International Journal of Biology; Vol. 7, No. 1; 2015
ISSN 1916-9671 E-ISSN 1916-968X
Published by Canadian Center of Science and Education

Optimal Algorithm for Metabolomics Classification and Feature Selection varies by Dataset

Charles E. Determan Jr.[1]

[1] University of Minnesota, Department of Surgery, Division of Critical Care and Acute Care Surgery, USA

Correspondence: Charles E. Determan Jr., University of Minnesota, Department of Surgery, Division of Critical Care and Acute Care Surgery, USA. Tel: 612-624-3940. E-mail: deter088@umn.edu

Received: November 13, 2014 Accepted: December 6, 2014 Online Published: December 12, 2014

doi:10.5539/ijb.v7n1p100 URL: http://dx.doi.org/10.5539/ijb.v7n1p100

Abstract

Metabolomics, the systematic identification and quantification of all metabolites in a biological system, is increasingly applied towards identification of biomarkers for disease diagnosis, prognosis and risk prediction. Applications of metabolomics extend across the health spectrum including Alzheimer's, cancer, diabetes, and trauma. Despite the continued interest in metabolomics there are numerous techniques for analyzing metabolomics datasets with the intent to classify group membership (e.g. Control or Treated). These include Partial Least Squares Discriminant Analysis, Support Vector Machines, Random Forest, Regularized Generalized Linear Models, and Prediction Analysis for Microarrays. Each classification algorithm is dependent upon different assumptions and can potentially lead to alternate conclusions. This project seeks to conduct an in depth comparison of algorithm performance on both simulated and real datasets to determine which algorithms perform best given alternate dataset structures. Three simulated datasets were generated to validate algorithm performance and mimic 'real' metabolomics data: (Han et al., 2011) independent null dataset (no correlation, no discriminatory variables), (Davis, Schiller, Eurich, & Sawyer, 2012) correlated null (no discriminating variables), (Guan et al., 2009) correlated discriminatory. This comparison is also applied to 3 open-access datasets including two Nuclear Magnetic Resonance (NMR) and one Mass Spectrometry (MS) dataset. Performance was evaluated based on the Robustness-Performance-Trade-off (RPT) incorporating a balance between model classification accuracy and feature selection stability. We also provide a free, open-source R Bioconductor package (OmicsMarkeR) that conducts the analyses herein. The proposed work provides an important advancement in metabolomics analysis and helps alleviate the confusion of potentially paradoxical analyses thereby leading to improved exploration of disease states and identification of clinically important biomarkers.

Keywords: feature selection, machine learning, metabolomics, multivariate analysis

1. Introduction

Metabolomics, similar to the other two common ‘omics’ approaches (i.e. transcriptomics and proteomics), is defined as the systematic identification and quantification of all metabolites in a biological system. Such data is commonly acquired via Nuclear Magnetic Resonance spectroscopy (NMR) or Mass Spectrometry (MS). Metabolomics has been increasingly applied towards identification of biomarkers for disease diagnosis, prognosis and risk prediction. Applications extend across the health spectrum including Alzheimer’s (Han et al., 2011), cancer (Davis et al., 2012; Guan et al., 2009; Nishiumi et al., 2010), diabetes (Bain et al., 2009), and trauma (Determan et al. 2014).

Following the initial pre-processing (e.g. peak picking, deconvolution, integration, etc.), a metabolomics dataset must ultimately be analyzed to typically classify two or more classes/conditions in addition to identifying the most important metabolites for the discrimination (e.g. biomarker studies). The availability and use of multivariate approaches is rapidly becoming critical with decreased cost and increased access to high-throughput metabolomics platforms including NMR and MS resulting in “large p, small n” problems (i.e. many more variables than samples). The common univariate tests become grossly underpowered to assess every feature and require a secondary model if classification is desired. The restrictive assumptions of univariate tests (e.g. normality) are typically avoided with more sophisticated multivariate, machine learning algorithms.

But despite the continued interest in metabolomics there is no standard statistical approach resulting in numerous techniques applied inconsistently across experiments. Common methods include Partial Least Squares Discriminant Analysis (PLSDA), Lasso and Elastic-Net Regularized Generalized Linear Models (GLMNET),

Support Vector Machines (SVM), Random Forests (RF), Gradient Boosting Machines (GBM), and Prediction Analysis for Microarrays (PAM). Although each method is an effective classification algorithm, previous comparisons of algorithms in gene expression experiments report that different algorithms provide improved accuracy for different datasets (JW Lee, JB Lee, Park, Song, 2005).

Limited algorithm comparisons in metabolomics studies (i.e. comparing two or three methods) often measure performance solely on accuracy and neglect the stability of features selected (i.e. how consistently the same features are identified). Even though an analysis reports high accuracy, repeating the biomarker discovery procedures can result in different feature subsets even within the same datasets (Ein-Dor, Zuk, & Domany, 2006; Michiels, Koscielny, & Hill, 2005; Zucknick, Richardson, & Stronach, 2008). Feature selection can also introduce optimistic bias into statistical inference because the signal-noise ratio of the data set is increased by the feature selection procedure. An extreme example is that all features are irrelevant to a response, but the selected features will still appear fairly predictive to the response which is, however, completely by chance (Ambroise & McLachlan, 2002). Therefore, to determine which algorithms perform optimally, both feature selection stability and overall classification accuracy must be evaluated together.

In this work, we evaluate the six aforementioned classification algorithms performance and stability on both *in silico* and experimentally acquired datasets. Metabolomics datasets are inherently multivariate with both independent and multicollinear variables in addition to possessing a mix of Gaussian and non-Gaussian distributions. To evaluate algorithm performance on such datasets it is necessary to generate standardized datasets that mimic true metabolomics data and possess known results as a benchmark. Furthermore, an application to previously acquired datasets from multiple platforms with previous results is provided herein.

The goal of this experiment has been to demonstrate variability in algorithm performance across multiple datasets of different characteristics (e.g. sample size, number variables, etc.) and various methodologies (e.g. ensemble). To our knowledge, this is also the first integration of feature stability with model performance to metabolomics datasets. Furthermore, the ability to conduct multiple analyses utilizing accepted methods resulting in similar conclusions adds further support to any conclusions in a field where reproducibility is often a major concern.

2. Methods

2.1 Datasets

Despite the growth of metabolomics, there is no commonly accepted gold standard dataset for algorithm evaluation. This necessitates the production of simulated datasets that accurately mimic typical metabolomics datasets from both NMR and MS. This requires the perturbation of normality in multiple variables and inclusion of multicollinearity as is typical of metabolite distributions and relationships. It is also necessary to determine the performance of algorithms when examining the null condition wherein there is no difference between conditions. Therefore, three simulated datasets were generated (null independent, null correlated, and correlated discriminatory) to analyze algorithm performance that may also be used by others for further performance evaluations with respect to mimicking biological datasets as opposed to datasets designed to evaluate specific algorithm performances. This was repeated twice, once at the NMR scale and once at the MS scale, as the number of resolvable features between the two techniques can be an order of magnitude; NMR typically can resolve 50-75 metabolites whereas MS can resolve 100's to 1000's of metabolites (Wishart 2010). Although exceedingly large datasets are possible with *in silico* data the sample sizes herein were selected to more accurately reflect empirical datasets given the limits from costs and/or sample availability. Low and high sample sizes were set at 25 and 50 samples per group respectively. Although these are still high for many applications, this allows the use of leave k-fold out cross-validation. For very small datasets one may use leave-one-out cross-validation or the option to forgo validation, if appropriate parameters are known.

2.1.1 Simulated Datasets

Each simulated dataset consists of N samples and p variables representing the individual samples and metabolites of an experiment. In addition, for the purposes of classification, groups are assigned to represent classes within a dataset (e.g. Control, Cancer).

