets to have partial membership across multiple classes. The result is that each subject may have a degree of membership in each class. This implies that the glucose prediction model for each subject could be a linear mixture of linear models.

3.1 Extension to Existing SB Approach

First consider the situation where there are N classes. In the existing SB approach a subject feature set is in either one or the other class. Alternatively,

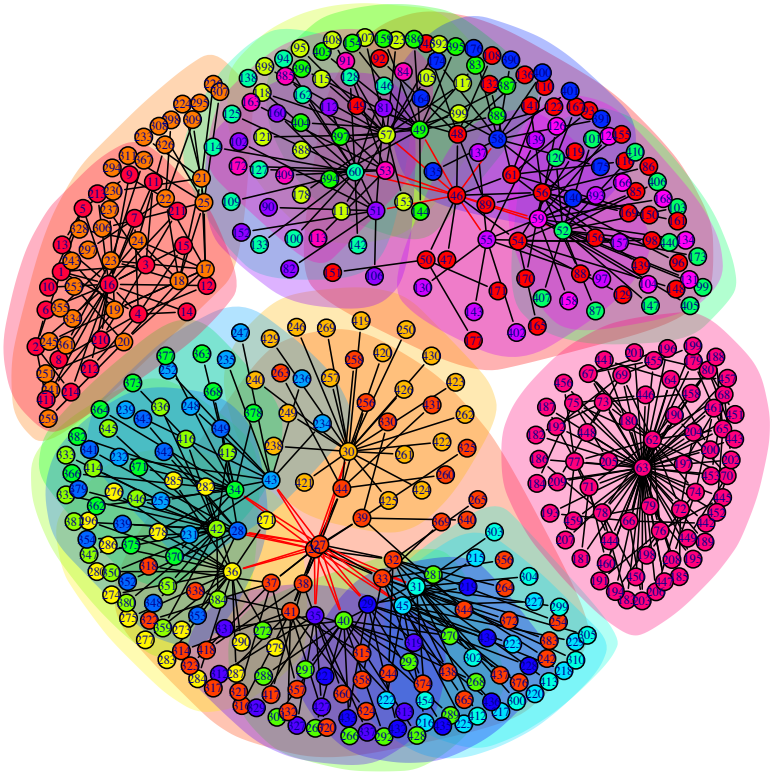

Figure 2: Crisp Tissue Classification

consider relaxing this constraint such that there is a probability $p_1, p_2, \ldots, p_N$ that a a feature set is in one class or the other and $\sum_N p_i = 1$.

The only distributional assumption with this method is that the log likelihood ratio of class distributions is linear in the observations. This assumption is verified by a large range of exponential density families, e.g. normal, beta, gamma, etc. Further, the approach:

- does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables.
- can be used to analyze relationships between a dichotomous (categorical) dependent variable and metric or dichotomous independent variables. The dependent variable can take on multiple values e.g. {Success, Failure}, {red, blue, green, purple}, etc. Independent variables may be either interval or categorical.
- estimates the probability that a particular set of values for the independent variables is a member of a response category.

The minimum number of cases per independent variable is 10, using a guideline provided by Hosmer and Lemeshow, while for stepwise logistic regression a 50-1 ratio is preferred. The SB has multiple, repeated observations for subject. Given:

$$\log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \beta_0 + \beta_1 \mathbf{x} + \cdots \tag{1}$$

where f_i are class conditional parametric densities and β are model parameters. Equivalently:

$$f(\mathbf{x}) = g\left(\beta_0 + \beta_1 \mathbf{x} + \cdots\right) \tag{2}$$

where: $g(z)$ is the logistic function $g(z) = 1/\left(1 + \exp^{-z}\right)$ and $f(\mathbf{x}) = P(y = 1|\mathbf{x})$ is the probability of being in class 1. The likelihood function for the data:

$$L(D, \beta) = P(D|\beta) = \prod_{j=1}^{n} P(y = y_j|\mathbf{x}_j, \beta)$$

where $\mathbf{x}_j$ is the vector of independent variables associated with the j-th observation y_j.

