Relevant metabolites' selection strategies

Metabolomics Perspectives

Hageman, Jos

https://doi.org/10.1016/B978-0-323-85062-9.00010-6

CHAPTER

Relevant metabolites' selection strategies

10

Jos Hageman

Biometris, Applied Statistics, Wageningen University & Research, Wageningen, The Netherlands

Introduction

Metabolomics is the study of the metabolome or all small molecules or metabolites present in plants or any other organisms (Madsen et al., 2010; Weckwerth, 2003). The metabolome is dynamic and is the best reflection of the phenotype of the organism under study (Hageman, Hendriks, et al., 2008). A major goal of metabolomics studies is, besides gathering general knowledge of the study objects, finding biomarkers related to different kinds of traits (Hageman, van den Berg, et al., 2008; Hendriks et al., 2011). These traits can be very diverse ranging from biomarkers for (the onset of) disease states (Madsen et al., 2010) to predictors for crop ripening (Lombardo et al., 2011) or sensory perceptions (Lindinger et al., 2009). Metabolomics aims to simultaneously measure as many metabolites as possible. Currently, there is no single platform that can measure all metabolites, but by combining different platforms a comprehensive view of the metabolome is obtained. Metabolomics experiments typically measures hundreds of metabolites (Dettmer et al., 2007).

After obtaining a snapshot of the metabolome, a statistical model relating metabolites to the trait of interest is created. Statistical models are created to assess the association between metabolites (the predictors) and the trait of interest (the response variable). In statistical terms, the response variable is the variable whose variation depends on the values of the predictors (the metabolites) and is the outcome of statistical models. It can be a class membership (e.g., control vs infected), or a quantitative trait of interest (e.g., taste sweet). When creating these models, there are several moments and reasons why selection of metabolites is useful.

1. Usually, a limited number of metabolites is biochemically connected to the trait of interest. Therefor it is not expected that all metabolites contribute equally to the prediction of the trait of interest (Doeswijk et al., 2011; Hageman, van den Berg, et al., 2008; Saccenti et al., 2014). The (most) predictive metabolites in these models are typically identified after creation of the model.
2. Some statistical techniques suffer greatly from the presence of many metabolites (predictors in the model): it hinders the creation of the models, makes them prone to overfitting. Overfitting happens when statistical models

Metabolomics Perspectives. DOI: https://doi.org/10.1016/B978-0-323-85062-9.00010-6

no longer learn the general trend of the data but learn the peculiarities of the data at hand. Large numbers of metabolites also slow down the creation and validation of the models. It is therefore important to limit the number of metabolites at the start of the modeling.

3. Another reason is that small metabolite sets are desirable as they make it easier to obtain insights into mechanisms (as compared to considering the whole observed metabolome). A small set of relevant metabolites connects better to research devoted to identifying biomarkers for the prediction of class membership (e.g., control vs infected) or any small set of metabolites that predict a quantitative trait the best (Hageman, Hendriks, et al., 2008; Saccenti et al., 2011). The purpose of the biomarkers is typically to use them routinely in a targeted, single platform setup for fast classification. As metabolomics can measure hundreds of metabolites, variable selection is a must.

This chapter provides an overview of different strategies to reduce the high dimensional dataset derived from a metabolomics experiment. For ease of the overview, all variable selection methods and techniques have been divided into one of three categories (He & Yu, 2010; Shahrjooihaghighi et al., 2017), see for an overview: Table 10.1.

1. Low-level variable selection: deselect metabolites that are completely uninteresting. Low-level variable selection is primarily focused on removing non-informative or redundant variables (Shahrjooihaghighi et al., 2017). These approaches are sometimes referred to as filter methods as they act as a filter to separate the promising metabolites from the not-so-promising ones.
2. Medium-level variable selection: methods that intrinsically select metabolites. Methods in this category select important metabolites or deselect the ones that are not important as part of their inner workings. A typical result would be a statistical model that uses not all metabolites as predictors but is using a small subset of metabolites.
3. High-level variable selection: Assess explicit importance of metabolites. High-level variable selection entails methods that intrinsically select metabolites as part of the inner working of the modelling technique itself. This category also includes algorithms or heuristics that indicate the importance of metabolites through some importance criterion. So, methods in this category can possess innate variable selection properties but this is not necessary as this can also be obtained through external criteria. The latter typically results in a ranked list.

Low-level variable selection

Low-level variable selection or filter methods can be divided into two categories: supervised and unsupervised methods. The difference between the two is whether they consider the trait of interest in their decision to exclude metabolites or to retain them.

Table 10.1 Overview of the three different variable selection levels, including a description and example categories.