1) ***Null***: A null dataset was generated to provide an absolute base where no classification should be found. Simulated data were generated with the *create.random.matrix* function following previously established methods (Wongravee et al., 2009) with the following noted modifications. The initial datasets was generated with random numbers from a normal distribution, each dataset consisting of N samples (*nsamp,* 25/50 per group, low and high) and p variables (*nvar*, NMR = 50, MS = 1000). The normal distribution was also perturbed by adding a second matrix containing uniform random numbers between -0.2 and 0.2 (*perturb* = 0.2). The samples were then assigned to groups of equal numbers.

2) ***Null Correlated***: To mimic 'real' metabolomics datasets, correlations were induced for the null correlated datasets with the *create.corr.matrix* function. This provided a dataset that could be used to evaluate the effect of correlations on classification and feature selection performance. Blocks of variables of size b (*min.block.size* = 2, *max.block.size* = 5) were randomly assigned and had values replaced with correlated values derived from the first variable of the specific block. We elected to incorporate blocks of size 1 (*min.block.size = 1)* as the smaller metabolomics datasets in NMR more likely possess independent variables. This induced correlation was also perturbed to more accurately represent real data. Derived correlation coefficients were compared to real metabolomics datasets to validate the method (data not shown).

3) ***Discriminatory***: To facilitate discriminatory analysis, *D* variables (NMR = 10, MS = 20) were randomly induced to be discriminatory with the *create.discr.matrix* function. A discriminatory index (*l*) was selected and for each variable D whereby a random number between –l and l was added to one group and subtracted from the other.

2.1.2 Real Datasets

NMR datasets included a binary (i.e. 2 groups) urine dataset analyzing cachexia (Eisner et al., 2011) and multi-class (i.e. 4 groups) rumen fluid dataset investigating the impact of altered diets of cows (Ametaj et al., 2010). The MS explores potential biomarkers of Hepatocellular Carcinoma in serum samples (Xiao et al. 2012) and was accessed from the open source metabolomics data repository Metabolights (Haug et al., 2013) accession number MTBLS19.

2.2 Classifier and Feature Selection Algorithms

Below we describe six classification algorithms utilized in the metabolomics literature, with built-in or added feature selection capability. We briefly describe how we used each for classification and feature selection. With respect to feature selection, investigators also may or may not have an approximate idea of how many features (i.e. metabolites) they expect to be discriminating. As such two methods, defined herein as 'subset' and 'model derived', are applied whereby a specific number of variables are specified *a priori* or the specific model is allowed to return an internally determined number of variables respectively.

2.2.1 Partial Least Squares Discriminant Analysis (PLSDA)

Partial Least Squares Discriminant Analysis is a dimension reduction technique analogous to principal component analysis. The algorithm focuses on maximizing the variance of the dependent variables explained by the independent variables (Wold, 1975). It is robust to multicollinearity, missing data, and skewed distributions (Cassel, Hackl, & Westlund, 1999). These models were tuned on the number of components to be retained. Feature selection was accomplished by ranking features on the sum of squares of their loading weights, a technique known as variable importance of projection (VIP)(Wold, Johansson, & Cocchi, 1993). Model derived features were selected as those with a VIP score >= 1.0. PLSDA is commonly used in metabolomics investigations including multiple forms of cancer, cardiac ischemia, parkinson's disease and asthma (Nishiumi et al., 2010; Bodi et al., 2012; Bogdanov et al., 2008; Carraro et al., 2007; Chen et al., 2011; Duarte et al., 2010; Qiu et al., 2010). This technique has also been previously implemented in our lab investigating hemorrhagic shock (Lexcen, Lusczek, Witowski, Mulier, & Beilman, 2012). It is readily available in the R package DiscriMiner (Sanchez, 2012).

2.2.2 Regularized General Linear Model (GLMNET)

Generalized linear models are a more flexible form of linear regression that allows the response variables to have non-parametric distributions. To avoid the risk of overfitting data in multiple linear regression a regularization method can be applied. We have selected to use the elastic-net penalty, which is a compromise between the LASSO and Ridge shrinkage methods and has been shown to outperform LASSO (Zou & Hastie, 2005). In brief, elastic net is a weighted average of the lasso and ridge solutions. The LASSO penalty encourages non-discriminative features to have a coefficient that is exactly zero, thus it performs feature selection; and the Ridge penalty is used to overcome multicollinearity. Having both of these penalties integrated facilitates analysis of data with collinearity and internal feature selection. These models were tuned on the lambda penalty and the elastic-net parameter alpha. Important features are identified as those with non-zero coefficients. These coefficients were also ranked for subset feature selection. Although it is less common than other techniques it has been used in recent metabolomics studies (Rohart et al., 2012, Wahl et al., 2013). GLMNET is readily available in the R package glmnet (Friedman, Hastie, & Tibshirani, 2010).

2.2.3 Random Forest (RF)

Random forest is a machine learning algorithm that uses a combination of tree predictors (i.e. forest) such that each tree is constructed on a random sample of the observations thereby independently ensuring that the distributions

are the same for all the trees in the forest. Each tree in the forest provides a 'vote' for the best class. This is constructed on a training subset of the data and tested against the remaining test data known as the 'out-of-bag' (OOB) data. The scaled sum of the votes derived from the trained trees determines the final "score" (Breiman, 2001). Random Forests were tuned on the number of trees and the number of variables randomly sampled at each split. The feature selection is determined by permuting variables in the OOB and observing increases in error. A variable score indicates greater importance to the model. These scores were ranked for subset feature selection and those exceeding a score of 1.0 for model derived results. It is robust to noise and outliers and computationally faster than bagging or boosting. Prior studies have reported error rates comparable if not better than other predictors such as logistic regression, linear discriminant analysis, quadratic discriminant analysis (QDA), K-nearest neighbors (KNN), Support Vector Machines (SVM), classification and regression trees (CART) and Naïve Bayes (Breiman, 2001; Folleco, Khoshgoftaar, Van Hulse, & Bullard, 2008; Svetnik et al., 2003). However, consistency of selected feature rankings has been shown to be problematic for high dimensional problems (Verikas, Gelzinis, & Bacauskiene, 2011). It has been used in several metabolomics studies (Hische et al., 2012; Houtkooper et al., 2011; Patterson et al., 2011) and is readily available in the R package randomForest (Liaw & Wiener, 2002).

2.2.4 Gradient Boosting Machine (GBM)

Gradient boosting is another machine learning technique applied most commonly to decision trees that produces robust and interpretable procedures for both regression and classification (Freidman, 2001). Unlike the bagging approach (e.g. random forest), where trees are constructed independently and are thus assumed to make prediction errors independently, gradient boosting trees are constructed sequentially and each new tree is fitted to compensate for errors committed by previous trees. As with random forest, feature selection is determined by permuting variables in the OOB and observing increases in error resulting in a subsequent variable score. Gradient Boosting models were tuned on the number of trees, the interaction depth, and learning rate (i.e. shrinkage, step-size reduction). Feature selection was accomplished via ranking scores and those exceeding 1.0 for model derived results. Boosting has become known as one of the most powerful learning ideas in the last twenty years (Hastie, Tibshirani, & Friedman, 2009) but curiously has never been applied to metabolomics settings. To our knowledge, this is the first application of boosting to analyze metabolomics data. Freidman's gradient boosting machine algorithm is available in the R package gbm (Ridgeway, 2013).

2.2.5 Support Vector Machines (SVM)

Support vector machine is based on the structural risk minimization principle from statistical learning theory (Vapnik, 1998). It can be applied to classification problems with the idea of structural risk minimization to find a hypothesis that has the lowest probability of error. It has been shown to be robust to high dimensionality, noisy data, and outliers. Prior comparisons with PLSDA report improved overall accuracy with less features (Mahadevan, Shah, Marrie, Slupsky, 2008) but feature selection consistency is unknown. This classification algorithm is readily available within the R e1071 package (Meyer, Dimiriadou, Hornik, Weingessel, & Leisch, 2012) where we elected to use the common linear kernel and tuned on the cost parameter. Feature selection was accomplished via recursive feature elimination (RFE) as detailed by Guyon as to our knowledge there is no metric specifically designed for SVM (Guyon, Weston, Barnhill, & Vapnik, 2002).