Let $\pi_j = p(y_j = 1|\mathbf{x}_j, \beta) = g(z_j) = g(\beta^T \mathbf{x})$. The likelihood function can then be expressed:

$$L(D, \beta) = \prod_{j=1}^{n} P(y = y_j|\mathbf{x}_j, \beta) = \prod_{j=1}^{n} \pi_j^{y_j}(1 - \pi_j)^{1-y_j}$$

The log-likelihood is often more convenient: $l(D, \beta) = \log L(D, \beta)$. Our goal is to maximize the likelihood function

$$L(D, \beta) = \prod_{j=1}^{n} P(y = y_j|\mathbf{x}_j, \beta) = \prod_{j=1}^{n} \pi_j^{y_j}(1 - \pi_j)^{1-y_j}$$

or, equivalently, minimize the error function: $E(D, \beta) = -\sum_i l(D_j, \beta)$. Figure 3 depicts the conditional relationships between each of the variables in a hierarchical Bayesian logistic regression.

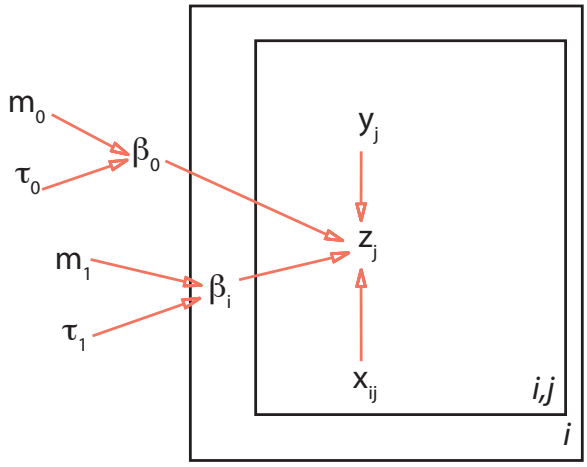

Figure 3: Hierarchical Model for Logistic Regression Classification

3.1.1 Logistic Regression with Measurement Error

Traditional regression models assume that the covariates are observed without error; the only error being in the response. Alternatively, measurement error models are regression models where the covariates are measured with error.

In the case of interest to the SB, the possibility exists that spectral features are measured with error. In particular, the measurement error problems arise

if instead of observing the covariate without error, at least one is measured as an error-prone surrogate. Such problems are primarily concerned with inference on regression parameters for an outcome Y on covariates X where measurements on X are available only through the recording of an imperfect surrogate w. It is well known that regressing Y on W and thus ignoring errors can be seriously misleading.

Statistical models must be fit to data formulated in terms of well-defined but unobservable variables X, using information on measurements W that are less than perfectly correlated with X. W is said to be a surrogate for X. There are a number of ways in which X and W can be related.

- Classical Error $W = X + U$, $E(W|X) = X$, $E(U) = 0$, U is independent of X
- Berkson Error $X = W + U$, $E(X|W) = W$, $E(U) = 0$, U is independent of X

A classical rrror model is commonly assumed. Figure 4 depicts the conditional relationships between each of the variables in a logistic regression accounting for observational error.

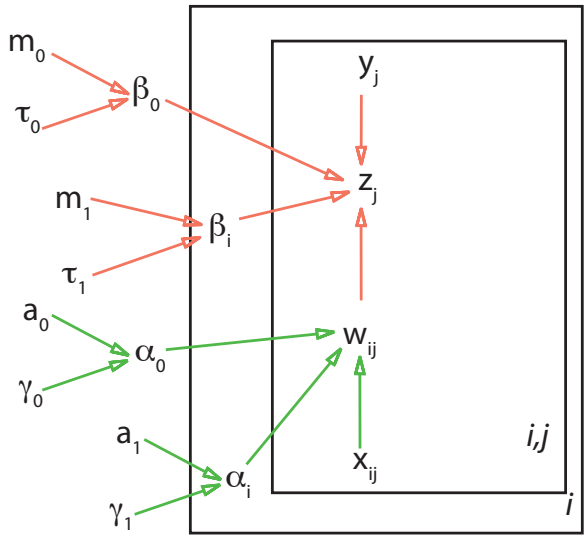

Figure 4: Hierarchical Model with Measurement Error

Formally, we are trying to predict the probability that a feature set is in class 1: $Y = 1$.

$$P\{Y = 1|w_1, w_2, ...\} = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 w_1 + \beta_2 w_2 + \ldots)]} \quad (3)$$

$$= \frac{\exp(\beta_0 + \beta_1 w_1 + \beta_2 w_2 + \ldots)}{1 + \exp(\beta_0 + \beta_1 w_1 + \beta_2 w_2 + \ldots)} \quad (4)$$