Variable selection level	Focus	Examples
Low	Removal of noninformative information	Unsupervised —percentage observed —variance based supervised —correlation with trait —fold change —hypothesis tests
Medium	Selection of important metabolites	Wrapper methods —stepwise regression —Global optimization techniques —genetic algorithms —simulated annealing —tabu search
High	Variable selection is intrinsically to the method or variable importance is indicated with an auxiliary criterion	Embedded techniques —regularized regression —latent variable techniques —tree based methods —support vector machines Heuristic approaches —bootstrapping —cross validation

Unsupervised low-level variable selection

In unsupervised variable selection, the response variable is not considered in the decision to select or deselect a metabolite. Metabolites are selected purely on the observed metabolic profile. The methods in this category have in common that they assess the observed variation in each metabolite. When the observed variation is deemed too low, the metabolite will be discarded.

Percentage observed

Not all metabolites are always observed in every sample. This can have different causes. Metabolites can be completely absent in some samples while being present in others. When a metabolite is present it can sometimes be present in a concentration that does not reach the detection limit. Another cause for not observing a metabolite could be of a more technical nature, like misalignments in processing of the measurements. All these cases lead to missing values for certain metabolites. When a metabolite has not been observed in a certain number of samples, it usually gets thrown out. Thresholds vary between 30%−60%.

Missing values in metabolomics is very common. Many statistical methods require a complete data set with no missing values. Imputing a modest amount of missing values is typically not a big problem. Metabolites that are not observed in a large part of the samples require too many values to be substituted. Metabolites

that have many imputed values may not be the most reliable ones to use in subsequent statistical analysis. Removing them from future statistical analysis reduces the chance of false positives.

Researchers should be very careful with removing metabolites using the percentage observed criterium. The absence of a metabolite (or rather being below the detection limit) in a group of samples could, in principle, correlate with different experimental factors making these metabolites the well sought biomarkers.

Variance based

Another criterion to consider in the selection of the metabolites fit for future analysis is assessing if the variation contained in metabolites is sufficient. It is expected that not all metabolites will respond to experimental or observational factors. In principle this means that the metabolite concentrations remain very much the same throughout all samples. Including these metabolites in statistical analysis does not make sense as they cannot be expected to function as biomarkers or otherwise as predictive metabolites. General values for a minimal threshold variance are difficult to give and are data set and probably even metabolite dependent. Thresholds are connected to the relative standard deviation of metabolites and general noise levels of the data.

Supervised low-level variable selection

Where unsupervised low-level variable selection methods for the most part deselect metabolites that do not have sufficient variation, supervised low-level variable selection methods select metabolites with sufficient relevant variation. This time the response variable is considered when deciding which variables to include or exclude as methods in this category assess the relationship between individual metabolites and the response. Metabolites that show a low degree of association can be discarded before any statistical model is even created. The idea is only to retain the metabolites that loosely show some association with the response. Based on the nature of the response, qualitative (like class membership) or quantitative (like taste sweet), different criteria can be used. All criteria calculate scores which are the basis on which to select the metabolites that will be used in subsequent statistical modeling.

Quantitative response

Several correlation metrics are available for expressing relatedness between a metabolite and a quantitative response.

Pearson's correlation coefficient

This expresses the linear association between two quantitative variables (in our case a metabolite and a quantitative response variable). The correlation coefficient is a dimensionless number between -1 and $+1$. Values close to -1 and $+1$ indicate a strong linear association, while outcomes close to 0 indicates that there

is no linear association. Correlation coefficient estimates can be tested using a *t*-test to test if the estimate is significantly different from 0. Selection of metabolites can be done in two ways, a threshold on the correlation coefficient estimate can be used, say the selection of metabolites with an absolute correlation coefficient of for example, >0.3. Alternatively, metabolites for which the hypothesis test has shown the estimate to be significantly different from zero can be selected. If desired, partial correlation coefficients can be calculated. These are Pearson correlation coefficients but corrected for the influence of other metabolites. In the case of suspected outlying values, nonnormally distributed metabolites/traits or nonlinear relations between the two, it can be useful to use Spearman rank correlation. In this procedure, metabolite and traits values are first converted to rank numbers followed by the usual Pearson correlation coefficient.

Other metrics for establishing the degree of association between a metabolite and a quantitative trait are for example, distance related criteria such as Euclidian or Manhattan distance and Fisher's score.

Qualitative response

When the response variable is qualitative (like the classes *control* and *infected* but this is easily extended to more than two groups), we would like to know if metabolites appear to be up or down regulated in one of the groups. We would like to select metabolites for further analysis that show some degree of difference between our classes and thereby deselect the ones that do not show any relatedness with the treatments.

Fold change

A fold change describes how much the average metabolite concentration has changed between two groups (van den Berg et al., 2006). It is the ratio of the average metabolite concentrations of the groups. It is calculated for each metabolite separately. Fold change is connected to effect size, the bigger the fold change, the larger the difference between the two groups and thus the larger the effect size. The idea is to retain only metabolites that have a fold change above a certain threshold. Metabolites with large fold changes show clear differences between the two groups and could potentially be of interest. An important caveat is that metabolites showing a large fold change are not necessarily more important than metabolites showing a small fold change.