2.2.6 Prediction Analysis for Microarrays (PAM)

Prediction Analysis for Microarrays is a modified nearest centroid classification method to include centroid shrinkage and contains an embedded feature selection step (Tibshirani, Hastie, Narasimhan, & Chu, 2002). In brief, the average value for each variable is divided by the within-class standard deviation to provide class centroids. These class centroids are then shrunk towards zero by a defined threshold to reduce noise and facilitate variable selection. Then a new sample profile is compared to each of the class centroids. The class whose centroid is closest is the predicted class. These models were tuned on the 'threshold' parameter for the centroid shrinkage. The internal feature selection is accomplished by identifying features with non-zero coefficients which are subsequently ranked for subset selection. This technique has not been used widely in metabolomics investigations; however, as the name implies it has been successfully been used for classification in gene expression experiments (Ray et al., 2007; Sadanandam et al., 2013). This algorithm is readily available in the R package pamr**.**

2.3 Evaluate Stability of Feature Selection Techniques

The high-dimensional datasets of metabolomics often necessitate feature selection techniques to reduce dimensionality to the most important features to facilitate subsequent analysis. Although many approaches rely exclusively on classification accuracy of feature subsets to facilitate biomarker selection this is problematic where several different feature subsets may yield equally optimal results (Saeys, Inza, & Larrañaga, 2007). It is therefore necessary to evaluate the robustness of feature selection techniques applied to metabolomics data to facilitate

improved reproducibility and confidence in identified biomarkers. In brief, algorithm robustness were evaluated via instance (bootstrapped data subsets) and function (alternate algorithms) perturbation and evaluated by the Jaccard's Index (Real & Vargas, 1996). Other options include the Dice-Sorensen's Index (Dice, 1945; Sorensen, 1948), Ochiai's Index (Ochiai, 1957), Percent of Overlapping Features (Shi et al., 2005), Kuncheva's Index (Kuncheva, 2007), Spearman Rank Correlation, and Canberra Distance (Jurman et al., 2008). A comparison of these metrics is beyond the scope of this article.

Two common approaches are applied within instance perturbation to evaluate the robustness of feature selection techniques: perturbation at the instance level (i.e. removing or adding samples) or at the feature level (i.e. adding noise). We have selected to evaluate robustness of feature selection algorithms by estimating stability following perturbation at the instance level as the number of samples is the most likely problem facing metabolomics investigations.

2.4 Single Feature Selection Stability and Classification Performance

For each feature selection algorithm we estimated stability via instance perturbation with the *fs.stability* function. Instance perturbation was conducted via bootstrapping, without replacement, 90% ($p = 0.9$) of the data 10 times ($k = 10$) thereby creating a training and testing dataset for each iteration. For each training dataset all 6 feature selection algorithms were run simultaneously to provide an ordered list of selected feature rankings. Each iteration tunes the full model (*optimize* = TRUE) with a tuning grid of a specified resolution determining how fine the tuning parameters are optimized (*resolution = 5*). To avoid overfitting, 10-fold cross-validation was utilized (*k.fold* = 10) wherein $1/10^{th}$ of the each training data subset is randomly removed and the model evaluated on this test fold. Results were averaged over all 10 folds to provide the confusion matrix for subsequent performance metrics evaluated on the typical prediction accuracy (*metric = "Accuracy"*) which is the proportion of true results (true positive and true negative) in the sample population. The optimized models were then used to extract feature subsets of a user specified length (*f,* NMR = 10, MS = 20) or optionally by the model defined cutoff (*model.features = FALSE*). These feature subsets are compared via the Jaccard index (Equation 1). The overall stability is defined as the average over all pairwise similarity comparisons between each of the feature selection runs. The final model is refit using the extracted feature subset from the individual method and re-optimized using the initial tuning grid generated. Lastly, this trimmed model is used to predict the initial testing dataset generated at the start of the iteration. Accuracy was extracted with the *performance.metrics* function to compare each algorithms performance. This is repeated for the additional 9 times utilizing the previously optimized parameters for the full model generation (*optimize.resample = FALSE*).

Equation 1 Jaccard Index

$$\frac{|x \cap y|}{|x \cup y|}$$

2.5 Balance Stability and Classification Performance

In every scientific investigation which sample size is a limitation (i.e. most studies), researchers must balance power and sensitivity. The same principle is applied to balancing feature selection robustness and classifier performance as both are integral to confident biomarker identification. We utilized the robustness-performance trade-off (RPT) to balance feature selection stability and classification performance (Saeys, Abeel, & de Peer, 2008). In brief, the user can specify the parameter β to control the relative importance of feature stability versus classification. The default value of β = 1 which represents equal importance between stability and classification.

Equation 2 Robustness-Performance Trade-off (RPT)

$$\frac{\beta^2 * stability * performance}{\beta^2 * stability + performance}$$

2.6 Ensemble Feature Selection Stability and Classification Performance

Ensemble feature selection has been shown to improve stability in gene expression studies (Abeel, Helleputte, de Peer, Dupont, & Saeys, 2010; Davis et al. 2006). Therefore it is important to incorporate such analysis into metabolomics analysis for each algorithm. In essence, ensemble approaches use different data subsets and aggregating the results following feature selection. As described in the 'Single Feature Selection Stability and Classification Performance' section, stability was evaluated via instance perturbation with the *fs.ensembl.stability* function. For each subsample a second level of instance perturbation generated 40 (*bags* = 40) further datasets via bootstrap aggregation (aka. Bagging) (Breiman, 1996). For each bag a separate feature ranking was performed. The resulting list of selected feature rankings from each bag were combined via linear aggregation

(*aggregation.metrics* = "CLA") whereby the sum of the individual feature ranks within each bag contribute linearly to the overall final rank. The Jaccard Index was used to measure similarity and averaged over all pairwise comparisons for an overall measure of stability. Function perturbation, the use of multiple feature selection algorithms, was also conducted by the list of methods chosen within *fs.stability* and *fs.ensembl.stability*. Lastly, in contrast to non-ensemble approaches, there are no model derived runs because all features must be ranked for aggregation methods.

3. Results

3.1 Simulated Data

3.1.1 Binary Classification - Low Samples

3.1.1.1 Random and correlated datasets (N=50, p=50/1000 (NMR/MS))

Non-ensemble analysis of random and correlated dataset analyses provided generally expected results. Accuracy exceeded 0.700 for SVM, RF, GLMNET and PAM but stability remained low (>=0.47). While RF achieved the highest accuracy (often in excess of 0.900), it had the lowest stability warranting caution in interpreting results. Notably, accuracy was generally higher with the MS-scale dataset where accuracy exceeded 0.900 for the same four algorithms. However the stability of the feature subsets was also lower. Ensemble analysis of NMR-scale and MS-scale random and correlated datasets again reflected previous analysis with high accuracy levels for SVM, RF, GLMNET and PAM but low stability (Supplementary File - S1).

3.1.1.2 Discriminatory datasets

Analysis of the NMR-scale discriminatory dataset determined PAM as the optimal model with the highest RPT and PPI% (percent of discriminating features positively identified, Equation 3) (Table 1). GLMNET performed similarly with better accuracy but lower stability and PPI%. The model derived analysis also provided PAM with the highest RPT, however, the low sample size resulted in a conservative trimming of features resulting in many remaining in the model and decreasing the PPI%. The highest PPI% was reported by SVM which also had the very high accuracy; however it is also noted that there is no internal trimming metric for SVM and only the top 10% of features are returned making this a more restricted subset model. This suggests that experiments with lower sample sizes may need to restrict to only a few of the most discriminate features. The MS-scale datasets also reported PAM with the highest RPT and stability but the highest PPI% was reported by PLSDA. Such a situation supports the value of using multiple algorithms to determine consistent results.

