

Chapter 17

Analysis and Predictive Modeling of Asthma Phenotypes

Allan R. Brasier and Hyunsu Ju

Abstract Molecular classification using robust biochemical measurements provides a level of diagnostic precision that is unattainable using indirect phenotypic measurements. Multidimensional measurements of proteins, genes, or metabolites (analytes) can identify subtle differences in the pathophysiology of patients with asthma in a way that is not otherwise possible using physiological or clinical assessments. We overview a method for relating biochemical analyte measurements to generate predictive models of discrete (categorical) clinical outcomes, a process referred to as “supervised classification.” We consider problems inherent in wide (small n and large p) high-dimensional data, including the curse of dimensionality, collinearity and lack of information content. We suggest methods for reducing the data to the most informative features. We describe different approaches for phenotypic modeling, using logistic regression, classification and regression trees, random forest and nonparametric regression spline modeling. We provide guidance on post hoc model evaluation and methods to evaluate model performance using ROC curves and generalized additive models. The application of validated predictive models for outcome prediction will significantly impact the clinical management of asthma.

Keywords Multivariate analysis • Supervised learning • False discovery rate • Feature reduction • Significance of microarrays (SAM) • Receiver operating characteristic (ROC) curve • Logistic regression • Random forest • Multivariate adaptive regression splines (MARS) • Generalized additive models (GAMs)

A.R. Brasier, M.D. (✉)

University of Texas Medical Branch, 8.128 Medical Research Building,
301 University Blvd, Galveston, TX, USA
e-mail: arbrasie@utmb.edu

H. Ju, Ph.D.

University of Texas Medical Branch, 1.134 Ewing Hall, 301 University Blvd,
Galveston, TX, USA
e-mail: hyju@utmb.edu

17.1 Introduction

The application of multidimensional profiling is a disruptive technology that enables identification of molecular phenotypes of asthma (Bhavnani et al. 2011; Brasier et al. 2008, 2010). When properly conducted and interpreted, multidimensional measurements of proteins, genes, or metabolites can identify subtle differences in the pathophysiology of asthma in a way that is not otherwise possible using physiological or clinical assessments. The molecular profiling approach holds promise for revolutionizing the diagnosis and management of asthma because it can identify robust objective, measurable, and quantitative features associated with disease phenotypes, disease progression, and treatment response.

Currently, the diagnosis and management of asthma is based on what could be considered coarse physiological assessments. Phenotypic classifications of asthma using age of onset, exacerbating factors (aspirin sensitivity, exercise induced, and occupational) cellular type of inflammation (eosinophilic vs. noneosinophilic), pattern of severity, or lack of response to conventional therapy provide little insight into mechanisms of disease or strategies for its management. In particular, the conventional ATS criteria for the diagnosis of severe asthma is based on a constellation of findings and clinical features that is better considered a syndrome than a specific disease diagnosis (ad hoc writing committee of the Assembly on 2000; Wenzel and Busse 2007). In the absence of biochemical marker measurement, the multiple underlying pathological processes that contribute to the clinical syndrome of asthma, such as allergy, inflammation, airway remodeling, smooth muscle hypertrophy, and neurogenic dysregulation, are not considered. This situation leads to the obvious result that patients carrying the diagnosis of asthma will exhibit heterogeneous clinical courses and variable medication responses. As a result, there is no reliable method for early recognition of treatment-resistant subtypes. Precise quantification of these processes at the biochemical level would significantly impact clinical management.

The ability to use precise molecular measurements to identify subtypes of disease is an exciting arena that opens the door for applications of personalized medicine. Approaches for the analysis of complex molecular profiling can be broadly broken into two distinct disciplines that have complementary, but distinct, outcomes. Supervised learning, or classification, is an analysis in which profiling is conducted in an experimental design where the outcome (phenotype) is known. Classification employs analytical techniques that use the outcomes of test cases for the generation of predictive models that classify outcome of unknown cases. One example of supervised learning would be to predict the risk of disease, or response to treatment. Unsupervised learning, by contrast, is conducted when the phenotype is not known and employs techniques to discover emergent characteristics of the data. An example of unsupervised learning would be hierarchical clustering, where natural groupings in the data are sought. This chapter focuses on the approach of classification (supervised learning). The accompanying chapter by Bhavnani et al. (Chap. 18) will treat the issue of unsupervised learning.

Here, we describe approaches for how high-dimensional data sets can be analyzed and used to develop models that link analyte measurement to discrete (categorical) outcomes. We consider problems inherent in “wide” (small n and large p) high-dimensional data, including the curse of dimensionality, collinearity, and lack of information content. We suggest methods for reducing the data to the most informative features. We describe different approaches for molecular phenotyping using logistic regression, classification and regression trees, random forest and nonparametric spline modeling. We provide guidance on post hoc model evaluation and discuss advantages and limitations of this approach.