The summary of the hierarchical model with typical assumptions regarding the prior distributions:

$$\begin{aligned} Y_j &\sim Bern(\pi_j) \\ \log\left(\frac{\pi_j}{1-\pi_j}\right) = Z_i &= \beta_0 + \beta_1 w_1 + \beta_2 w_2 + \ldots, i = 1, \ldots, n \\ \\ X_j &\sim Bern(\pi_{x_j}), j = 1, \ldots, n_c \\ W_{x_j} &= \alpha_0 + \alpha_j X_j \\ \pi_{x_j} &= 1/[1 + \exp(-W_{x_i})] \\ W_j &\sim Bern(\pi_{w_j}) \\ \\ E_k &\sim Cat(n_j, \pi_w), k = 1, \ldots, n_e \\ \\ \alpha_0, \alpha_j &\sim N(0, 0.001) \\ \beta_0, \beta_i &\sim N(0, 0.001) \end{aligned}$$

The following section describes an alternative method for direct prediction of the analyte or tissue component of interest. This was the second task of the effort.

4 Direct Prediction

As summarized previously, the intent of the project was to identify and, if possible, implement improved methods for classification, feature selection, and training to improve the performance of predictive algorithms used for

tissue classification. In particular, the second task of the effort was to develop and demonstrate the improved classification capabilities on either simulated or actual tissue data sets.

4.1 Approach

The approach described here is sometimes referred to in regression analysis as a *random effects* analysis. However, in Bayesian analysis this can be a bit of a misnomer since the parameters are already assumed to be random.

Since the mathematics behind this approach are well established, the discussion here provides a brief introduction. A thorough discussion to the methods is provided in: [3, 2, 1].

It is assumed that the relationship between glucose measurements and the spectral features involves nonlinear relationship where the parameters of the function are random variables. The relationship between the variables is conceptually the same as that depicted in Figure 3. The difference is that the response is not the probability of class membership, but the actual glucose observation.

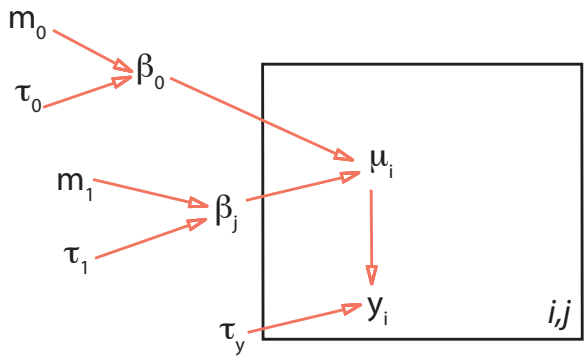

Figure 5: Hierarchical Model for Prediction

Hierarchical model summary with typical assumptions regarding the prior distributions:

$$\begin{aligned} Z_j &\sim N(\mu_i, \tau) \\ \mu_i &= \beta_{0i} + \beta_{1i} w_1 + \beta_{2i} w_2 + \ldots, i = 1, \ldots, n \end{aligned}$$

$$\begin{aligned}
\beta_j &\sim N(\mu_i, \tau_i), j = 1, \ldots, 196 \\
\tau &\sim Uniform(0, 10^{-4}) \\
\alpha_0, \alpha_j &\sim N(0, 0.001) \\
\beta_0, \beta_i &\sim N(0, 0.001)
\end{aligned}$$

where n is the total number of observations, and 196 refers to the number of spectra available (size of the feature space).

As with the classification model discussed previously, it is possible to extend this model to include the possibility of error in measuring the spectra. (Error in the glucose measurement is already accounted for in the model above.) Insufficient information regarding the statistical characteristics of the spectra was available (it has not been characterized by the SB). Unfortunately, data was provided by the SB too late in the effort to implement and test the above algorithms.

5 Acknowledgements

Prepared by Sandia National Laboratories, Albuquerque, New Mexico 87185 and Livermore, California 94550. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] A. Brasoveanu. Estimating general random-effect structures with jags. http://people.ucsc.edu/ abrsvn/general-random-effects-jags.html, 2014.

[2] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis*. CRC Texts in Statistical Science. Chapman and Hall, third edition, 2013.

[3] J. K. Kruschke. *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*. Elsevier, second edition, 2015.