Dimension Reduction Using the Rule Ensemble Machine Learning Method

Orianna DeMasi*, Juan Meza+, David Bailey*
*Lawrence Berkeley National Laboratory, +UC Merced
DOE Applied Mathematics Program Meeting
October 2011

Motivation

Ensemble methods for supervised machine learning have become popular due to their ability to accurately predict class labels with groups of simple "base learners." Many ensemble methods are computationally efficient, but offer little insight into the structure of a dataset. We consider an ensemble technique that returns a model of ranked rules. The model accurately predicts class labels and has the advantage of indicating which parameter constraints are most useful for predicting those labels. An example is given with a dataset containing images of potential supernovas where the number of necessary features is reduced from 39 to 21.

Problem Formulation

Suppose we are given a dataset $S = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ where the label y_i is either +1 or -1 and $\mathbf{x}_i$ is a vector $(x_1, \ldots, x_K)$ of K features. We want to predict which class and unlabeled observation $\mathbf{x}$ came from.

We use the rule ensemble method developed by Friedman and Popescu (2003, 2004, 2005) to construct a function $F(\mathbf{x})$ such that $sign(F(\mathbf{x}))$ predicts the true label y.

We assume $F(\mathbf{x})$ is a linear combination of base learners f_k

$$F(\mathbf{x}; \mathbf{a}) = a_0 + \sum_{k=1}^{K} a_k f_k(\mathbf{x})$$

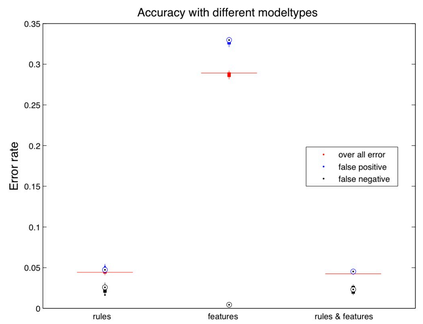
and approximate the coefficients $\mathbf{a}$ by minimizing the risk of using $F(\mathbf{x})$ on the sample of training data S.

The $l_1$lasso penalty is added to the minimization problem to make $\mathbf{a}$ sparse and prevent overfitting the training sample. A sparse solution returns a model with fewer terms that is easier to interpret. As $\mathbf{a}$ is a vector of weights for each rule, we can also interpret which rules and coefficients are most influential.

Application: Supernova

Dataset is a collection of images
Positive labels indicate that the picture is of a supernova
Dataset: 5,000 positive & 20,000 negative observations
Training set: 2,500 positive & 2,500 negative observations
Testing set: all observations not in the training set
10 cross validation tests

Rules as Base Learners



Using rules as terms was six times more accurate than only using features as linear terms, a classical multiple linear regression of the labels on the attributes.

- The Rule Ensemble Method uses rules $r_k(\mathbf{x}_i)$ as base learners $f_k(\mathbf{x}_i)$
- Each rule is a node in a decision tree

$$r_k(\mathbf{x}_i) = \prod_j I(x_{ij} \in p_{kj})$$

- Rules define hypercubes in parameter space and evaluate to 1 if an observation is in that cube, i.e. an observation's jth attribute abides by the rule's constraint p_j
- Rules are diverse, quick to build, and allow interactions

Calculation of Coefficients

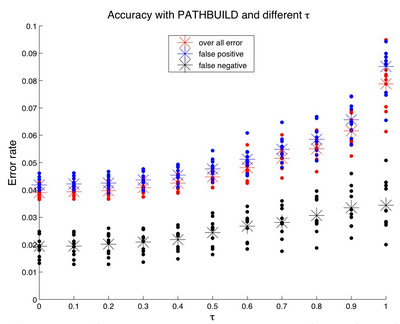
PATHBUILD:
(Friedman & Popescu 2004)
Uses a constrained gradient descent method to approximate $\mathbf{a}$ by solving

$$\min_{\{\mathbf{a}\}} \sum_{i=1}^{N} L\left(a_0 + \sum_{k=1}^{K} a_k f_k(\mathbf{x}_i), y_i\right) + \lambda \sum_{k=1}^{K} |a_k|$$