Hypothesis testing

Using hypothesis tests, it is investigated if metabolites show a significant difference between treatments groups. Hypothesis tests can be used to select metabolites showing a significantly different response between two groups (using two independent samples *t*-test) or more than two groups (using analysis of variance, Anova models). In contrast to regular hypothesis testing, a very modest confidence level must be used, say 50% (or even lower) to ensure not only the most significant metabolites are retained. That could easily lead to missing a

multivariate set of metabolites describing group differences. By using a modest confidence level, even loosely associated metabolites can be selected.

Fold changes and *t*-tests are connected in a sense. Where fold changes only consider the means between groups (in the form of a ratio), *t*-tests also consider the variability of these means (using so called standard errors). When fold changes are large, but the variability of the estimates is also large, hypothesis tests will likely indicate there is no significant difference. Selection using a fold change criterion could still happen as it ignores the variability of the estimates. Graphically fold changes and *t*-test can be connected in Volcano plots, showing the magnitude of the difference and the significance of this difference (Hur et al., 2013).

Medium-level variable selection

The previous section described mainly methods for reducing the number of metabolites to be used in subsequent modeling. Subsequent modeling could be for example, a prediction model using some form of regression or a classification model. The idea is to only use the most promising metabolites in these analyses. Metabolites with general insufficient variation or variation insufficiently related to the trait of interest are discarded not to hamper these analyses. However, this still does not mean that all available metabolites for these analyses are all relevant and have equally important predictive properties. At this point it is also likely that we'll still have more metabolites present than we have samples or observations. In general, when we have more variables than data points, we have an ill-posed, or undetermined problem. There is simply not one unique solution that gives the best model fit. There are several ways of addressing this problem. One way is to use variable selection methods. The other is to use modeling techniques that explicitly handle large number of variables and have an embedded way of selecting variables.

Variable selection or wrapper methods

These methods have in common that they evaluate multiple models using different subsets of the metabolites. The purpose of analyzing different subsets of metabolites is to identify metabolites with the most predictive properties (Saeys et al., 2007). By analyzing the results from these subsets and combining different metabolites, they come to a (near) optimal set of metabolites for the prediction of the trait under investigation. Wrapper methods are multivariate in nature, in the sense that they study the predictive properties of (small) sets of metabolites simultaneously. Sets of metabolites that are chosen form a combination that together predict the trait of interest the best. They supplement each other in their predictive behavior.

Variable selection methods use ordinary least squares regression [or multiple linear regression (MLR) as it is sometimes called] for creating a model predicting

the trait from the metabolites. However, since we have more metabolites than samples, we need to reduce the number of metabolites in the model. Variable selection techniques aim at selecting only the most predictive metabolites.

There are two important approaches of variable selection: local methods for variable selection like for example, stepwise regression or global optimization algorithms.

Stepwise regression

Stepwise regression starts with an "empty" model, a model with only an intercept. In subsequent iterations metabolites are added one at a time to the model. The metabolite that gives the largest improvement to the model, as measured using F-tests, will be added to the model in each iteration. This iterative procedure of adding variables is repeated until the model cannot be improved anymore (the model does not change significantly). It can happen that a set of metabolites makes another metabolite obsolete as its variance is contained in the other metabolites. A metabolite is removed from the model when the model is not significantly different from a model with that metabolite present. There are two variants of stepwise regression. One is forward stepwise regression, and it does not allow the removal of variables. A second one is backward elimination. It starts with a model with all variables present. For metabolomics data, we typically have more metabolites compared to the number observations meaning this model cannot be calculated.

The result from stepwise regression is a small and compact set of metabolites that predict the trait of interest the best. It is not guaranteed to be the best subset possible. Stepwise is a greedy algorithm and can get stuck in local optima. Adaptations to stepwise regression are possible, one is e.g., to limit the number of selected metabolites to a prefixed number. The idea is that the first few selected metabolites are most important and explain the large majority of variance. Later added metabolites do not explain a lot of variance.

Global optimization algorithms

Stepwise regression can get stuck in a local optimum, meaning that the obtained solution (a set of metabolites) is not the best one possible. One way of overcoming this is the use of global optimization algorithms such as genetic algorithm (GA's) (Wehrens & Buydens, 1998), simulated annealing (SA) (Kirkpatrick et al., 1983) and tabu search (TS) (Glover, 1990). A complete discussion of these techniques would be beyond the scope of this chapter. In short, GA's mimic evolution and work on a group of trial solutions, called a population (Hageman, van den Berg, et al., 2008). A trial solution in our context would be a subset of metabolites. These trial solutions are recombined with each other (a process called crossover) and mutated into new trial solutions (called mutation). All trial solutions are evaluated, in our context that means their predictive properties are assessed. The best trials solutions, the ones with the best predictive properties, are kept and

serve as a starting point for the next generation. The process is repeated until, for a prespecified number of generations, no improvement has been encountered.