Equation 3 Percent Features Positively Identified (PPI%)

$$\frac{|x \cap y|}{|y|}$$

Ensemble analysis of NMR-scale discriminatory dataset reported SVM with the highest RPT with PLSDA, GLMNET and PAM performing similarly. The MS-scale analysis was less conclusive with mixed performance among SVM, GLMNET and PAM (Supplementary File - S2).

3.1.2 Binary Classification – High Samples

3.1.2.1 Random and correlated datasets (N=100, p=50/1000 (NMR/MS))

Increased sample size had little effect on accuracy and stability of NMR-scale random and correlated datasets but worsened models of MS-scale data (Supplementary File – S1). This should be expected as having more data should increase the likelihood of calculating the 'true' condition. As expected, there was also little effect of high samples on the ensemble analysis of the both random and correlated datasets of MS-scale and NMR-scale sizes.

3.1.2.2 Discriminatory Datasets

Overall performance greatly improved in the NMR-scale discriminatory dataset with accuracy exceeding 0.9 for three algorithms, stability exceeding 0.8, and PPI% to 80% (Table 1). Performance in the MS-scale dataset also improved with stability and PPI% increasing most notably for GBM and RF respectively (Supplementary File – S2). Both datasets provide a circumstance whereby no algorithm performs best and multiple methods are beneficial.

The results of the ensemble NMR-scale discriminatory dataset improved RF and SVM but worsened GLMNET and PAM (Table 1). Likewise, the ensemble MS-scale discriminatory dataset also improved performance with increased samples most notably in RF stability and PPI% (Supplementary File – S2). However, the best performing algorithms were PLSDA, PAM and GLMNET. Although ensemble aggregation tends to improve model performance it may also have no effect or even worsen model performance.

Table 1. Results from NMR-scale Binary Classification Simulations. RPT – Robustness-Performance Trade-off, PPI% - Percent features positively identified. *SVM doesn't have internal cutoff so defaults to top 10%.

			Method	RPT	Accuracy	Stability	# Features	PPI%
Binary	High Sample	Subset	PLSDA	0.923	0.982	0.87	10	80.0%
			GBM	0.748	0.710	0.79	10	76.0%
			SVM	0.742	1.000	0.59	10	61.0%
			RF	0.529	1.000	0.36	10	59.0%
			GLMNET	0.942	1.000	0.89	10	80.0%
			PAM	0.936	1.000	0.88	10	80.0%
		Model Derived	PLSDA	0.931	0.965	0.90	10	80.0%
			GBM	0.552	0.480	0.65	14	55.0%
			SVM	0.770	0.990	0.63	5*	68.0%
			RF	0.549	0.990	0.38	17	38.8%
			GLMNET	0.927	0.980	0.88	8	92.5%
			PAM	0.860	0.990	0.76	44	22.5%
		Ensemble	PLSDA	0.928	0.982	0.88	10	80.0%
			GBM	0.562	0.430	0.81	10	80.0%
			SVM	0.773	1.000	0.63	10	76.0%
			RF	0.765	1.000	0.62	10	76.0%
			GLMNET	0.817	1.000	0.69	10	66.0%
			PAM	0.930	1.000	0.87	10	80.0%
	Low Sample	Subset	PLSDA	0.690	0.875	0.57	10	61.0%
			GBM	0.395	0.300	0.58	10	56.0%
			SVM	0.667	1.000	0.50	10	48.0%
			RF	0.359	0.975	0.22	10	48.0%
			GLMNET	0.801	0.975	0.68	10	68.0%
			PAM	0.806	0.900	0.73	10	78.0%
		Model Derived	PLSDA	0.792	0.895	0.71	14	50.0%
			GBM	0.585	0.525	0.66	20	36.5%
			SVM	0.802	1.000	0.67	5*	68.0%
			RF	0.406	0.925	0.26	17	25.9%
			GLMNET	0.830	1.000	0.71	32	28.4%
			PAM	0.925	1.000	0.86	48	20.8%
		Ensemble	PLSDA	0.694	0.863	0.58	10	66.0%
			GBM	0.455	0.350	0.65	10	62.0%
			SVM	0.788	1.000	0.65	10	70.0%
			RF	0.606	0.975	0.44	10	56.0%
			GLMNET	0.765	0.975	0.63	10	57.0%
			PAM	0.748	0.900	0.64	10	72.0%

3.1.3 Multiclass Classification - Low-Samples

3.1.3.1 Random and correlated and discriminatory datasets (N=100, p=50/1000 (NMR/MS))

NMR-scale random and correlated datasets generally had low accuracy and stability whereas MS-scale had four algorithms consistently with accuracy >= 0.700 but stability still remained low. Ensemble analysis of random and correlated datasets had little effect on performance of both NMR-scale and MS-scale (Supplementary File – S1).

3.1.3.2 Discriminatory Datasets

Analysis of the NMR-scale discriminatory dataset determined SVM as the best performing algorithm with the highest RPT whereas PAM and RF had the highest stability and accuracy respectively. The model derived results report GLMNET and PAM among the best; however, as with the binary classification problems the low sample size resulted in untrimmed features wherein nearly all were retained resulting in a decreased PPI%. Curiously, SVM reported the highest PPI% (64%) again suggesting that the lower sample size may restrict to only a few of the most discriminating features (Table 2). The MS-scale dataset had high predictive accuracy with SVM, GLMNET and PAM but very low with PLSDA and GBM. Curiously, PLSDA had the highest PPI% and stability (0.53)

suggesting a potential use in comparison to better classifying algorithms for feature selection (Supplementary File – S2).

Ensemble analysis of the NMR-scale discriminatory dataset reported GLMNET with the highest RPT but PLSDA with the highest PPI% (Table 2). This reflects the single run analysis whereby the improved stability of ensemble methods provided improved classification and stability of GLMNET. The MS-scale analysis reported SVM as the optimum algorithm with the highest RPT and accuracy but PAM reported the highest PPI%.

Table 2. Results from NMR-scale Multi-class Classification Simulations. RPT – Robustness-Performance Trade-off, PPI% - Percent features positively identified. *SVM doesn't have internal cutoff so defaults to top 10%

			Method	RPT	Accuracy	Stability	# Features	PPI%
Multi-class	High Sample	Subset	PLSDA	0.432	0.380	0.50	10	49.0%
			GBM	0.398	0.340	0.48	10	46.0%
			SVM	0.531	0.565	0.50	10	42.0%
			RF	0.257	0.900	0.15	10	30.0%
			GLMNET	0.615	0.575	0.66	10	52.0%
			PAM	0.586	0.520	0.67	10	53.0%
		Model Derived	PLSDA	0.441	0.373	0.54	12	44.2%
			GBM	0.926	0.895	0.96	50	19.6%
			SVM	0.559	0.485	0.66	5	60.0%
			RF	0.326	0.875	0.20	19	16.3%
			GLMNET	0.786	0.675	0.94	50	20.0%
			PAM	0.780	0.640	1.00	50	20.0%
		Ensemble	PLSDA	0.456	0.380	0.57	10	54.0%
			GBM	0.344	0.265	0.49	10	44.0%
			SVM	0.467	0.510	0.43	10	39.0%
			RF	0.299	0.890	0.18	10	35.0%
			GLMNET	0.538	0.495	0.59	10	34.0%
			PAM	0.557	0.520	0.60	10	52.0%
	Low Sample	Subset	PLSDA	0.485	0.394	0.63	10	58.0%
			GBM	0.397	0.325	0.51	10	38.0%
			SVM	0.723	0.888	0.61	10	52.0%
			RF	0.356	0.938	0.22	10	39.0%
			GLMNET	0.673	0.688	0.66	10	52.0%
			PAM	0.715	0.675	0.76	10	53.0%
		Model Derived	PLSDA	0.501	0.425	0.61	13	40.8%
			GBM	0.517	0.375	0.83	47	20.4%
			SVM	0.752	0.800	0.71	5	64.0%
			RF	0.365	0.888	0.23	17	21.2%
			GLMNET	0.871	0.863	0.88	49	19.4%
			PAM	0.900	0.825	0.99	50	20.0%
		Ensemble	PLSDA	0.485	0.394	0.63	10	62.0%
			GBM	0.388	0.300	0.55	10	38.0%
			SVM	0.654	0.788	0.56	10	53.0%
			RF	0.389	0.875	0.25	10	39.0%
			GLMNET	0.751	0.763	0.74	10	47.0%
			PAM	0.746	0.775	0.72	10	53.0%

