

Streaming Feature Selection

Lyle Ungar (ungar@cis.upenn.edu)

Computer and Information Science

Dean Foster and Bob Stine(foster, stine@wharton.upenn.edu)

Statistics Department

University of Pennsylvania

Philadelphia, PA 19104

When learning predictive models that may require testing hundred of thousands of features in order to find tens of significant features, it is often desirable to interleave feature generation with feature selection. New features (e.g. interaction terms) can then be generated lazily based on which features have already proven significant. We address three issues: 1) when feature selection is preferable to regularization or other smoothing methods, 2) what form of complexity penalty to use to decide when to add a variable to the model and 3) the relative performance of streaming vs. stepwise feature selection.

When only a handful out of thousands of potential features are expected to be significant in a model, methods which try to smoothly combine all of the features are doomed to poor performance [?]. Testing all possible subsets of the features is also clearly intractable. The standard approach in such circumstances has been stepwise addition of features to the model. Stepwise methods try *all* potential features at each iteration of the algorithm. We propose that streaming methods, in which each feature is considered only once, are competitive in accuracy, and allow dynamic selection of which features to consider for model inclusion.

Careful feature selection is critical in either stepwise or streaming feature selection. Most current methods for selecting features for inclusion in a model break down when the number of features to be considered for inclusion in a model, p , becomes very large. Inclusion rules which are not a function of p (e.g. AIC or BIC) inevitably overfit as p becomes large (with n fixed) and inevitably underfit when n becomes large (with p fixed). Hence it is important to have rules such as Bonferonni or RIC that depend on p . However, even these rules must be used carefully since they often underfit [?]. If both p and the correct number of variables q in the true model were known a variant of an adaptive rule we call *eBIC*, gives good performance. Of course, in a streaming setting neither p nor q are known, but we do know at each point how many features have been considered and how many added to the model. This allows us to create a streaming version of eBIC.

The different penalty methods and their associated penalties are given in the following table:

	<i>p</i> -independent	<i>p</i> -dependent	adaptive
streaming	streaming AIC, BIC	α -spending	SFS
stepwise	standard AIC, BIC	RIC	eBIC
penalty	$2, \log(n)$	$2\log(p)$	$2\log(p/q)$

Table 1: Types of feature selection. n is the number of observations, p the number of features, and q the number of features selected for model inclusion

The Streaming Feature Selection (SFS) algorithm proposed in this talk is a streaming version of eBIC which considers a stream of potential features for model inclusion, and incrementally adjusts the criterion for including features in the model based on the success in including the features seen previously in the stream. We show, using an information theoretic argument, that SFS is guaranteed never to overfit, and empirically compare its performance against more standard methods such as stepwise regression using a BIC-based penalty. In streaming regression, the order in which features are tested matters; it is desirable to consider better features earlier. Fortunately, simple heuristics are often available to order features: one can consider individual terms before interactions, unigrams before bigrams, or features resulting from simple logic or SQL queries before those resulting from longer, more complex queries.

We empirically compare streaming and stepwise regression for both constant (BIC) and adaptive (eBIC) complexity penalties. For a simple synthetic data set containing 100 observations of 1,000 independent predictors, where y is a linear combination of three of the predictors plus a small amount of noise, Stepwise AIC and BIC both select models with 100 predictors (perfect fit on the training data), while SFS finds the correct model regardless of whether the three correct variables are at the start or the end of the stream. We also examine a real data set of 100 observations in which we predict venue in which a document was published based on 10,000 features derived using a structural relational learning (SRL)-style search over the relations between documents contained in the CiteSeer data base[?].

We find that SFS, in spite of making only a single pass through the features, is competitive with stepwise regression, which repeatedly considers all variables for inclusion in the model. SFS performs extremely well when the good variables are encountered relatively early (e.g., in the first 1000 features), and suffers little loss when the good features are distributed randomly throughout the feature stream. As predicted by theory, non-adaptive penalty methods such as BIC perform extremely poorly in regimes where the number of features that should be in the model, q is very different than p/n^2 .

References

- D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.
- D. P. Foster and R. A. Stine. Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association (JASA)*, 2004. in press.
- A. Popescul, L. H. Ungar, S. Lawrence, and D. M. Pennock. Statistical relational learning for document mining. In *Proc. of IEEE International Conference on Data Mining (ICDM-2003)*, 2003.