17.2 Analytic Approach

17.2.1 Data Visualization

The focus of this discussion is how to develop robust models of physiological phenotypes. Analytic approaches for high-throughput data should recognize some of the unique features of high-dimensional data. Prior to the application of any modeling technique, high-dimensional data sets should be carefully analyzed by visualization and conventional statistical approaches to ensure data quality, detect missing data, identify outliers, determine data distributions and suggest batch effects (Spratt et al 2012). This approach is schematically diagrammed in Fig. 17.1.

17.2.2 Data Preprocessing

Data preprocessing is a method to manage high-dimensional data prior to classification approaches. High-dimensional data frequently may contain missing values, exhibit measurements with a broad dynamic range, and exhibit skewed population distributions.

The presence of missing data is characteristic of high-throughput data sets. Missing data frequently arises in mass spectrometry, for example, where observations are not made in every sample due to sampling bias, or in immunoassays, where the analyte concentration is below the limit of detection. The presence of a significant amount of missing data complicates many statistical analytic and machine learning approaches. If a few data points are missing, imputation can be used, where missing values are substituted with reasonable numerical values. However, if significant amount of data is missing, we frequently will remove the analyte from consideration, because imputation then can introduce bias.

Several different approaches can be used for imputing missing values, the selection of which depends on the data type and understanding of why the data is absent. In highly quantitative assays, such as immunoassays, values may be missing because

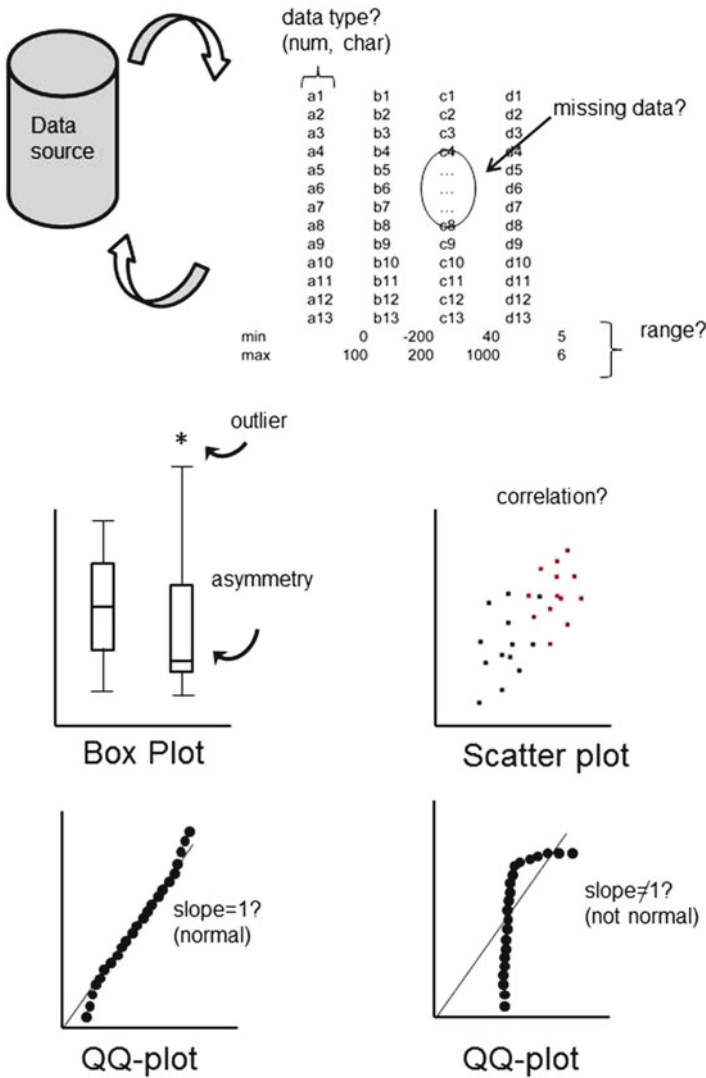


Fig. 17.1 Data preprocessing. A schematic for the general approach to the analysis of a high-throughput data set. Shown is an example of a spreadsheet example with common problems identified by data visualization. The spreadsheet contains missing values (...). Identification of data type, and distribution of data can be performed using min and max analysis, or box-plot determinations. In a box-plot determination, the 25–75 % quartile is shown as well as the mean. Inspection of the *box-plot* suggests nonnormally distributed data sets. Also illustrated are *q-q* plots that suggest data that is normally distributed (*left*) and data that is not (*right*)