3.1.4 Multiclass Classification - High Samples

3.1.4.1 Random and correlated and discriminatory datasets (N=200, p=50/1000 (NMR/MS))

NMR-scale random and correlated datasets continued to provide expected results, whereby most models had poor classification. In contrast, MS-scale data provided high accuracy but continued to provide very poor stability. Ensemble analysis of random and correlated datasets again reflected previous analysis with low accuracy and stability for both NMR-scale and MS-scale data (Supplementary File – S1).

3.1.4.2 Discriminatory Datasets

Analysis of the NMR-scale discriminatory dataset determined GLMNET as the best performing algorithm despite lower accuracy (Table 2). Model derived results determined GLMNET and PAM as the best models however they did not successfully extract any discriminating features resulting in a depleted PPI% whereas SVM reported the highest PPI% at 60%.

Ensemble analysis of the discriminatory dataset did not significantly improve performance for the NMR-scale dataset. This is also reflected in the MS-scale data where only RF and PLSDA stability improved slightly (Supplementary File – S2).

3.2 Real Datasets

3.2.1 Eisner – Urine analysis of Cachexia via NMR

Non-ensemble analysis determined PAM and GLMNET as the best overall models according the RPT (0.711, 0.705). Important features were extracted with the *feature.table* function. Adipate, Glucose, 3-Hydroxyisovalerate were identified in all subsamples by both models. PAM also identified creatine and succinate consistently (Table 3). GLMNET identified leucine, quinolinate, and valine as important features (Table 4). Seven of these metabolites were within the top 10 metabolites identified by Eisner et. al. with creatine being the 11th (Eisner et al., 2011). Furthermore, the other 3 metabolites in the top 10 (myo-inositol, betaine, and *N,N*-dimethylglycine) were also identified in the top 10 by GLMNET and PAM. Random forest provided the best classification accuracy at the expense of stability. SVM performed similarly well with respect to classification accuracy (Supplementary File – S3), however, stability was also quite low (0.31).

Ensemble methods improved stability of GBM, SVM, and RF; however, it noticeably decreased GLMNET stability from 0.64 to 0.49. Performance decreased slightly for PLSDA but improved for PAM; both proved the best overall models. Overall 8 of the 9 identified metabolites (frequency >= 0.9) by PLSDA and PAM were in the top 10 identified by Eisner et.al. (Supplementary File – S3). For this particular dataset the ensemble approach does not appear necessary as neither model performance nor feature stability was significantly improved. Irrespective, the applied methods provide further validation and support for the classification and metabolites selected.

Table 3. Feature table of PAM analysis consisting of consistency (i.e. number of times the feature identified as important) and frequency (i.e. percentage of iterations feature identified)

PAM Feature Table		
features	consistency	frequency
Adipate	10	1
Creatine	10	1
Glucose	10	1
Succinate	10	1
X3.Hydroxyisovalerate	10	1
myo.Inositol	9	0.9
Betaine	7	0.7
Glutamine	7	0.7
Quinolinate	6	0.6
cis.Aconitate	6	0.6
Acetate	5	0.5
N.N.Dimethylglycine	5	0.5
Lysine	3	0.3
Leucine	2	0.2

Table 4. Feature table of GLMNET analysis consisting of consistency (i.e. number of times the feature identified as important) and frequency (i.e. percentage of iterations feature identified)

GLMNET Feature Table		
features	consistency	frequency
Adipate	10	1
Glucose	10	1
Leucine	10	1
Quinolinate	10	1
Valine	10	1
X3.Hydroxyisovalerate	10	1
myo.Inositol	9	0.9
Succinate	6	0.6
Betaine	4	0.4
Glutamine	4	0.4
Lysine	4	0.4
Creatine	3	0.3
N.N.Dimethylglycine	3	0.3
Acetate	3	0.3
Alanine	2	0.2
Formate	1	0.1
Xylose	1	0.1

3.2.2 Ametaj – Analysis of Rumen Metabolism via NMR

The PAM and GLMNET models performed very well with the Ametaj dataset with RPT values of 0.850 and 0.806 respectively. Although PLSDA did not have a high RPT (0.497), the stability was very high (0.77) and was therefore evaluated for consistency with PAM and GLMNET. Glucose, endotoxin, and methylamine were consistently identified by PAM, GLMNET and PLSDA (Supplementary File – S4). Glucose and endotoxin were expected and methylamine was the first statistically significant metabolite discussed by Ametaj et.al. (Ametaj et al., 2010). In addition, uracil, acetate, fumarate, and lactate were also consistently identified by at least two models. All of these metabolites are identified by Ametaj et.al. except for lactate which is reported as non-significant but was approaching significance (0.149). This suggests a potential power issue although the authors' comment on the conversion of lactate to proprionate in ruminates is well supported.

Ensemble methods improved stability of PLSDA, GBM, SVM, and RF; however, it slightly decreased GLMNET and PAM stability by 0.02 and 0.07 respectively (Supplementary File – S4). This general improvement is expected with smaller sample sizes as variability often greater. The stability of all models was very high with all >= 0.67 except for RF. Stability of PLSDA was highest; however, classification accuracy was poor (0.438). Accuracy increased slightly with PAM and GLMNET but decreased in SVM to match GLMNET. These three proved to be the best overall models. Although the stability was high, identified features did vary between models creating what we refer to as a 'hierarchy of confidence' whereby the greatest confidence would be placed in the most consistently identified features within and across algorithms. Endotoxin, glucose, and methylamine were once more identified by four models (PAM, SVM, PLSDA and GBM). Alanine was also identified by 3 models (PAM, GLMNET, PLSDA). In addition, acetate, 3-PP, and uracil were also identified by at least two models consistently. All of these metabolites were identified as important by Ametaj et.al. Lastly, although most results were consistent, ferulate was also identified by PAM and SVM which was not discussed previously. It is also apparent that given the small sample size and greater variability within the data, this analysis benefits from ensemble methods.

3.2.3 Xiao – Serum Analysis of Hepatocellular Carcinoma via MS

Within the negative mode comparisons PAM mostly performed better than all other algorithms applied with respect to RPT (Supplementary File – S5). There was only one exception where PAM and GLMNET had almost

identical RPT values (0.679, 0.680). However, PAM stability was consistently the highest suggesting it to be the ideal method to explore this particular dataset. Additionally, it should be noted that PAM did not have very high accuracy compared to other methods such as GLMNET and SVM. Random Forest also had very high accuracy but exceptionally low stability making it a good tool for analyses that require high classification accuracy but do not require further feature identification. The higher accuracy makes GLMNET a close second to the PAM approach. An ensemble analysis was performed, however only RF improved stability significantly rendering the ensemble analysis of little help. Identified features were largely consistent with previous analyses (Supplementary File – S6), however, each comparison identified at least an additional 15 metabolites not mentioned in the prior manuscript.

