Stability Selection

Stability selection

A problem of many variable selection procedures is that noise variables might be erroneously selected. To improve the selection process and to obtain an error control for the number of falsely selected noise variables Meinshausen and Bühlmann (2010) proposed *stability selection*, which was later enhanced by Shah and Samworth (2013).

Stability selection is a versatile approach, which can be combined with all highdimensional variable selection approaches. It is based on sub-sampling and controls the *per-family error rate* $\mathbb{E}(V)$, where V is the number of false positive variables.

Consider a data set with p variables $x_1, \ldots, x_p$ and a response y. Let $S \subseteq \{1, \ldots, p\}$ be the set of *signal* variables, and let $N = \{1, \ldots, p\} \setminus S$ be the set of *noise* variables.

The set of variables that are *selected* by the statistical learning procedure is denoted by $\hat{S}_n \subseteq \{1, \ldots, p\}$. This set $\hat{S}_n$ can be considered to be an estimator of S, based on a data set with n observations.

In short, for stability selection one proceeds as follows:

1. Select a random subset of size $\lfloor n/2 \rfloor$ of the data, where $\lfloor x \rfloor$ denotes the largest integer $\leq x$.
2. Run a *selection procedure* by using the $\lfloor n/2 \rfloor$ observations obtained in step 1 until $q \leq p$ variables are selected. Let $\hat{S}^b_{\lfloor n/2 \rfloor}$ denotes the set of selected variables.

3.Repeat the steps 1. and 2. for $b = 1, \ldots, B$

4. Compute the relative selection frequencies $\hat{\pi}_j = \frac{1}{B} \sum_{b=1}^{B} I\{j \in \hat{S}^b_{\lfloor n/2 \rfloor}\}$ per variable $j = 1, \ldots, B$, where $I\{\cdot\}$ is the indicator function.
5. Select all base-learners that were selected with a frequency of at least π_{thr}, , where π_{thr} is a pre-specified threshold value. Thus, we obtain a set of stable variables $\hat{S}_{\text{stable}} = \{j : \hat{\pi}_j \geq \pi_{\text{thr}}\}$.

Meinshausen and Bühlmann (2010) show that this selection procedure controls the *per-family error rate* PFER= $\mathbb{E}(V)$. In general it holds that FWER $\leq$ PFER, thus, for a fixed significance level α it holds that PFER-control is more conservative than FWER-control.

An upper bound is given by

$$\mathbb{E}(V) \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

where q is the number of selected variables per run, p is the number of (possible) variables and π_{thr} is the threshold for selection probability. The theory requires two assumptions to ensure that the error bound holds:

(i) The distribution $\{I_{j \in \hat{S}_{\text{stable}}}, j \in N\}$ needs to be *exchangeable* for all noise variables N

(ii) The original selection procedure must not be worse than random guessing

In practice, assumption (i) essentially means that each noise variable has the same selection probability. Thus, all noise variables should, for example, have the same correlation with the signal variables (and the outcome). For examples of situations where exchangeability is given see Meinshausen and Bühlmann.

Assumption (ii) means that signal variables should be selected with higher probability than noise variables. This assumption is usually not very restrictive as we wouuld expect it to hold for any sensible selection procedure.

Choice of parameters

The stability selection procedure mainly depends on two parameters: the number of selected variables per run q and the threshold value for stable variables π_{thr}.

Meinshausen and Bühlmann (2010) propose to chose $\pi_{\text{thr}} \in (0.6, 0.9)$ and claim that the threshold has little influence on the selection procedure. In general, any value $\in (0.5, 1)$ is potentially acceptable, i.e. a variable should be selected in more than half of the fitted models in order to be considered stable.

The number of selected variables q should be chosen so high that in theory all signal variables S can be chosen. If q was too small, one would inevitably select only a small subset of the signal variables S in the set $\hat{S}_{\text{stable}}$ as $\#\hat{S}_{\text{stable}} \leq \#\hat{S}^b_{\lfloor n/2 \rfloor} = q$ (if $\pi_{\text{thr}} > 0.5$).

The choice of the number of subsamples B is of minor importance as long as it is large enough. Meinshausen and Bühlmann (2010) propose to use $B = 100$ replicates, which seems to be sufficient for an accurate estimation of $\hat{\pi}_j$ in most situations.

In general, we would recommend to choose an upper bound PFER_{max} for PFER and specify either q or π_{thr}, preferably q. The missing parameter can then be computed from

$$\text{PFER}_{\text{max}} = \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

For a fixed value q, we can easily vary the desired error bound PFER_{max} by varying the threshold π_{thr} accordingly. As we do not need to re-run the subsampling procedure, this is very easy and fast. In a second step, one should check that the computed value is sensible, i.e. that $\pi_{\text{thr}} \in (0.5, 1)$, or that q is not too small, or that PFER_{max} is not too small or too large. Note that the PFER can be greater than one as it resembles the tolerable expected number of falsely selected noise variables.

The size of the subsamples is no tuning parameter but should always be chosen to be $\lfloor n/2 \rfloor$. This an essential requirement for the derivation of the error bound. Other (larger) subsample sizes would theoretically be possible but would require the derivation of a different error bound for that situation.

One should keep in mind that stability selection controls the per-family error rate, which is very conservative. Specifying the error rate such that $\alpha \leq \text{PFER}_{\text{max}} \leq m\alpha$, with significance level α and m hypothesis tests, might provide a good idea for a sensible error control in high-dimensional settings with FWER control ($\text{PFER}_{\text{max}} = \alpha$) and no multiplicity adjustment ($\text{PFER}_{\text{max}} = m\alpha$) as the extreme cases.

Furthermore, prediction models might not always benefit from stability selection. If the error control is tight, i.e. PFER_{max} is small, the true positive rate is usually smaller than in a cross-validated prediction model without stability selection and the prediction accuracy suffers. Prediction and variable selection are two different goals.