a measurement is below the limit of detection. In this case, an appropriate imputation method could involve substitution of a value below the assay's lower limit of quantification. In earlier studies of bronchoalveolar lavage cytokines, we evaluated the effect of substituting a value of 1/10th of the lower limit of detection for each assay. This conservative estimate of the analyte measurement has been widely used (Brasier et al. 2008, 2011). Another imputation technique, appropriate for missing values that could be any value, is K nearest neighbors (KNN); KNN is a technique where values are imputed by an algorithm in which neighbors are the variables of interest and the distance between neighbors is based on the correlation between them (Troyanskaya et al. 2001) (KNN implementations are available in R package and Matlab program).

Another issue in data preprocessing for classification is how to treat data that fails analysis of normality. Typical normality tests would include box-plot or q - q graphical analysis (Spratt et al. 2012). This point is important because comparing feature expression in a nonnormally distributed data set requires the use of less powerful nonparametric analytical approaches. Because parametric testing is more powerful, where possible, we will try to employ statistical manipulations to make the data set behave more like normally distributed data. Popular data transformations include taking the logarithm of the data (log base 10 or natural logarithm) or taking the square root of the data. Log transformation is an approach widely used in gene expression analysis and protein abundance estimation.

17.2.3 Feature Reduction Strategies

A characteristic of complex, high-dimensional data sets is that many analytes are identified, only some of which are informative features, whereas others only contribute noise. The measurement of multiple features leads to the “curse of dimensionality,” where the presence of many features often leads to poorer, rather than better, classifier performance. One potential solution to this problem is to reduce the dimensionality and simplify model complexity. This approach is frequently referred to as feature reduction and schematically diagrammed in Fig. 17.2.

Feature reduction seeks to reduce dimensionality by selecting an informative subset from the existing set of analytes. In practice, feature reduction can be accomplished by first selecting a subset of the most informative features that distinguish the two groups. This selection can be done by rank-ordering analytes by statistical significance based on groupwise comparison, by calculating the fold change between the two comparison groups (ratio of group means), by a statistical analysis of microarrays (SAM), Bayes methods or heuristics (biological intuition of the investigator). Selection of the feature reduction method is a decision made on the basis of the data set characteristics, the goals of the analysis and/or heuristics. Some considerations for this decision are shown in Table 17.1.

Statistical methods for feature reduction is approached by comparing different group means either using the two sample t -test or Wilcoxon rank sum test at a given appropriate level of significance. The Wilcoxon rank sum test is an alternative

Fig. 17.2 Model development and assessment. A schematic for the use of feature reduction and downstream supervised or unsupervised classifications. Abbreviations: *CART* classification and regression trees, *RF* random forests, *SVM* support vector machines, *MARS* multivariate adaptive regression splines

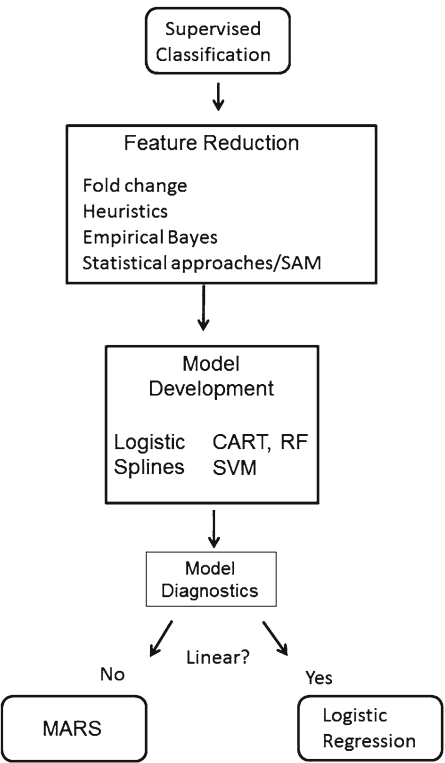


Table 17.1 Selection of feature reduction techniques

Technique	Considerations
Rank ordered <i>t</i> -test	• Useful for large sample number (>30) • Sensitive to features with small group mean variance
Fold change (FC) cut-off	• Cut-off value selection can be seen as arbitrary • Sensitive to error if class is unmeasurable
SAM	• Appropriate for small sample number or correlated measurements • Enables FDR estimation • Must specify comparator for group differences
Empirical Bayesian	• Model uncertainty; model misspecification • Limited available software options (R or Winbugs); long time to compute; data structure dependent
PCA	• Good for large feature set; • Correlated data; • Unreliable if first three principal components do not account for at least 80 % of the variability
Heuristics	• Understand biological relevance or interpretation • Subject to bias