Within the positive mode comparisons GLMNET performed slightly better than PAM in all except the 1 comparison; however, PAM consistently maintained the highest stability demonstrating that even with very high-dimensional datasets no single algorithm dominates. As with the negative mode, an ensemble analysis was performed but only RF improved stability significantly making ensemble aggregation unnecessary. As with the NMR studies the two models provided further support to mutually identified features (Supplementary File – S7). These features included GDCA, Oleoylcarnitine, GCDCA, L-N_2-(2-Carboxyethyl) arginine, Tetracosahexaenoic acid, Palmitoyl carnintine and Linoelaidyl carnitine which support the author's interpretations of bile and fatty acid metabolism. In addition, 48 metabolites were consistently identified by PAM but were not mentioned in the paper however it is possible they were identified as the authors do not present all identified metabolites.

4. Discussion

The results presented focused on determining if a single classification algorithm performs better than other commonly used algorithms in the field of metabolomics. Although there is strong support for several algorithms, there are additional papers stating one algorithm outperforms another. This is expected according to the concept of No Free Lunch Theorem (Wolpert & Macready, 1997) which, in essence, states that there is no single model that is appropriate for all problems. Our analyses support this theory and suggest a comparative approach to evaluate algorithm performance as well as to provide additional support via multiple algorithms for future conclusions. As such, the Bioconductor package OmicsMarkeR has been created to facilitate the rapid comparison of algorithm performances on individual datasets.

The null datasets initially appeared to demonstrate abnormally high accuracy and AUROC values when such metrics theoretically should result in ~0.50. However, these final models were re-tuned and built upon the 'best' features identified likely resulting in an inflated model metric. This has provided an example of the 'optimistic bias' that may result from feature selection (Ambroise & McLachlan, 2002). Furthermore, increased dimensionality of mass spectrometry datasets also increases the likelihood of finding discriminating variables by chance. Increased noise may also result in an inflated model metric. It is important to note that despite high model metrics, the low stability and RPT values provide strong evidence that the relationship has been found by chance and no further information can be reliably drawn from the model thereby increasing the value of such metrics in metabolomics investigations. Using the *fit.only.model* function on the null datasets confirmed the datasets as a whole did not possess any inherent classification whereby Accuracy and AUROC were ~0.50 (data not shown).

The binary and multivariate simulation results demonstrate the variation in optimal models. Depending on the users approach to a dataset (i.e. feature subset or model derived), number of samples relative to features, the number of groups, or application of ensemble methods, a different algorithm may be more appropriate. Within the simulation analyses PLSDA, SVM, PAM, RF, and GLMNET all proved to be optimal algorithms for different situations. Curiously, the reduction in features generally decreased GBM performance but feature selection was relatively consistent with other methods. Random Forest, in most cases, did not prove to be the optimal algorithm but consistently had among the highest accuracy but very low stability; which could be partially mitigated by an ensemble approach. This low stability is expected given the algorithms search encompasses interactions between variables resulting in a larger search space. If a user wished to weigh accuracy more than stability, the RPT value can be easily recalculated with the *RPT* function by increasing the beta parameter (e.g. beta = 1.5 increases the weight of accuracy by 50%).

The Eisner cachexia NMR dataset provided an example of the value behind using multiple algorithms. In the event that algorithms have similar overall performance, analysis of features selected by both may provide further support for identified features. Additionally, this dataset demonstrated the circumstance whereby ensemble approaches are superfluous. Consistent with published results, the Eisner dataset did reflect the high accuracy of SVM but stability was low. The GLMNET and PAM analyses provided respectable accuracy and high stability. The ability to assess stability of identified features for classification models is valuable if rapid and clinically applicable tests are to be developed.

The Ametaj cow diet NMR dataset also benefited from multiple algorithms adding further support to identified features. Furthermore, the multi-class experimental design and low number of samples likely resulted in higher variability. As such the ensemble approach was beneficial by improving most algorithm performances.

The Xiao Hepatocellular Carcinoma MS dataset was analyzed in four separate comparisons within each mode (i.e. positive or negative) following the manuscripts analysis. The negative mode results suggest that PAM was generally the optimum algorithm with respect to RPT values. The positive mode results suggest that GLMNET was the optimum algorithm. However, the goal behind the study was the identification of biomarkers. As such, the consistently higher stability of PAM is noteworthy. A potential extension would be to use the features identified by PAM and build the GLMENT model (or other method) for class prediction to improve accuracy.

As an example, we provide a very concise sample summary of the final positive comparison which could be used as a general guideline: *Six commonly applied metabolomics algorithms were tuned and cross-validated with the OmicsMarkeR bioconductor package via 10-fold cross-validation (k.fold = 10). Feature selection was accomplished with default model parameters. After comparison of overall performance Prediction Analysis of Microarrays (PAM) provided the highest Robustness-Performance-Trade-off (RPT) balancing both classification accuracy (73%) and stability of identified metabolites (Jaccard = 0.74) reflecting good prediction and reproducible metabolite identification. Metabolites identified in all data perturbation runs (k = 10) include GDCA, Oleoylcarnitine, GCDCA, L-N_2-(2-Carboxyethyl) arginine, Tetracosahexaenoic acid, Palmitoyl carnintine and Linoelaidyl carnitine. Forty-eight (48) additional metabolites were consistently identified that may be pursued via tandom mass spectrometry (MS/MS).*

5. Conclusion

Within this paper, simulated datasets were used to demonstrate that algorithm performance varies depending on the investigator's approach to a dataset (i.e. defined number or unknown), number of samples relative to features, the known groups, and/or if ensemble methods are applied. We also applied this comparative approach to three published, typically designed and freely accessible datasets including two NMR and one MS datasets. Following tuning and cross-validation, the optimal algorithms were used to compare selected features to the respective studies. Results proved very consistent whereby most discriminating features were identified; however, additional features were also identified that demonstrate the variation in applied methods and complexity of these large datasets.

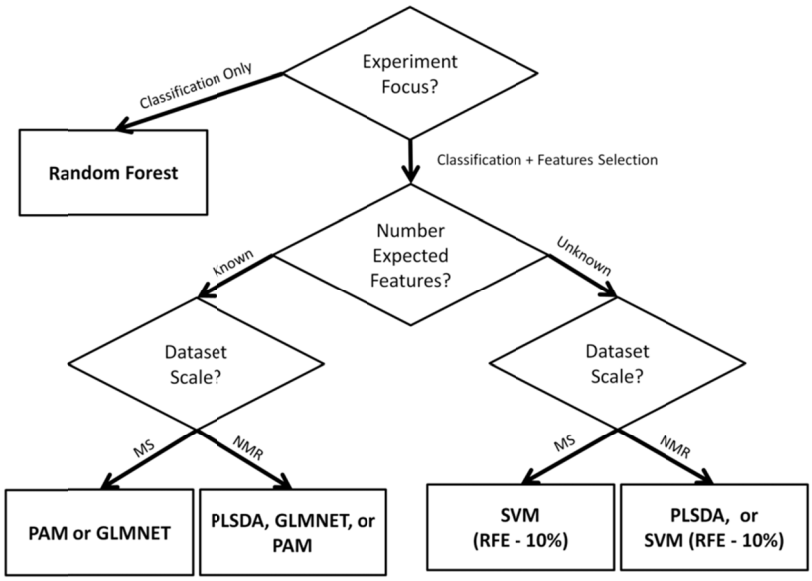

Figure 1. General guideline for which algorithms to apply to a given dataset. The reader is still encouraged to compare multiple algorithms and employ ensemble analysis given the variability in biological datasets

This comparative analysis provides a means to objectively choose a particular algorithm in addition to stability metrics to provide the highest confidence in potentially identified biomarkers and direct more focused independent validation. From these results the choice of the 'best' algorithm appears dependent upon the goals of the experiment in addition to the structure of the dataset. If classification is the primary goal random forest is an excellent option. If feature selection is also important there a few options the must be considered. These include the

number of expected discriminatory features and the datasets scale (i.e. MS or NMR). The diagram in Figure 1 is meant to be a general guide; however, the results herein strongly suggest using multiple algorithms and comparing performance on each unique dataset as well as ensemble methods which have been made far more accessible with the *OmicsMarkeR* package. It is important to emphasize, however, that there are far more considerations to be considered including the strengths and weaknesses of the available technology such as sensitivity, reproducibility, costs, and sample preparation (Robertson, 2005). In addition, preprocessing methods from normalization techniques to NMR preprocessing (e.g. binning) to the diverse software implementations of mass spectrometry processing (Castillo, Gpalacharyulu, Yetukuri, & Orešič, 2011) likely will impact results and further comparisons are encouraged.