SA works on a single solution at a time. A solution in our context is a subset of metabolites. By taking small steps in the search space (and in this context steps in the search space are for example, changing one metabolite for another one) predictive properties of the model will change. Some metabolites will improve the model, whilst others will deteriorate the model. Changes that improve the model are kept, while changes that deteriorate the model are kept with a certain probability. During the optimization, this probability of accepting a worse solution is getting smaller and smaller. This decreasing probability mimics the cooling of a metal, hence the name simulated annealing. By allowing steps that deteriorate the model's performance, the algorithm can overcome local optima and eventually (or hopefully) reach the global optimum.

TS also works on a single solution at a time (Hageman et al., 2000, 2003; Hageman, Wehrens, et al., 2003). By modifying this solution (modifying in this context is adding/removing/changing selected metabolites), the search space around the current solution is investigated. Parts of the search space that have been visited are stored in memory and are never revisited again (they are taboo, hence the name). By forcing the algorithm to always explore a new part of the search space, the algorithm will eventually overcome local optima and should be able to reach the global optimum.

All three of these algorithms are finicky to use, because of the many settings they have. When used correctly they can be very powerful and deliver a compact set of metabolites able to best predict a certain trait.

High-level variable selection

Embedded methods for the selection of variables

Many statistical methods have been devised to get around the problem of having more variables compared to the number of observations. They do this by implicitly or explicitly selecting variables used in the model or by the creation of a set of low dimensional latent variables (LVs) (Gareth et al., 2013).

Regularization techniques

When we have more metabolites compared to the number of samples, the problem is ill-conditioned. The use of MLR would result in many models that fit the data equally well and would highly overfit the data. It has become likely that MLR would achieve low errors by fitting random fluctuations in the metabolite data that do not represent the true relationships between metabolites and the response. One way to overcome ill-conditioning is using constraints or penalty functions on the regression coefficients. This is called regularization. The idea is that besides the original loss function from MLR an added regularization term or

penalty is added to the equation (see Eq. (10.1)). Here n is the number of observations and p the number of predictors.

$$\text{Error} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_p x_{ij}\right)^2 + \text{penalty term on } \beta\text{'s} \tag{10.1}$$

Two main regularization strategies are ridge regression and the lasso both differing in the definition of the penalty term. Ridge regression uses the so-called L2-norm, while the lasso using the L1-norm (Tibshirani, 1996). The L2-norm minimizes the sum of squared regression coefficients; the L1-norm minimizes the sum of absolute regression coefficients (see Eqs. (10.2) and (10.3)).

$$\text{Ridge penalty term (L2 norm)}: \lambda\sum_{j=1}^{p}\beta_j^2 \tag{10.2}$$

$$\text{Lasso penalty term (L1 norm)}: \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{10.3}$$

So, regularization methods do not only optimize the fit of the data using the ordinary loss function, but they also minimize the regression coefficients themselves (Gareth et al., 2013). The lasso is of special interest for us since it has variable selection properties. While the L2-norm has the effect that many regression coefficients are shrunken toward zero, they never reach exactly zero. With the lasso and the L1-norm this is different, the L1-norm enables the deselection of variables by setting their corresponding regression coefficients exactly to zero. When regression coefficients reach exactly zero, they are effectively removed from the model, hence the variable selection properties of the lasso. The result of the lasso is a small subset of metabolites best able to predict the trait of interest. The balance between the loss function and the penalty term needs to be optimized by a meta-parameter (typically called λ) (Bujak et al., 2016).

Latent variable methods

A different strategy to deal with ill-conditioned problems is the use of LVs regression methods. Here, the metabolites are mathematically transformed into a low dimensional representation of the data. Each new dimension is to contain as much variation of the original the data. There are several ways these new dimensions can be constructed. With regression modeling in mind, two are of special interest, Principal Components Regression (PCR) and Partial Least Squares (PLS).

Principal component regression

Principal Component Analysis is a method for deriving dimension reduction by combining variables (metabolites in our case) into a small number of principal components (PCs) (Antonelli et al., 2019; Gareth et al., 2013). The PCs are constructed in such a way that each component describes as much of the variation of the data at hand. The components are ranked in decreasing order of explained

variance and are all uncorrelated (they are said to be orthogonal). Usually only a small number of PCs is needed to describe all relevant variation in a dataset; the remaining PCs describe very small amounts of variation usually associated with random noise. After the creation of the PCs, they are used in MLR as the explanatory variables, they take over the role of the metabolites.

Partial least squares

PCR involves the creation of PCs that describe as much of the variation as possible. In this step only the metabolites are involved, and the resulting PCs summarize these metabolites the best (Antonelli et al., 2019; Bujak et al., 2016). By focusing only on the variation in the metabolites, the important variation related to the prediction of the trait is not always retained in the first PCs. PLS solves this problem by the creation of LVs that do not only capture the variation of the metabolites but capture the variation in the metabolites that is relevant for the prediction of the response. PLS finds LVs that best summarize the metabolites and the response simultaneously.