approach to the Student t -test in two-sample comparisons that does not require a parametric distribution. Accommodation for effects of multiple hypothesis testing should be routinely applied to high-dimensional data that contains a large number of analyte measurements on a small number of samples [small n –large p , or wide data set (Hastie et al. 2009)]. In using this approach, it is important to recognize that simultaneous testing of more than one hypothesis leads to inflated numbers of false positives. Common methods for reducing errors from multiple hypothesis testing are to employ family-wise error rate corrections such as the Bonferroni correction or the false discovery rate (FDR). Because the Bonferroni correction is extremely conservative and leads to the rejection of true positives, FDR procedures have been developed. FDR use a resampling method and have improved ability to reject false hypotheses compared to the Bonferroni correction (Hastie et al. 2009). The FDR analysis results in a quantile (q)-value, a number that indicates the proportion of false positives. For example, if a feature has a q -value of 0.05, it means that 5 % of features that show p -values at least as small as that are false positives. Feature reduction can also be based on a fold change value (values of 1.5 or 2 are often used) or statistical significance of mean difference of two groups (a q -value threshold of 1 or 5 % are often used). The fold change threshold will change the number of significant features when the analysis is followed by the use of a parametric test with multiple testing corrections.

Significance analysis of microarrays (SAM) is a widely used permutation-based approach to identifying differentially expressed features in high-dimensional data sets. One of the advantages of SAM is that it is able to assess statistical significance using FDR adjustment (Tusher et al. 2001). In the SAM analysis, each feature is assigned a score on the basis of its change relative to the standard deviation of its repeated measurements. SAM accounts for feature-specific fluctuations in signals and adjusts for increasing variation in features with low signal-to-noise ratios. Data are presented as a scatter plot of expected vs. observed relative differences, where significant deviations that exceed a threshold from expected relative differences are identified and considered “significant.”

An alternative approach for feature reduction is an empirical Bayesian technique (Efron et al. 2001). This approach uses noninformative priors and derives the posterior probability difference for each predictor. An advantage of empirical Bayes is that it does not rely on initial feature selection using t -tests or Wilcoxon tests to identify which features are differentially expressed. However, there are limited software implementations of this method and it requires significant computational time.

Principal component analysis (PCA) is a classical method for reducing the dimensionality of data to identify informative features. Here, features are transformed into orthogonal linear representation to identify uncorrelated components that most contribute to the variability in the experimental data. The features employed in the major principal components are then used in order of the magnitude of their corresponding eigenvalues.

Finally, intuition of the biological system (heuristics) can be used to select features based on their plausibility or prior knowledge about the system under investigation. Heuristics is also frequently used in combination with various

statistical methods (Table 17.1), where, for example, known biological relevance to disease pathophysiology would be used as a criterion to select some analytes over others.

17.3 Classification Approaches

17.3.1 *Logistic Regression*

Logistic Regression (LR) seeks to model the relationship between a set of continuous, categorical, or dichotomous variables and the probability of a dichotomous outcome as a “logit” function. Stated in another way, logistic regression is the method used for a binary, rather than a continuous outcome. The logistic regression model does not necessarily require the assumptions of some other regression models, like assuming that the variables are normally distributed in linear discriminant analysis. Maximum likelihood estimation is the method used to solve for the logistic regression equation. Recent techniques, penalized shrinkage and regularization estimation, lasso-type regularization logistic regression models have been developed to improve prediction accuracy in classification (Hastie et al. 2009).

17.3.2 *Variable Selection with the Lasso*

Regularized and shrinkage estimation methods, such as the least absolute shrinkage and selection operator (lasso) estimator (Tibshirani 1997), are methods of choice to address the problems of variable selection and multicollinearity. The lasso utilizes the L1 penalty to perform simultaneous continuous feature set shrinkage and automated variable selection. For binary response problems, the lasso estimator is estimated by penalized the negative log-likelihood with the L1-norm. The penalty term is chosen by a cross-validation technique for evaluating the out-of-sample negative log-likelihood. The elastic net (Zou and Hastie 2005) combines both L1 and L2 penalizing terms and possess a grouping effect, i.e., if there is a set of variables which have high pairwise correlations, the elastic net groups the correlated variables together. Zou proposed the adaptive lasso (Zou 2006), an approach that permits different weights for different parameters. The adaptive lasso also has been shown to have the oracle property (consistency in variable selection and asymptotic normality).