In addition, the package contains two Monte-Carlo permutation functions (*perm.class* and *perm.features*) for assessing model performance and identified features for further evaluation. Current areas of improvement include the addition of further algorithms, database searching, improving memory efficiency, and easy to access graphics (e.g. scores plots, variable importance plots, etc.). The R package *OmicsMarkeR* for this analysis is publically accessible from the bioconductor platform (www.bioconductor.org).

References

Abeel, T., Helleputte, T., Peer, Y. V. de., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics, 26*(3), 392-398. http://dx.doi.org/10.1093/bioinformatics/btp630

Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci, 99*(10), 6562-6566. http://dx.doi.org/10.1073/pnas. 102102699

Ametaj, BN., Zebeli, Q., Saleem, F., Psychogios, N., & Lewis, M. J. (2010). Metabolomics reveals unhealthy alterations in rumen metabolism with increased proportion of cereal grain in the diet of dairy cows. *Metabolomics, 6*(4), 583-594. http://dx.doi.org/10.1007/s11306-010-0227-6

Bain, J. R., Stevens, R. D., Wenner, B. R., Ilkayeva, O., & Muoio, D. M. (2009). Metabolomics Applied to Diabetes Research Moving From Information to Knowledge. *Diabetes, 58*(11), 2429-2443. http://dx.doi.org/10.2337/db09-0580

Bodi, V., Sanchis, J., Morales, J. M., Marrachelli, V. G., & Nunez, J. (2012). Metabolomic Profile of Human Myocardial Ischemia by Nuclear Magnetic Resonance Spectroscopy of Peripheral Blood SerumA Translational Study Based on Transient Coronary Occlusion Models. *J Am Coll Cardiol, 59*(18), 1629-1641. http://dx.doi.org/10.1016/j.jacc.2011.09.083

Bogdanov, M., Matson, W. R., Wang, L., Matson, T., & Saunders-Pullman, R. (2008). Metabolomic profiling to develop blood biomarkers for Parkinson's disease. *Brain, 131*(2), 389-396. http://dx.doi.org/10.1093/brain/awm304

Breiman, L. (1996). Bagging predictors. *Mach Learn, 24*(2), 123-140. http://dx.doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Mach Learn, 45*(1), 5-32. http://dx.doi.org/10.1023/A:1010933404324

Carraro, S., Rezzi, S., Reniero, F., Héberger, K., & Giordano, G. (2007). Metabolomics Applied to Exhaled Breath Condensate in Childhood Asthma. *Am J Respir Crit Care Med, 175*(10), 986-990. http://dx.doi.org/10.1164/rccm.200606-769OC

Cassel, C., Hackl, P., & Westlund, A. H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *J Appl Stat, 26*(4), 435-446. http://dx.doi.org/10.1080/02664769922322

Castillo, S., Gopalacharyulu, P., Yetukuri, L., & Orešič, M. (2011). Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemom Intell Lab Syst, 108*(1), 23-32. http://dx.doi.org/10.1016/j.chemolab.2011.03.010

Chen, T., Xie, G., Wang, X., Fan, J., & Qiu, Y. (2011). Serum and Urine Metabolite Profiling Reveals Potential Biomarkers of Human Hepatocellular Carcinoma. *Mol Cell Proteomics, 10*(7), M110-004945. Retrieved February 27, 2013, from http://www.mcponline.org/content/10/7/M110.004945

Cramer, R. D., DePriest, S. A., Patterson, D. E., Hecht, P., & Kubinyi, H. (1993). 3D QSAR in drug design: theory, methods and applications. *3D QSAR in Drug Design: Theory, Methods and Applications* (pp. 523-550).

Davis, C. A., Gerick, F., Hintermair, V., Friedel, C. C., & Fundel, K. (2006). Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics, 22*(19), 2356-2363. http://dx.doi.org/10.1093/bioinformatics/btl400

Davis, V. W., Schiller, D. E., Eurich, D., & Sawyer, M. B. (2012). Urinary metabolomic signature of esophageal cancer and Barrett's esophagus. *World J Surg Oncol, 10*(1), 271-283. http://dx.doi.org/10.1186/1477-7819-10-271

Determan, Jr. C., Lusczek, E. R., Witowski, N. E., Lexcen, D., & Mulier, K. E. (2014). Carbohydrate fed state alters the metabolomic response to hemorrhagic shock and resuscitation in liver. *Metabolomics,* 1-8. http://dx.doi.org/10.1007/s11306-014-0621-6

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology, 26*(3), 297-302. http://dx.doi.org/10.2307/1932409

Duarte, I. F., Rocha, C. M., Barros, A. S., Gil, A. M., & Goodfellow, B. J. (2010). Can nuclear magnetic resonance (NMR) spectroscopy reveal different metabolic signatures for lung tumours? *Virchows Arch, 457*(6), 715-725. http://dx.doi.org/10.1007/s00428-010-0993-6

Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci, 103*(15), 5923-5928. http://dx.doi.org/10.1073/pnas.0601231103

Eisner, R., Stretch, C., Eastman, T., Xia, J., & Hau, D. (2011). Learning to predict cancer-associated skeletal muscle wasting from 1H-NMR profiles of urinary metabolites. *Metabolomics, 7*(1), 25-34. http://dx.doi.org/10.1007/s11306-010-0232-9

Folleco, A., Khoshgoftaar, T. M., Van Hulse, J., & Bullard, L. (2008). Software quality modeling: The impact of class noise on the random forest classifier. In Evolutionary Computation, 2008. CEC 2008. *IEEE World Congress on Computational Intelligence* (pp. 3853-3859). http://dx.doi.org/10.1109/CEC.2008.4631321

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

Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., & Walker, L D. (2009). Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics, 10*(1), 259. http://dx.doi.org/10.1186/1471-2105-10-259

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Mach Learn, 46*(1-3), 389-422. http://dx.doi.org/10.1023/A:1012487302797.

