Maximal Subset Feature Selection for BioInformatics

Dean P. Foster and Lyle H. Ungar
Statistics and Computer and Information Science Departments
University of Pennsylvania
Philadelphia, PA 19104, USA
foster@wharton.upenn.edu and ungar@cis.upenn.edu

Feature selection methods generally try to select the minimal set of features needed to give good predictive accuracy in a model, for example by minimizing training error plus a penalty on either the number of features (e.g., using AIC, BIC, RIC) or the L_1 norm of the features (LARS/Lasso), sometimes combined with an L_2 penalty (elastic net). For many problems in biology, however, it is more important to find as many features as possible that are significantly predictive, rather than to select a minimal subset. Example applications include selecting among SNPs in a genome wide assay to determine which mutations may lead to increased danger of mental disease, or selecting genes of different yeast strains based on their expression levels with the goal of identifying which genes effect the growth rate of the yeasts. When the goal is science – e.g., finding which mutations "cause" a given disease – rather than prediction, the objective is not predictive accuracy, but to maximize the number of true features identified while minimizing the number of false positives. More precisely, we seek to maximize the number of features selected subject to a constraint on the fraction of the features selected which are false positives – a sort of false discovery rate (FDR).

Finding the maximal feature set poses some difficulties not present in predictive modeling, since the problem is non-falsifiable; There is no way to determine based only on the training data if the correct causal features have been identified. If several features are highly correlated in the training data, as neighboring SNPs on a chromosome are, one cannot tell which of the correlated features is causal. (Note that we do not assume that future features will maintain the same correlation structure as the training data.) Thus, to test a maximal feature set selection method, one must use synthetic data where the features are known – or do experiments where one changes features and observes the consequences. More formally, We assume that data is generated by an unknown linear model $y = \sum w_i x_i$, where the sum is over some subset of a much larger number of features X. If a true feature is perfectly correlated with a "false" feature, it is impossible to determine which of the two is the "causal" or correct feature.

It is common practice in biology to find candidate causal features using a univariate screen, in which the correlation of each feature with a response (the label) is calculated, features are sorted from most to least correlated with the response, and those features whose correlation is above some threshold are taken to significant, e.g. the probability of the correlation arising by chance (the p-value) is compared to Bonferroni corrected threshold. A variety of extensions ranging from the Simes procedure to more modern FDR step-up and step-down procedures (e.g., Benjamini and Hochberg) give more power by sequentially adjusting the threshold, but all still look at univariate correlations, where each x is compared individually to y.

Such univariate screening methods can miss features because one feature can mask another. Consider the case where $y = c_1 x_1 + c_2 x_2$ with $c_1 \gg c_1$ and x_1 and x_2 of the same size. In this case, $c_1 x_1$ effectively serves as noise to the univariate correlation of x_2 with y, masking x_2 and making it difficult to discover, even though it would be shown to be significant if a multivariate regression were done. Naive FDR methods test the hypothesis that each individual feature is correlated with y, and so will miss masked features.

Regression methods such as LARS and elastic net attempt to address the masking problem. However, for the biological data sets we have looked at, the elastic net does relatively poorly at identifying maximal ffeature sets in part because setting the strength of the L_1 and L_2 penalties to minimize the cross validation prediction error fails to to maximize the score of interest: the number of features selected subject to a bound on the false discovery rate.

How might one, if computational power were no object, estimate the score, since it is unmeasurable on real data sets? Assume, for simplicity, that we know the number, q, of causal features. We could then compute all possible subsets of features of size q such that all features in each subset are significant (in that set) after a Bonferroni correction and none of the left out features are significant. Counting how often each feature is selected gives a measure of how likely it is to be causal. We prove that if all the features of the correct model are statistically significant, then the minimax estimate of the probabilities of a variable being included are those given by this "subset counting" algorithm. One can, of course pick a threshold on the probability to select features at a desired precision/recall tradeoff.

Table 1 compares the performance of the above "subset counting" method with several competing methods on a synthetic data set, and illustrates many interesting points. The first column gives the L_2 norm of the difference between the estimated probability of each feature being in the model and the correct answer (1 or 0, respectively, if the feature is nonzero or not). Almost all feature selection methods assume a feature is either in a model (with probability one) or not, in which case the L_2 norm is just the number of true positives missed plus the number of false positives (i.e., the error rate). We also compare against single pass screening, where all features that have a p-value for the correlation with y of less than $0.5/p$ (where p is the number of features, the standard Bonferroni correction) are selected. Multi-pass screening, where we repeatedly test for the correlation of features with the residuals resulting from subtracting off the effect of the previously selected features, shows how important masking can be, even when feature weights are all the same size. By way of contrast, using the Simes procedure on univariate correlations gives only a small benefit in finding more true features. Elastic net adds too many false features, in spite (or because) of the regularization parameters being chosen by cross validation. Choosing the L_1 and L_2 penalties in elastic net to minimize prediction error leads to it adding far more false positives than is desirable for minimizing error in selecting true features. An interesting research question is how to modify the elastic net to do a better job of maximal feature set selection. Zeroing small coefficient values (those less than 0.001) greatly improves the performance of elastic net (to Error = 12.2), suggesting that significant improvements can be made.

Error	TP	FP	Train Err	Test Err	Method
0	10	0	0.00136	0.00338	true model
10	0	0	0.29046	0.02236	null model
190	10	190	0.00013	0.01992	full model
8.8	1.6	0.4	0.03893	0.02126	screening
4.7	8.0	2.7	0.00432	0.00919	multipass screening
11.1	2.2	3.3	0.03490	0.02296	Simes
9.6	3.4	4.4	0.01616	0.02093	subset counting
24.1	9.1	23.2	0.00149	0.01219	elastic net

Table 1. L_2 Error in the probability of features being in model (Error), True positives (TP), False positives (FP), Training error (Train Err), Testing error (Test Err) for each method. In all methods, the confidence thresholds are set at 0.5 (before Bonferroni correction). 40 observations, 200 features, 10 true features, all with equal sized weights. Results are averaged over 100 randomly generated data sets.

We believe that explicitly designing algorithms to search for maximal feature sets is important and practical. We show that in spite of the nonidentifiability of the maximal feature set problem, one can formally characterize the performance of some of these methods. We are applying maximal feature selection to the large data sets that arise in genome wide association (GWA) studies. On these large data sets (tens to hundreds of thousands of features), neither LARS nor the subset regression and counting methods are feasible, but they can be well approximated using multipass methods.