17.3.3 *Classification and Regression Trees*

Classification and Regression Trees (CART) is a nonparametric modeling algorithm for building decision trees (Steinberg and Colla 1997). Decision trees have a human-readable split at each decision point that represents a binary response of an informative feature. CART is highly useful for because it is one of the modeling techniques that does not require initial variable selection. Some other advantages of CART are that it can easily handle data sets which are complex in structure, it is extremely robust and not very effected by outliers, it can use a combination of both categorical and continuous data, missing data values do not pose an obstacle to CART classification.

The three main steps in CART modeling are creating a set of rules for splitting each node in a tree, deciding when a tree is fully grown, and assigning a classification to each terminal node of the tree. The basic algorithm for building the decision tree seeks some feature of the data that maximizes the difference between the classes contained in the parent node. Once CART has decided on an appropriate split resulting in two child nodes, the child nodes then become the new parent nodes, and the process is carried on down the branches of the tree. Cross-validation is frequently used to determine decision tree accuracy.

17.3.4 *Random Forests*

Random forests (RF) is an ensemble of CART models that offers several unique and extremely useful features which include built-in estimation of prediction accuracy, measures of feature importance, and a measure of similarity between sample inputs (Breiman 2001a, b). Decision trees are known for their ability to select the most informative descriptors among many and to ignore irrelevant ones. The RF algorithm is very efficient, especially when the number of descriptors is very large. This efficiency over traditional CART methods arises from two general areas: the first is that CART requires some amount of pruning of the tree to reach optimal prediction strength; RF, however, does not do any pruning, which reduces performance time. Second, RF uses only a small number of descriptors to test the splitting performance at each node instead of doing an exhaustive search as does CART.

RF thus builds many trees and determines the most likely splits based upon a comparison within the ensemble of trees. The RF modeling uses first a bootstrapped sample from the training data set. Then, for each bootstrapped sample, a classification tree is grown. Here, RF modifies the CART algorithm by randomly selecting from a subset of the descriptors, instead of choosing the best split among all samples. This procedure is repeated until a sufficiently large number of trees have been computed, usually 500 or more.

RF performs a bootstrapping cross-validation procedure in parallel with the training step of its algorithm. This method allows some of the data to be left out at each step and then used later to estimate the accuracy of the classifier after each instance (i.e., tree) has been completed.

17.3.5 *Multivariate Adaptive Regression Splines*

Multivariate adaptive regression splines (MARS) is a robust nonparametric modeling approach for feature reduction and model building (Friedman 1991). MARS is a nonparametric, multivariate regression method that can estimate complex nonlinear relationships by a series of spline functions of the predictor variables. Regression splines seek to find thresholds and breaks in relationships between variables and are suited for identifying changes in the behavior of individuals or processes over time. Some of the advantages of MARS are that it can model predictor variables of many data types, continuous or categorical, and it can tolerate large numbers of input predictor variables. As a nonparametric approach, MARS does not make any underlying assumptions about the distribution of the predictor variables of interest. This characteristic is extremely important in high-dimensional modeling because many of the cytokine and protein expression values are not normally distributed, as would be required for the application of classical modeling techniques. The basic concept behind spline models is to model using potentially discrete linear or nonlinear functions of any analyte over differing intervals. The resulting piecewise curve, referred to as a spline, is represented by basis functions.

Each MARS basis function takes one of the following three forms (1) a constant, where there is just one such term, the intercept; (2) a hinge function. A hinge function has the form $\max(0, xi - ci)$ or $\max(0, ci - xi)$. MARS automatically selects variables and values of those variables for knots of the hinge functions; or (3) a product of two or more hinge functions. These basis functions can model interaction between two or more variables.

The MARS algorithm has the ability to search through a large number of candidate predictor variables to determine those most relevant to the classification model. The specific variables to use and their exact parameters are identified by an intensive search procedure that is fast in comparison to other methods. The optimal functional form for the variables in the model is based on regression splines called basis functions. MARS uses a two-stage process for constructing the optimal classification model. The first half of the process involves creating an overly large model by adding basis functions that represent either single variable transformations or multivariate interaction terms. The model becomes more flexible and complex as additional basis functions are added. The process is complete when a user-specified number of basis functions have been added. In the second stage, MARS deletes basis functions in order of least contribution to the model until the optimum one is reached.

Depending on the modeling assumptions, method for model selection and acceptance of interaction terms, MARS models are sensitive to overfitting. Overfitting is a phenomenon where the model describes noise in the data set as well as the underlying relationships, and as a result, will not generalize to other data sets. As with the other modeling techniques, MARS uses cross-validation avoid overfitting. The optimal end result is a classification model based on single variables and interaction terms which will optimally determine class identity.