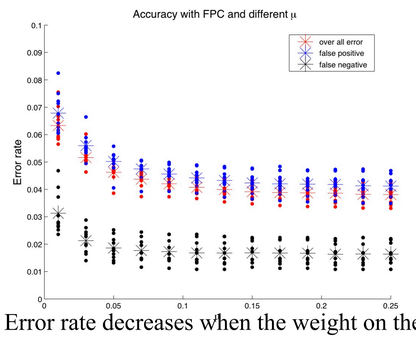
where L() is ramp loss error. Initializes all coefficients to zero and at each iteration updates only coefficients k that have large enough gradient

$$\{k : |g_k(\mathbf{X}; \mathbf{a}^\ell)| \geq \tau * ||g_k(\mathbf{X}; \mathbf{a}^\ell)||_\infty\}.$$

Constraint parameter τ is in [0, 1].



Error rate increases as τ increases and restricts the number of coordinates that PATHBUILD advances in at each iteration.



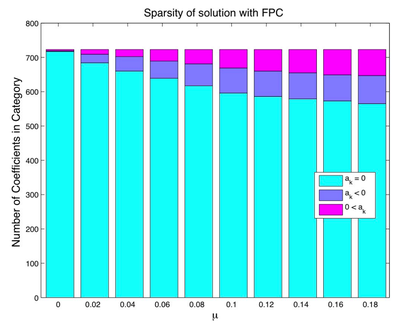
Error rate decreases when the weight on the risk is increased (μ increased).

Fixed Point Continuation:
(Hale, Yin & Zhang 2007)
Uses a shrinkage operator to approximate $\mathbf{a}$ by solving

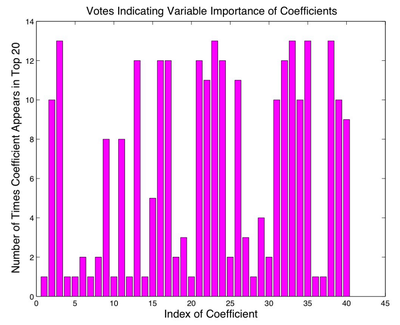
$$\min_{\{\mathbf{a}\}} \frac{\mu}{2} * \sum_{i=1}^{N} \left(a_0 + \sum_{k=1}^{K} a_k f_k(\mathbf{x}_i) - y_i\right)^2 + \sum_{k=1}^{K} |a_k|.$$

This problem is a reformulation of the l_1-penalized regression problem. It uses the l_2 norm for a loss function and the sparsity is controlled by μ > 0

Rule Importance



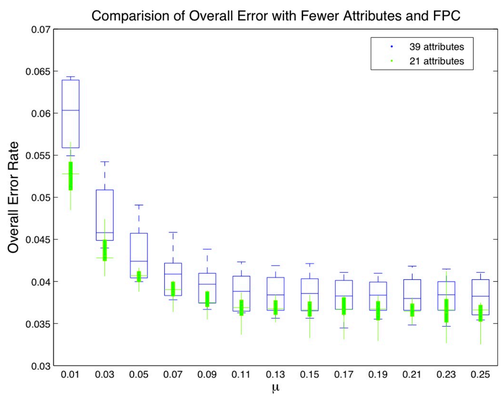
As μ is increases, **a** becomes less sparse and more terms are included in the model. The sparsity of the solution stops decreasing when μ is large. Here 78% of the coefficients are trivial when μ = 0.19.



Solutions were generated at 13 different values of μ. Rules that received 13 votes were one of the 20 most influential rules for every value of μ tried. Only rules that received at least one vote are shown. The attributes used to compose these rules were used to find a smaller subset of attributes to train on.

Reduce 39 to 21 features

Using the magnitude of a coefficient to indicate the importance of a rule, the rule ensemble method indicated that 21 out of the original 39 attributes were more influential than the rest. Retraining and testing with the restricted set of attributes gave a lower overall error rate which indicates that the rule ensemble method successfully identified important attributes in the dataset.



Preliminary tests used all 39 attributes in the dataset. Then new tests were run using a subset of 21 attributes that the rule ensemble method had selected as being most influential.

We would like to thank Sean Peisert and Peter Nugent for their extremely valuable comments and suggestions.
This research was supported in part by the Director, Office of Computational and Technology Research, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy, under contract number DE-AC02-05CH11231