Discussion of 'Stability Selection', by Nicolai Meinshausen and Peter Bühlmann

John Shawe-Taylor Shiliang Sun
Department of Computer Science
University College London

February, 2010

We congratulate the authors on a paper with an exciting mix of novel theoretical insights and practical experimental testing and verification of the ideas. We provide a personal view of the developments introduced by the paper, mentioning some areas where further work might be usefully undertaken, before presenting some results assessing the generalisation performance of stability selection on a medical dataset.

The paper introduces a general method for assessing the reliability of including component features in a model. They independently follow a similar line to that proposed by Bach (2008), in which the author proposes to run the Lasso algorithm using boostrap samples and only include features that occur in all of the models thus created. Meinshausen and Bühlmann refine this idea by assessing the probability that a feature is included in models created with random subsets of $\lfloor n/2 \rfloor$ training examples. Features are included if this probability exceeds a threshold π_{thr}.

Theorem 1 provides a theoretical bound on the expected number of falsely selected variables in terms of π_{thr} and q_Λ the expected number of features to be included in the models for a fixed subset of the training data, but range of values of the regularisation parameter $\lambda \in \Lambda$. The theorem is quite general, but makes one non-trivial assumption: that the distribution over the inclusion of false variables is exchangeable. In their evaluation of this bound on a range of real-world training sets, albeit with artificial regression functions, they demonstrate a remarkable agreement between the bound value (chosen to equal 2.5) and the true number of falsely included variables.

We would have liked to have seen further assessment of the reliability of

the bound in different regimes, that is bound values as fixed by different q_Λ and π_{thr}. The experimental results indicate that in the datasets considered the exchangeability assumption either holds, or if it fails to hold, does not adversely affect the quality of the bound. We believe that it would have been useful to explore in greater detail which of these explanations is more probable.

One relatively minor misfit between the theory and practical experiments was the fact that the theoretical results are in terms of the expected value of the quantities over random subsets, while in practice a small sample is used to estimate the features to include as well as quantities such as q_Λ. Perhaps finite sample methods for estimating fit with the assumption of exchangeability could also be considered. This might lead to an online approach where samples are generated until the required accuracy is achieved.

Theorem 2 provides a more refined analysis in that it also provides guarantees that relevant examples are included provided they play a significant part in the true model, something that Theorem 1 does not address. Though stability selection as defined refers to the use of random subsampling and all the experiments make use of this strategy, Theorem 2 analyses the effect of a 'randomised Lasso' algorithm that randomly rescales the features before training on the full set. Furthermore, the proof of Theorem 2 does not make it easy for the reader to gain an intuitive understanding of the key ideas behind the result.

Our final suggestion for further elucidation of the usefulness of the ideas presented in the paper is to look at the effects of stability selection on the generalisation performance of the resulting models. As an example we have applied the approach to a dataset concerned with predicting the cholesterol level of subjects based on risk factors and SNP genotype features.

The data set includes 1842 subjects/examples. The feature set (input) includes six risk factors (age, smo, bmi, apob, apoa, hdl) and 787 genotypes. Each genotype takes a value in $\{1, 2, 3\}$. As preprocessing, each risk factor is normalized to have mean 0 and variance 1. For each example, its output is the averaged cholesterol level over five successive years. The whole data were divided into a training set of 1200 examples and a test set of the remaining 642 examples. We will report the test performance averaged across ten different random divisions of training and test sets. The performance is evaluated through RMSE (root mean square error). In addition to standard 'stability selection' we report performance for a variant in which complementary pairs of subsets are used as proposed in the discussion of Shah and Samworth.

We report results for four methods: (1) Ridge regression with the original features (M1); (2) Lasso with the original features (M2); (3) Ridge regression with the features identified by stability selection (M3); (4) Lasso with the features identified by stability selection (M4). The variants of M3 and M4 based on complementary pairs of subsets are denoted (M3c) and (M4c). The performance of the first two methods are independent of π_{thr} and provide a baseline given in Table 1.

	M1	M2
RMSE	0.752 (0.017)	0.707 (0.017)
# retained features	792 (0.66)	109 (5.22)

Table 1: Mean (standard deviation) of the test performance and number of retained features for methods M1 and M2.

For the two methods involving stability selection we experiment with values of π_{thr} from the set $\{0.2, 0, 25, 0.3, 0.35, 0.4, 0.45, 0.5\}$. The results for different values of π_{thr} for the methods M3 and M4 using standard subsampling and randomised Lasso are given in Table 2, while using the complementary sampling gives the results of Table 3.

π_{thr}	# features	M3	M4
0.20	117.4 (6.2)	0.722 (0.017)	0.716 (0.017)
0.25	86.8 (5.2)	0.720 (0.016)	0.715 (0.016)
0.30	64.7 (4.1)	0.719 (0.017)	0.715 (0.017)
0.35	45.3 (4.1)	0.716 (0.016)	0.715 (0.017)
0.40	27.3 (3.8)	0.714 (0.016)	0.713 (0.016)
0.45	17.7 (1.9)	0.712 (0.016)	0.710 (0.016)
0.50	11.4 (1.6)	0.714 (0.019)	0.713 (0.019)

Table 2: Mean (standard deviation) of the test performance and number of retained features for methods M3 and M4.

The results suggest that the stability selection has not improved the generalisation ability of the resulting regressors, though clearly the Lasso methods outperform Ridge regression. The performance is remarkably stable across different values of π_{thr} despite the number of stable variables undergoing an order of magnitude reduction.

π_{thr}	# features	M3c	M4c
0.20	116.5 (4.4)	0.721 (0.017)	0.715 (0.017)
0.25	83.4 (3.0)	0.720 (0.017)	0.715 (0.017)
0.30	62.4 (3.6)	0.718 (0.017)	0.714 (0.016)
0.35	44.2 (3.2)	0.717 (0.015)	0.716 (0.016)
0.40	27.4 (3.4)	0.714 (0.015)	0.713 (0.015)
0.45	18.2 (1.7)	0.714 (0.012)	0.712 (0.013)
0.50	11.8 (1.8)	0.715 (0.014)	0.714 (0.014)

Table 3: Mean (standard deviation) of the test performance and number of retained features for methods M3c and M4c.