17.4 Methods for Model Evaluation

In practice, it is often customary for a supervised analysis to be conducted using several modeling approaches. The investigator will then examine the model performance using a variety of criteria, as well as look for convergence of informative features.

A widely accepted approach in model evaluation is to evaluate the area under the receiver operating characteristic (ROC) curve (Hanley and McNeil 1982). The area under the ROC curve (AUC) measures indicate the ability of a model to discriminate amongst the outcome groups. An AUC score of 0.5 indicates that a model has no discriminatory ability and a score of 1 indicates the two groups are perfectly discriminated.

Model performance of machine learning algorithms can also be examined using a confusion matrix. A confusion matrix shows the actual number of cases within each group, as well as the number of cases predicted by the model to lie within each group. This analysis allows the modeler to determine model performance for prediction of each class. In addition, the overall percent correct prediction is also evaluated.

17.4.1 Classifier Performance in Asthma Phenotypes

Recently, we have evaluated the performance of various predictive models using multiplex BAL cytokine measurements in bronchoalveolar lavage (BAL) to predict intermediary phenotypes of atopic asthma (Brasier et al. 2010). In our preliminary analysis of 41 severe, and 43 nonsevere asthmatics from the Severe Asthma Research Program, we observed that cytokine concentrations could be used as molecular fingerprints of asthmatic subtypes using unsupervised clustering. Unsupervised classification using hierarchical clustering suggested that subjects could be segregated into groups that showed statistically significant differences in type of BAL cellularity, including eosinophils and lymphocytes, as well as measures of static and dynamic pulmonary lung function (Brasier et al. 2008). Our aim was, therefore, to compare classification methods to identify the best methods that would allow predictive models of these distinct clinical phenotypes.

We conducted this systematic comparison of distinct statistical learning approaches because currently it is not possible presently to select a priori the machine learning tool that performs best for any given data set. All statistical learning approaches are sensitive to the underlying data structure. To evaluate the performance of these distinct classifiers, we performed supervised classification using BAL cytokines as input, and clinical phenotype of methacholine sensitivity (“hyperresponders”) as an outcome (as well as others).

We observed that an LR model constructed using the AIC information criterion, a method equivalent to cross-validation (Stone 1977), produced a good fit with the

Table 17.2 Model performance comparisons

Phenotype	Classifiers	Accuracy (%)	AUC	Sensitivity (%)	Specificity (%)
High eosinophils	LR	85	0.84	53	94
	MARS	85	0.89	46	93
	CART	67	0.48	30	74
	RF	49	0.40	46	49
High neutrophils	LR	97	0.76	60	98
	MARS	93	0.77	20	98
	CART	83	0.51	20	88
	RF	66	0.47	40	68
Bronchodilators	LR	76	0.71	23	85
	MARS	88	0.85	23	100
	CART	75	0.69	53	81
	RF	61	0.65	54	64
Hyper-responders	LR	85	0.85	53	94
	MARS	90	0.88	73	94
	CART	79	0.68	53	86
	RF	70	0.77	73	70

Shown is the model performance comparisons for the machine learning approaches for accuracy, AUC, sensitivity and specificity. Abbreviations: *AUC* area under the ROC curve, *LR* logistic regression, *MARS* multivariate adaptive regression splines, *CART* classification and regression trees, *RF* random forests. Reprinted with permission from (Brasier et al. 2010)

outcome. For the methacholine “hyperresponders” IL-1Ra, Eotaxin, IL-4, and IL-2R were selected (Table 17.2). MARS produced a slightly more accurate model than did LR with 90 % accuracy (Table 17.2, Fig. 17.3). Both of these models outperformed the CART and RF. The one cytokine with the greatest χ^2 score statistic in the LR and the greatest variable importance in the MARS was IL-1Ra. IL-1Ra was also highly statistically significant in the groupwise comparison of cytokines in the hyperresponders vs. nonhyperresponder groups.

The ROC curves for LR MARS, CART, and RF prediction of methacholine hypersensitivity are shown in Fig. 17.3. Note that MARS and LR show various degrees of leftward shift from the line of no discrimination (AUC of 0.88 and 0.85, respectively), indicating that these models perform better than random guessing. Finally, LR resulted in a model with 85 % overall accuracy, whereas MARS produced a model with 90 % overall accuracy.