Han, X., Rozen, S., Boyle, S. H., Hellegers, C., & Cheng, H. (2011). Metabolomics in Early Alzheimer's Disease: Identification of Altered Plasma Sphingolipidome Using Shotgun Lipidomics. *PLoS ONE, 6*, e21643. http://dx.doi.org/10.1371/journal.pone.0021643

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and Additive Trees. In *The Elements of Statistical Learning* (pp. 337-387). (Vol. 2, No. 1). New York. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-84858-7_10. Accessed 26 August 2013

Haug, K., Salek, R. M., Conesa, P., Hastings, J., & de Matos, P. (2013). MetaboLights--an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res, 41*, D781-D786. http://dx.doi.org/10.1093/nar/gks1004

Hische, M., Larhlimi, A., Schwarz, F., Fischer-Rosinský, A., & Bobbert, T. (2012). A distinct metabolic signature predicts development of fasting plasma glucose. *J Clin Bioinformatics, 2*(3), 1-10. http://dx.doi.org/10.1186/2043-9113-2-3

Houtkooper, R. H., Argmann, C., Houten, S. M., Cantó, C., & Jeninga, E. H. (2011). The metabolic footprint of aging in mice. Sci Rep, 1. Retrieved January 21, 2014, from http://www.nature.com/srep/2011/111031/srep00134/full/srep00134.html?WT.mc_id=FBK_SciReports

Jurman, G., Merler, S., Barla, A., Paoli, S., & Galea, A. (2008). Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics, 24*(2), 258-264. http://dx.doi.org/10.1093/bioinformatics/btm550

Kuncheva, L. I. (2007). *A Stability Index for Feature Selection ACTA* (pp. 421-427). Retrieved September 3, 2013, from http://www.actapress.com/Abstract.aspx?paperId=29484

Lee, J. W., Lee, J. B., Park, M., & Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statatistics & Data Analysis, 48*(4), 869-885. http://dx.doi.org/10.1016/j.csda.2004.03.017

Lexcen, D. R., Lusczek, E. R., Witowski, N. E., Mulier, K. E., & Beilman, G. J. (2012). Metabolomics classifies phase of care and identifies risk for mortality in a porcine model of multiple injuries and hemorrhagic shock. *J Trauma Acute Care Surg, 73*(2), S147-S155. http://dx.doi.org/10.1097/TA.0b013e3182609821

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News 2*(3), 18-22.

Mahadevan, S., Shah, S. L., Marrie, T. J., & Slupsky, C. M. (2008). Analysis of Metabolomic Data Using Support Vector Machines. *Anal Chem, 80*(19), 7562-7570. http://dx.doi.org/10.1021/ac800954c

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2012). *e1071: Misc Functions of the Department of Statistics* (e1071). Retrieved from http://CRAN.R-project.org/package=e1071

Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays. *The Lancet, 365*(9472), 1684-1685. http://dx.doi.org/10.1016/S0140-6736(05)66539-7

Nishiumi, S., Shinohara, M., Ikeda, A., Yoshie, T., & Hatano, N. (2010). Serum metabolomics as a novel diagnostic approach for pancreatic cancer. *Metabolomics, 6*(4), 518-528. http://dx.doi.org/10.1007/s11306-010-0224-9

Ochiai, A. (1957). Zoogeographical studies on the soleoid fishes found in Japan and its neigbouring regions. *Bull Jpn Soc Sci Fish, 22*(9), 526-530.

Patterson, A. D., Maurhofer, O., Beyoğlu, D., Lanz, C., & Krausz, K. W. (2011). Aberrant Lipid Metabolism in Hepatocellular Carcinoma Revealed by Plasma Metabolomics and Lipid Profiling. *Cancer Res, 71*(21), 6590-6600. http://dx.doi.org/10.1158/0008-5472.CAN-11-0885

Qiu, Y., Cai, G., Su, M., Chen, T., & Liu, Y. (2010). Urinary Metabonomic Study on Colorectal Cancer. *J Proteome Res, 9*(3), 1627-1634. http://dx.doi.org/10.1021/pr901081y

Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., & Boxer, A. (2007). Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. *Nat Med, 13*(11), 1359-1362. http://dx.doi.org/10.1038/nm1653

Real, R., & Vargas, J. M. (1996). The Probabilistic Basis of Jaccard's Index of Similarity. *Syst Biol, 45*(3), 380-385. http://dx.doi.org/10.2307/2413572

Ridgeway, G. (2013). gbm: Generalized Boosted Regression Models. Retrieved from http://CRAN.R-project.org/package=gbm

Robertson, D. G. (2005). Metabonomics in Toxicology: A Review. *Toxicol Sci, 85*(2), 809-822. http://dx.doi.org/10.1093/toxsci/kfi102

Rohart, F., Paris, A., Laurent, B., Canlet, C., & Molina, J. (2012). Phenotypic prediction based on metabolomic data for growing pigs from three main European breeds. *J Anim Sci, 90*(13), 4729-4740. http://dx.doi.org/10.2527/jas.2012-5338

Romero, R., Mazaki-Tovi, S., Vaisbuch, E., Kusanovic, J. P., & Chaiworapongsa, T. (2010). Metabolomics in premature labor: a novel approach to identify patients at risk for preterm delivery. *J Matern Fetal Neonatal Med, 23*(12), 1344-1359. http://dx.doi.org/10.3109/14767058.2010.482618

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., & Gibb, W. J. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med, 19*(5), 619-625. http://dx.doi.org/10.1038/nm.3175

Saeys, Y., Abeel, T., & Peer, Y. V. de (2008). Robust Feature Selection Using Ensemble Feature Selection Techniques. In: Machine Learning and Knowledge Discovery in Databases (pp. 313-325). Springer Berlin Heidelberg. Retrieved August 23, 2013, from http://link.springer.com/chapter/10.1007/978-3-540-87481-2_21

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517. http://dx.doi.org/10.1093/bioinformatics/btm344

Sanchez, G. (2012). DiscriMiner: Tools of the Trade for Discriminant Analysis. R Package Version 01-29. Available: http://CRAN.R-project.org/package=DiscriMiner

Shi, L., Tong, W., Fang, H., Scherf, U., & Han, J. (2005). Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics., 6*(Suppl 2), S12. http://dx.doi.org/10.1186/1471-2105-6-S2-S12

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol Skr, 5*, 1-34.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., & Sheridan, R. P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J Chem Inf Comput Sci, 43*(6), 1947-1958. http://dx.doi.org/10.1021/ci034160g

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci, 99*(10), 6567-6572. http://dx.doi.org/10.1073/pnas.082099299

Vapnik, V. (1998). The Support Vector Method of Function Estimation. In Nonlinear Modeling (pp. 55-85). Springer US.. Retrieved January 2, 2014, from http://link.springer.com/chapter/10.1007/978-1-4615-5703-6_3

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330-349. http://dx.doi.org/10.1016/j.patcog.2010.08.011

Wahl, S., Holzapfel, C., Yu, Z., Breier, M., & Kondofersky, I. (2013). Metabolomics reveals determinants of weight loss during lifestyle intervention in obese children. *Metabolomics, 9*(6), 1-11. http://dx.doi.org/10.1007/s11306-013-0550-9

Wishart, D. S. (2010). Computational Approaches to Metabolomics. In: Bioinformatics Methods in Clinical Research (pp. 283-313). Humana Press. Retrieved April 9, 2013, from http://link.springer.com/protocol/10.1007/978-1-60327-194-3_14

Wold, S. (1975). Path models with laten variables: The NIPALS approach (pp. 307-357). In H. M. Blalock, A. Aganbegian, F. M. Borodkin, R. Boudon, & V. Capecchi (Eds.), *Quantitative Sociology: International perspectives on mathematical and statistical modeling*. New York, NY, USA: Academic Press.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *Evol Comput, IEEE Trans, 1*(1), 67 -82. http://dx.doi.org/10.1109/4235.585893

Wongravee, K., Lloyd, G R., Hall, J., Holmboe, M. E., & Schaefer, M. L. (2009). Monte-Carlo methods for determining optimal number of significant variables. Application to mouse urinary profiles. *Metabolomics, 5*(4), 387-406. http://dx.doi.org/10.1007/s11306-009-0164-4

Xiao, J. F., Varghese, R. S., Zhou, B., Nezami Ranjbar, M. R., & Zhao, Y. (2012). LC-MS Based Serum Metabolomics for Identification of Hepatocellular Carcinoma Biomarkers in Egyptian Cohort. *J Proteome Res, 11*(12), 5914-5923. http://dx.doi.org/10.1021/pr300673x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B, 67*(2), 301-320.

Zucknick, M., Richardson, S., & Stronach, E. A. (2008). Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods. *Stat Appl Genet Mol Biol, 7*(1). http://dx.doi.org/10.2202/1544-6115.1307