For both methods, the number of PCs or the number of LVs is a metaparameter that needs to be optimized (Westerhuis et al., 2008).

PCs and LVs are typically difficult to interpret (Antonelli et al., 2019). They are a linear combination of all metabolites and as such all metabolites take part in the modeling. However, some metabolites are more important than other in the formation of the PCs/LVs. A metabolite's importance is expressed by so-called loadings. Loadings are weights that determine how much each individual metabolite contributes to a particular PC/LV. Inspection of the loadings will separate important metabolites from the unimportant ones; the higher the loading, the more important a certain metabolite is in that PC/LV. With PLS variable importance in projection (VIP) scores can be calculated. The importance of each variable is assessed, and VIP scores close to or greater to one are considered important in the PLS model.

There are several other modeling techniques that explicitly reduce or select variables to base their models on.

Decision trees

A decision tree is a flow chart like algorithm. It consists of nodes and branches where branches connect the nodes, and several branches emerge from a node. A node with no connection is typically called a leaf (Rokach, 2010). In decision trees, each node represents a test on a feature. Depending on the outcome of this test you traverse down different branches. This process is repeated several times until you reach a leaf (the end note). The leaf represents the predicted value. This can be a classification or a predicted variable in the case of a quantitative response. The test found on a nonend node relates to metabolites and contains a split point. It is a test on a single metabolite and depending on the outcome diverts to a certain branch. When following the node all the way to an end note, the end note contains a prediction for a certain object. An example tree is given in Fig. 10.1. Variable selection takes place at each

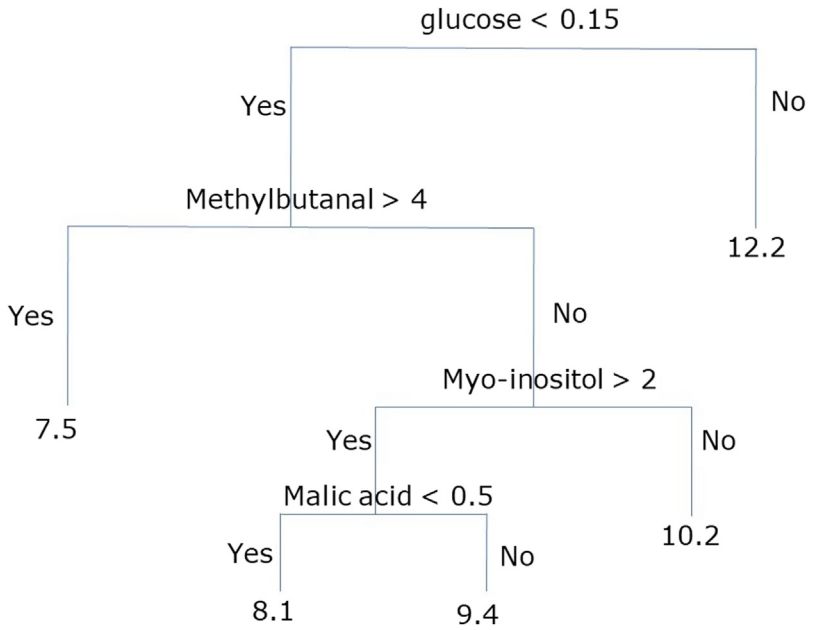

FIGURE 10.1

Decision tree. Example of decision tree. See text for details.

node since each node is a test on a single metabolite. Metabolites that do not appear in a node, are never selected, and do not influence the outcome. When training a tree, it is calculated how much each metabolite decreases the prediction error in a node. This is weighed by the chance to reach that node. The higher the value, called impurity, the more important a metabolite is.

Decision trees are easy to understand and interpret but are a bit unstable against small changes in the data (Gareth et al., 2013). They are also prone to overfitting which means they rather learn the specifics of the data set than generalize. This is something that can be reduced by pruning the tree which is the process of removing small noncritical branches.

Random forests

Random forests are an extension to decision trees. Random forests are an ensemble method and build and combines the output of many individual decision trees (Determan Jr, 2015; Gareth et al., 2013). Each tree in the forests is trained using a random subset of the original training set (a process called bootstrapping, see later in this chapter). As an addition, each time a tree is split, a random sample of the metabolites is chosen as split candidates. The number of metabolites available is typically the square root of the number of available metabolites. This has the

effect of forcing the decision tree to focus also on other important metabolites for predicting the response. It forces individual decision trees to look different from each other. This will make the trees uncorrelated and the average of the trees less variable and more reliable (Gareth et al., 2013). After training all trees in the forests (which is typically a large number, say 500−1000), the final prediction from the random forests is the average of all trees. The individual decision trees in random forests typically have many nodes and are not pruned. Variable selection in random forests is averaging all the impurity values from the individual decision trees. Alternative ways of variable selection in random forests are also present, one is the Boruta algorithm (Kursa & Rudnicki, 2010). Here, so called shadow features or randomized copies of the original variables are added to the data set. Random forests are trained, and a feature importance measure is applied. When the importance feature of actual metabolites is smaller compared to the ones from the randomized copies, these metabolites are removed. This process is repeated until convergence or a predefined number of forests.

