Review on boosting algorithms

Vu Tuan Hung and Do Quoc Khanh

8 mai 2012

1 Introduction

2 Two-class classification

In this first part, we present an overview on boosting methods in the two-class classification framework. From a *training* set $(\mathbf{x}_i, y_i)_{i=1,...,N}$ in which $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, we try to construct a function $F : \mathcal{X} \to \mathcal{Y}$ so that when a new value $\mathbf{x}$ is randomly introduced, we have the highest probability to predict correctly the value y corresponding to this value of $\mathbf{x}$. Formally, we want to minimize the probability :

$$\mathbb{P}_{(\mathbf{x},y)} \left(y \neq F(\mathbf{x}) \right)$$

The variable $\mathbf{x}$ is called explanatory variables ($\mathbf{x}$ may be multi-variational) and y is called response variable. In the two-class classification framework, $\mathcal{Y} = \{-1, 1\}$.

2.1 Boosting and optimization in function space

We exploit the point of view presented in [2] by considering this problem as an estimation and optimization in function space. Indeed, if there exists a function F^* which minimizes the above error :

$$\begin{aligned} F^* &= arg \min_F \mathbb{P}_{(\mathbf{x},y)}(y \neq F(\mathbf{x})) \\ &= arg \min_F \mathbb{E}_{(\mathbf{x},y)} \left[1(y \neq F(\mathbf{x})) \right] \end{aligned}$$

then we are trying to estimate F^* by a function $\hat{F}$ through the training set $(\mathbf{x}_i, y_i)_{i=1,...,N}$.

Base classifiers : An approach frequently employed by classification algorithms is to suppose F^* belongs to a function class parameterized by $\theta \in \Theta$:

$$F^* \in \mathcal{Q} = \{F(., \theta) | \theta \in \Theta\}$$

so that the problem of estimating F^* becomes an optimization of the parameters on Θ :

$$\hat{\theta} = arg \min_{\theta \in \Theta} \mathbb{E}_{(\mathbf{x},y)} \left[1(y \neq F(\mathbf{x}, \theta)) \right]$$

and then we will take $\hat{F} = F(., \hat{\theta}) \in \mathcal{Q}$. For example, with regression tree algorithms, we have :

$$\mathcal{Q} = \left\{ F(x, \theta) = \sum_{k=1}^{K} \lambda_k 1(\mathbf{x} \in R_k) | (\lambda_1,...,\lambda_K) \in \mathbb{R}^K, (R_1,...,R_K) \in \mathcal{P}_{\mathcal{X}} \right\}$$

in which $\theta = (\lambda_{1:K}, R_{1:K})$ and $\mathcal{P}_{\mathcal{X}}$ is the set of all partitions of $\mathcal{X}$ into K disjoint subsets by hyperplans which are orthogonal to axes. Similarly for support vector machines, K disjoint subsets R_1,...,R_K are divided by hyperplans in the reproducing kernel Hilbert space of $\mathcal{X}$ corresponding to some kernel.

We can see that a classifier is characterized by its function sub-space $\mathcal{Q}$ and the corresponding parameter space. Having the base classifiers $\mathcal{Q}_{1:M}$ with parameter spaces $\Theta_{1:M}$, instead of considering each of these classifiers separately, boosting methods consider functions of the following additive form :

$$\hat{F} \in \mathcal{F}_{\mathcal{Q}_1,\ldots,\mathcal{Q}_M} = \left\{ \sum_{m=1}^{M} \beta_m F(., \theta_m) | \theta_m \in \Theta_m, \forall m = 1,\ldots,M \right\}$$

so that the optimization problem becomes :

$$\left\{ \hat{\beta}_{1:M}, \hat{\theta}_{1:M} \right\} = \tag{1}$$

The paper explains boosting . Advantage

2.2 One-degree optimization

2.3 Two-degree optimization

3 Multi-class classification and some generalizations

3.1 A traditional approach

3.2 Some generalization of two-class algorithms

3.3 Other generalizations

4 Experiments

4.1 Experiments with simulated data

4.2 Experiments with real data

5 Conclusion

Références

[1] Friedman, J., Hastie, T. & Tibshirani, R. *Additive Logistic Regression : a Statistical View of Boosting*, 2000.

[2] Friedman, J. *Greedy Function Approximation : A Gradient Boosting Machine*, IMS 1999 Reitz Lecture, 2001.

[3] Schapire, R.E. & Singer, Y. *Improved Boosting Algorithms : Using Confidence-rated Predictions*, 1998.