17.5 Post Hoc Model and Feature Evaluation

We have found that post hoc model diagnostics is an important procedure to assess adequacy of machine learning predictors. In this regard, several techniques are used to examine model performance. In LR, often the interest is to assess the linear or nonlinear association of binary responses on its covariates (features). When we use an LR model, we assume that the logit of the outcome is a linear combination of

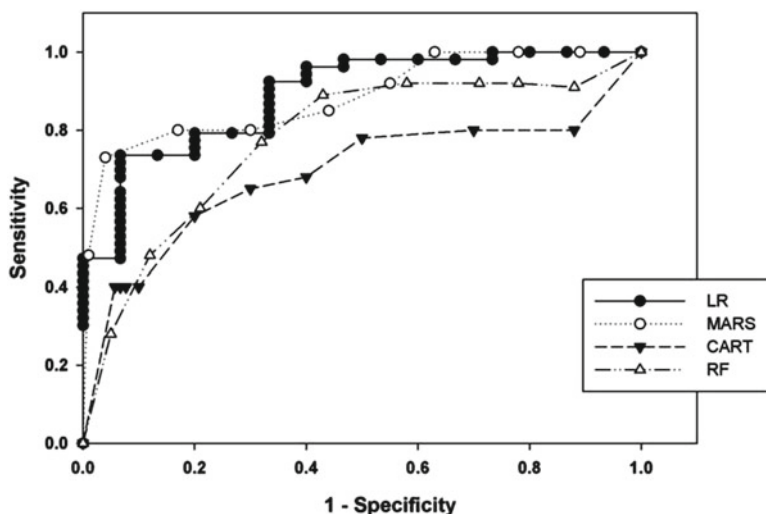


Fig. 17.3 Comparison of model performance in predicting asthma phenotype. Receiver operating characteristic (ROC) curve of methacholine hyperresponder models. The corresponding AUCs for each curve are presented. Reprinted with permission from (Brasier et al. 2010)

predictors. If the model assumptions are not satisfied, the model may contain biased coefficient estimators and very large standard errors for regression coefficients, and consequently not generalize well to other cases.

17.5.1 Generalized Additive Models

In practice, we examine whether our model has all relevant predictors and if the linear association with outcome is appropriate. One method for assessing this relationship is to use Generalized Additive Models (GAMs). GAMs are regression models which can evaluate complex nonlinear predictors on the outcomes in a generalized linear model setting (Hastie and Tibshirani 1995). A GAM plot is represented by the value of the covariate (feature) on the X-axis, and the partial residual represents the difference between the estimated and observed value on the Y-axis (Fig. 17.4). The relationship of the covariate with the outcome is indicated by a smoothed piecewise polynomial relationship contained within 95 % confidence limits.

We evaluate the partial residual plot as a diagnostic graphical tool for identifying nonlinear relationships between the response variable and covariates for generalized additive models. In the case of the IL-1Ra covariate, the piecewise polynomial relationship is not a linear relationship, and therefore is best estimated using a non-parametric modeling approach (Fig. 17.4). For linear regression models, we also use

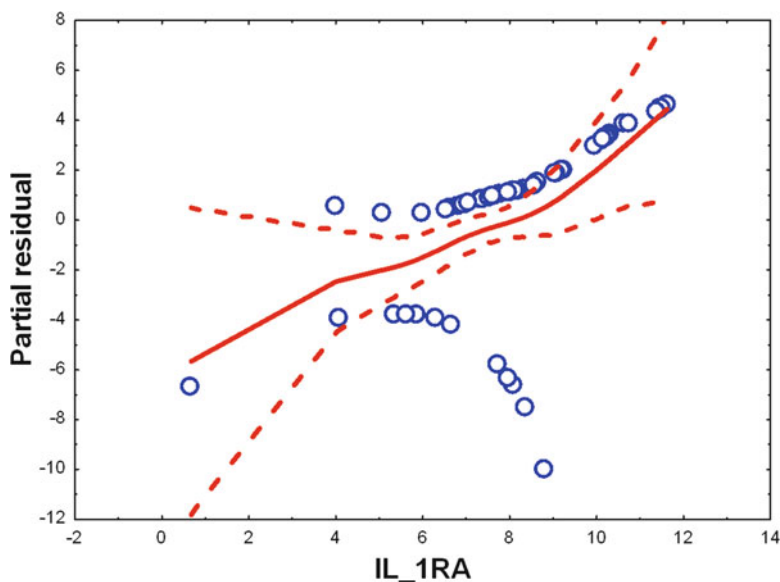


Fig. 17.4 Generalized Additive Model (GAM). In a GAM graph, the Y-axis is the partial residual and the X-axis is the value of the covariate (feature). *Dashed lines* are 95 % confidence intervals. *Open circles* are data points. Shown are the partial regressions for IL-1Ra in prediction of PC₂₀ methacholine hyperresponsiveness, demonstrating a nonparametric relationship with outcome. In this case, the partial residual plot deviates from a *straight line*

the projection (hat) matrix and Cook's distance to evaluate the effect of outliers and influential points on the model (Spratt et al. 2012).