Support vector machine

Super vector machine (SVM) is a machine learning algorithm typically used for classification but also able to predict quantitative traits (Grissa et al., 2016). SVMs are concerned with finding a hyperplane that best divides a dataset into two classes. Support vectors are the data points that are nearest to this hyperplane and thus define this hyperplane. Removal of these points would change the hyperplane. For a schematic overview see Fig. 10.2. The best hyperplane is the plane that separates the two classes while all data points are as far away as possible from the hyperplane (Vapnik, 1998). When no clear hyperplane exists, soft margins can be applied which allows for some misclassifications. When a nonlinear boundary is more appropriate, kernels can be used. Kernels allow for a transformation of the data after which a linear hyperplane can be found. Feature selection takes place by investigating the coefficients or weights of the model (Grissa et al., 2016). Large coefficients are deemed more important than smaller coefficients.

Heuristic approach

One recurring theme in variable selection with metabolomics is that the reported subset of metabolites is just one set out of many possible sets of metabolites that have a comparable fit. Owing to the correlated nature of many metabolites, metabolites have a certain interchangeable aspect with respect to their predictive properties. Methods with variable selection properties will include metabolites with high (or preferable the highest) explanatory properties in the model. Unfortunately, when sample sizes are low and the number of metabolites large (as is typically the case), the set of metabolites that is reported can show a

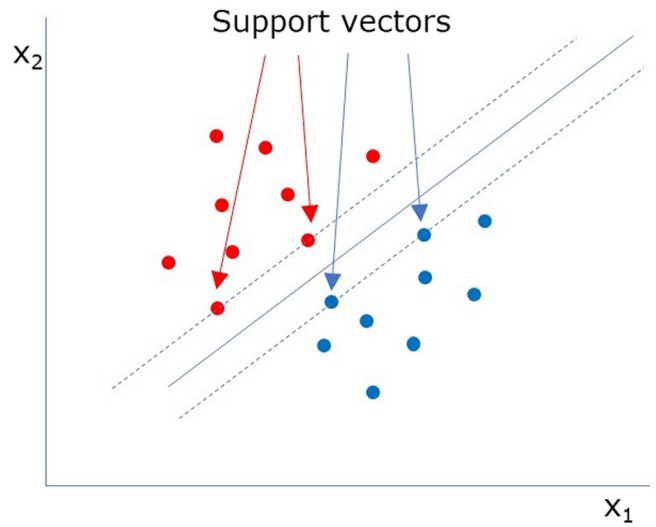

FIGURE 10.2

SVM hyperplane. Schematic overview of a separating (hyper)plane with several support vectors.

coincidental correlation with the trait of interest (Hageman et al., 2017; Westerhuis et al., 2008). This means that in the specific data set under study, metabolites appear to have important predictive properties but when the experiment and the statistical analysis is repeated, a (partly) different set of metabolites is reported. It is not always possible to repeat complete experiments, but we can mimic the process of analyzing different data sets using a computer. The central idea is that small changes in the data can result in substantial changes of the selected metabolites. Ideally, the set of selected metabolites would not change and would always be the same for every time we perturb our data set. Metabolites that are selected despite the perturbations of the data set are stable and, quite likely, the ones that carry the best predictive properties. On the other hand, metabolites that are not selected very often represent, most likely, coincidental correlations and probably do not hold up in future experiments.

Bootstrap and stability selection

Bootstrapping is a statistical procedure that resamples a single data set to create many simulated samples (Efron & Tibshirani, 1994). Bootstrapping allows you to calculate standard errors, confidence intervals and perform hypothesis tests for different kinds of sample statistics like for example, population mean, population difference of means etc. Under the correct assumptions these are sample statistics we could investigate using t distributions, but when we cannot make these

assumptions, bootstrapping can provide us with the required estimates. The power of the bootstrap lies in situations where there is no easy alternative available. We will not use the bootstrap only for estimation of for example, prediction quality but we will also study the metabolites that are selected during the bootstrap procedure.

During bootstrapping, a new and perturbed sample is generated by random drawing with replacement from the original sample. This approach allows different metabolites to be selected in the regression models and reveal their potential importance in each model.

The statistical analysis is performed on this newly created sample using a modeling method of choice. The only requirement for the statistical analysis is that it must perform some form of metabolite selection or be able to indicate the most important ones. The creation of a perturbed data set is repeated many times (e.g., 100 times) and each time the selected metabolites are stored. After the bootstrap procedure, we have 100 models. We can use the information from these models in two ways. First, we can use all 100 prediction errors to give us an idea of prediction variability, this is what the bootstrap is classically used for. Next, from the 100 models, we create an overview of the metabolites that have been selected the most. In follow up research we should devote our attention to the metabolites that have been selected the most. If no stable metabolites can be identified, the conclusion should be that, despite sets of metabolites being reported as predictive, none of them can be used reliably as small perturbations change the selected set.

Cross validation

Where bootstrapping perturbs the data randomly, we can also use a more systematic way for creating different "takes" on the existing data. One such mechanism is cross validation, sometimes called jackknife (Rubingh et al., 2006).

Cross validation is typically used to assess the prediction quality for objects that have not been used in the construction of the model (Hageman et al., 2017; Takahashi et al., 2020; Westerhuis et al., 2008). The rationale behind this is that when models predict truly unknown observations and not just observations used to train the model, we get the best indication of the prediction quality of the model. When data is scarce, we cannot sacrifice part of the data to serve as an independent test set. To solve this problem, we can divide the data into several groups, say k. What we do next is always leave out one group, build the model on the remaining groups and we predict the left-out group of observations. This gets repeated so that every group is left out once and serves as an independent test set to get predicted. The predictions of all left out objects is combined to give an impression of the prediction accuracy. This mechanism also allows us to study the variability between all the different k models with respect to the selected metabolites. Again, metabolites that are selected constantly in most of the models are the most worthwhile (Wehrens et al., 2011). Metabolites selected only

FIGURE 10.3

Schematic overview of fivefold cross validation. In each iteration one fifth, or one fold is left out the model building. Test error (here root mean square error) is calculated using the left out fold. Together with the prediction error, the most important metabolites (indicated by a letter) of each model in each iteration is stored. Metabolites most often encountered are the overall most important ones, in this example a, g and e.

incidentally probably represent accidental correlations and should not be pursued in follow up research. See for a schematic overview of this principle Fig. 10.3.

The number of groups, k, is typically something like 5 or 10, allowing for the creation of 5 or 10 models. It is possible to create as many groups as there are objects. This is referred to as leave-one-out cross validation or jackknife.

Concluding remarks

Typical metabolomics data have inherent statistical difficulties. It contains many more metabolites compared to the number of objects while it is expected that most metabolites are not related or responding to the phenomena under investigation. This poses some modeling difficulties: regular regression techniques do not work unless variables are selected. Variable selection can be difficult to execute correctly. Deselecting uninformative metabolites can be a useful first step. Next, when a small, predefined number of metabolites is required, variable selection using wrapper techniques can provide a solution with a predefined number of metabolites. If the exact number of selected metabolites is not an issue, intrinsic variable selection methods, like lasso or random forests, are good candidates. Techniques like cross validation in concert with stability selection criteria can

provide even more detailed information. Besides an estimate of the prediction quality for unknown observations, it prevents, to a certain extent, the selection of spurious metabolites, focusing on metabolites that have shown repeatedly a relationship to the trait under investigation.

References

Antonelli, J., Claggett, B. L., Henglin, M., Kim, A., Ovsak, G., Kim, N., Deng, K., Rao, K., Tyagi, O., & Watrous, J. D. (2019). Statistical workflow for feature selection in human metabolomics data. *Metabolites*, *9*(7), 143.

Bujak, R., Daghir-Wojtkowiak, E., Kaliszan, R., & Markuszewski, M. J. (2016). PLS-based and regularization-based methods for the selection of relevant variables in non-targeted metabolomics data. *Frontiers in Molecular Biosciences*, *3*, 35.

Determan, C. E., Jr (2015). Optimal algorithm for metabolomics classification and feature selection varies by dataset. *International Journal of Biology*, *7*(1), 100.

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, *26*(1), 51−78.

Doeswijk, T., Smilde, A., Hageman, J., Westerhuis, J., & Van Eeuwijk, F. (2011). On the increase of predictive performance with high-level data fusion. *Analytica Chimica Acta, 705*(1−2), 41−47.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: With applications in R*. Spinger.

Glover, F. (1990). Tabu search: A tutorial. *Interfaces*, *20*(4), 74−94.

Grissa, D., Pétéra, M., Brandolini, M., Napoli, A., Comte, B., & Pujos-Guillot, E. (2016). Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. *Frontiers in Molecular Biosciences*, *3*, 30.

Hageman, J., Wehrens, R., de Gelder, R., Leo Meerts, W., & Buydens, L. (2000). Direct determination of molecular constants from rovibronic spectra with genetic algorithms. *The Journal of Chemical Physics*, *113*(18), 7955−7962.

Hageman, J., Wehrens, R., Van Sprang, H., & Buydens, L. (2003). Hybrid genetic algorithm−tabu search approach for optimising multilayer optical coatings. *Analytica Chimica Acta*, *490*(1−2), 211−222.

Hageman, J., Streppel, M., Wehrens, R., & Buydens, L. (2003). Wavelength selection with Tabu search. *Journal of Chemometrics: A Journal of the Chemometrics Society, 17* (8−9), 427−437.