17.6 Strengths and Limitations of Supervised Approaches to Profiling

Supervised approaches can be extremely powerful to predict specific populations from unknown cases based on molecular profiling measurements. Informative analytes may provide information on potential mechanisms of dysregulation of disease; informative analytes are identified from statistical (between-group comparisons) and model analysis, where systematic relationships between analyte values a phenotype can be revealed. Additionally, the model may also suggest potential interfeature interactions that may further extend understanding of the pathophysiology of disease.

Supervised analysis critically depends on a method for analyte measurement that is quantitative and reproducible across the population. Moreover, an objective, reproducible phenotypic characterization of subjects in the data from which the model is developed is critical for success. It is obvious that if there is inconsistency

in outcome classification, the resulting model will not generalize (have wide applicability). For this reason, a parallel analysis of the data using unsupervised classification methods such as those described in the following chapter is advisable.

17.7 Conclusions and Future Directions

Supervised classification is critically dependent on the characteristics of the analyte distribution and the number of analytes being considered. These factors have an important contribution to the performance of different classifiers. For example, LR is a parametric approach that identifies main effects of candidate cytokines and assumes a global linear relationship between the independent (cytokine) and the dependent (asthma phenotype) variables. Although LR performs well if the relationship between the dependent and independent variables is well described, the performance of LR is reduced with high-dimensional data or when there are multiple interactions between independent variables. By contrast, MARS is a nonparametric, piecewise linear approach that can establish relationships over smaller intervals of independent variables, as well as detect interactions between independent variables.

Our experience so far shows that LR and MARS result in quite similar accuracy for relating proteomic profiles to clinical phenotypes in asthma, with both clearly outperforming CART and RF classifiers (Brasier et al. 2010). Interestingly to us, for many of the best performing models developed in LR and MARS, the cytokine features important for the model are different. The reasons for these discrepancies will require further investigation, but may be a consequence of the underlying data structure affecting feature selection.

References

- ad-hoc writing committee of the Assembly on Allergy, Immunology and Inflammation (2000) In: Proceedings of the ATS Workshop on Refractory Asthma. Current understanding, recommendations, and unanswered questions. *Am J Respir Crit Care Med* 162:2341–51
- Bhavnani S, Victor S, Calhoun WJ et al (2011) How cytokines co-occur across asthma patients: from Bipartite Network Analysis to a molecular-based classification. *J Biomed Inform* 44:24–30
- Brasier AR, Victor S, Boetticher G et al (2008) Molecular phenotyping of severe asthma using pattern recognition of bronchoalveolar lavage-derived cytokines. *J Allergy Clin Immunol* 121:30–37
- Brasier AR, Victor S, Ju H et al (2010) Predicting intermediate phenotypes in asthma using bronchoalveolar lavage-derived cytokines. *Clin Transl Sci* 13:147–57
- Brasier AR, Garcia J, Wiktorowicz JE et al (2011) A candidate biomarker panel for predicting dengue hemorrhagic fever using discovery proteomics and nonparametric modeling. *Clin Transl Sci* 5:8–20
- Breiman L (2001a) Random forests. *Machine Learning* 45:525–31

- Breiman L (2001b) Random forests, random features. University of California, Berkeley, CA
- Efron B, Tibshirani R, Storey JD et al (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96:1151–60
- Friedman JH (1991) Multivariate adaptive regression splines. *Ann Stat* 19:1–67
- Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic curve. *Radiology* 143:29–36
- Hastie T, Tibshirani R (1995) Generalized additive models for medical research. *Stat Methods Med Res* 4:187–96
- Hastie T, Tibshirani R, Friedman A (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
- Spratt H, Ju H, Brasier AR (2012) A structured approach to predictive modeling of a two-class problem using multidimensional data sets. *Methods* 61:73–85
- Steinberg D, Colla P (1997) CART classification and regression trees. Salford Systems, San Diego, CA
- Stone M (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J R Stat Soc, Series B (Methodological)* 39:44–47
- Tibshirani R (1997) The lasso method for variable selection in the Cox model. *Stat Med* 16:385–95
- Troyanskaya O, Cantor M, Sherlock G et al (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17:520–5
- Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98:5116–21
- Wenzel SE, Busse WW (2007) Severe asthma: lessons from the Severe Asthma Research Program. *J Allergy Clin Immunol* 119:14–21
- Zou H (2006) The adaptive lasso and its oracle properties. *J Am Stat Assoc* 101:1418–29
- Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J R Stat Soc B* 67:301–20