Hageman, J. A., Hendriks, M. M., Westerhuis, J. A., Van Der Werf, M. J., Berger, R., & Smilde, A. K. (2008). Simplivariate models: Ideas and first examples. *PLoS One*, *3*(9), e3259.

Hageman, J. A., Engel, B., de Vos, R. C. H., Mumm, R., Hall, R. D., Jwanro, H., Crouzillat, D., Spadone, J. C., & van Eeuwijk, F. A. (2017). *Robust and confident predictor selection in metabolomics* (pp. 239−257). Springer International Publishing. Available from https://doi.org/10.1007/978-3-319-45809-0_13.

Hageman, J. A., van den Berg, R. A., Westerhuis, J. A., van der Werf, M. J., & Smilde, A. K. (2008). Genetic algorithm based two-mode clustering of metabolomics data. *Metabolomics*, *4*(2), 141−149.

He, Z., & Yu, W. (2010). Stable feature selection for biomarker discovery. *Computational Biology and Chemistry*, *34*(4), 215–225.

Hendriks, M. M., van Eeuwijk, F. A., Jellema, R. H., Westerhuis, J. A., Reijmers, T. H., Hoefsloot, H. C., & Smilde, A. K. (2011). Data-processing strategies for metabolomics studies. *TrAC Trends in Analytical Chemistry, 30*(10), 1685–1698.

Hur, M., Campbell, A. A., Almeida-de-Macedo, M., Li, L., Ransom, N., Jose, A., Crispin, M., Nikolau, B. J., & Wurtele, E. S. (2013). A global approach to analysis and interpretation of metabolic data for plant natural product discovery. *Natural Product Reports*, *30*(4), 565–583.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*(11), 1–13.

Lindinger, C., Pollien, P., de Vos, R. C., Tikunov, Y., Hageman, J. A., Lambot, C., Fumeaux, R., Voirol-Baliguet, E., & Blank, I. (2009). Identification of ethyl formate as a quality marker of the fermented off-note in coffee by a nontargeted chemometric approach. *Journal of Agricultural and Food Chemistry*, *57*(21), 9972–9978.

Lombardo, V. A., Osorio, S., Borsani, J., Lauxmann, M. A., Bustamante, C. A., Budde, C. O., Andreo, C. S., Lara, M. V., Fernie, A. R., & Drincovich, M. F. (2011). Metabolic profiling during peach fruit development and ripening reveals the metabolic networks that underpin each developmental stage. *Plant Physiology*, *157*(4), 1696–1710.

Madsen, R., Lundstedt, T., & Trygg, J. (2010). Chemometrics in metabolomics—A review in human disease diagnosis. *Analytica Chimica Acta*, *659*(1), 23–33. Available from https://doi.org/10.1016/j.aca.2009.11.042.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1), 1–39.

Rubingh, C. M., Bijlsma, S., Derks, E. P., Bobeldijk, I., Verheij, E. R., Kochhar, S., & Smilde, A. K. (2006). Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics*, *2*(2), 53–61.

Saccenti, E., Westerhuis, J. A., Smilde, A. K., van der Werf, M. J., Hageman, J. A., & Hendriks, M. M. (2011). Simplivariate models: Uncovering the underlying biology in functional genomics data. *PloS One*, *6*(6), e20747.

Saccenti, E., Hoefsloot, H. C., Smilde, A. K., Westerhuis, J. A., & Hendriks, M. M. (2014). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics, 10*(3), 361–374.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517.

Shahrjooihaghighi, A., Frigui, H., Zhang, X., Wei, X., Shi, B., & Trabelsi, A. (2017). An ensemble feature selection method for biomarker discovery. In *2017 IEEE international symposium on signal processing and information technology (ISSPIT)* (pp. 416–421). IEEE.

Takahashi, Y., Ueki, M., Yamada, M., Tamiya, G., Motoike, I. N., Saigusa, D., Sakurai, M., Nagami, F., Ogishima, S., & Koshiba, S. (2020). Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. *Translational Psychiatry*, *10*(1), 1–12.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics*, *7*(1), 1–15.

Vapnik, V. (1998). *The support vector method of function estimation. Nonlinear modeling* (pp. 55–85). Springer.

Weckwerth, W. (2003). Metabolomics in systems biology. *Annual Review of Plant Biology*, *54*(1), 669–689.

Wehrens, R., & Buydens, L. M. (1998). Evolutionary optimisation: A tutorial. *TrAC Trends in Analytical Chemistry*, *17*(4), 193–203.

Wehrens, R., Franceschi, P., Vrhovsek, U., & Mattivi, F. (2011). Stability-based biomarker selection. *Analytica Chimica Acta*, *705*(1–2), 15–23.

Westerhuis, J. A., Hoefsloot, H. C., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J., van Duijnhoven, J. P., & van Dorsten, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics*, *4*(1), 81–89.