

Разработка эффективной реализации методов, основанных на поиске оптимального баланса между дивергенцией и точностью аппроксимации

Д. Д. Аристархов

ВМК МГУ

15 ноября 2024 г.

Постановка задачи

Дана выборка $S = (X_1, y_1), \dots, (X_m, y_m)$, где X_i — вектор признакового описания объекта, y_i — метка объекта. Рассматривается задача регрессии: $X_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$. Требуется построить ансамбль базовых алгоритмов $A_1(X), \dots, A_k(X)$, предсказывающих значения метки по вектору признаков.

Дисперсия ансамбля

Рассматриваем ошибку ансамбля $\mathcal{A}(x) = \frac{1}{k} \sum_i^k A_i(x)$

Разложение ошибки ансамбля

$$\begin{aligned} L(\mathcal{A}) = & \underbrace{\mathbb{E}_{x,y} \left[(y - \mathbb{E}[y|x])^2 \right]}_{\text{Шум}} + \underbrace{\mathbb{E}_{x,y} \left[(\mathbb{E}_X [A_i(X)(x)] - \mathbb{E}[y|x])^2 \right]}_{\text{Смещение}} \\ & + \underbrace{\frac{1}{k} \mathbb{E}_{x,y} \left[\mathbb{E}_X \left[\left(A_i(X)(x) - \mathbb{E}_X [A_i(X)(x)] \right)^2 \right] \right]}_{\text{Дисперсия } A_i} \\ & + \underbrace{\frac{k(k-1)}{k^2} \mathbb{E}_{x,y} \left[\mathbb{E}_X \left[\left(A_i(X)(x) - \mathbb{E}_X [A_i(X)(x)] \right) \left(A_j(X)(x) - \mathbb{E}_X [A_j(X)(x)] \right) \right] \right]}_{\text{Ковариация } A_i, A_j} \end{aligned}$$

Дисперсия ансамбля

- Шум — свойство выборки, не зависит от модели
- Смещение — равно смещению базового алгоритма, поэтому берем базовые алгоритмы с маленьким смещением, например, глубокие деревья
- Дисперсия A_i — уменьшается в k раз при увеличении количества базовых алгоритмов
- Ковариация A_i, A_j — ?

Дисперсия ансамбля

Для уменьшения ковариации используются следующие подходы:

- Бэггинг. Каждый алгоритм обучается на случайной подвыборке, сгенерированной из выборки с помощью бутстрэпа, т.е. выбираются m объектов с возвращениями. Таким образом, в одной выборке некоторые объекты встретятся несколько раз, а некоторые — ни разу.
- Рандомизация признаков. При построении очередного дерева в каждой вершине выбор наилучшего признака для разбиения происходит не из всех возможных признаков, а из случайно выбранной подвыборки.

Предлагаемый метод

Обозначим $L_k(X) = \frac{1}{k} \sum_{i=1}^k A_i(X)$, $Q_k(X) = \frac{1}{k} \sum_{i=1}^k A_i^2(X)$. Введем критерий, представляющие собой среднеквадратичную ошибку алгоритма и дисперсию прогнозов вычисляемых алгоритмов:

Критерий Φ_E

$$\Phi_E(A_1(X), \dots, A_k(X)) = \frac{1}{mk} \sum_{i=1}^k \sum_{j=1}^m (y_j - A_i(X_j))^2$$

Критерий Φ_V

$$\Phi_V(A_1(X), \dots, A_k(X)) = \frac{1}{mk} \sum_{i=1}^k \sum_{j=1}^m (L_k(X_j) - A_i(X_j))^2$$

Предлагаемый метод

При построении ансамбля предлагается предлагается явно минимизировать Φ_E и максимизировать Φ_V . Данная задача может быть сведена к минимизации Φ_G :

Критерий Φ_G

$$\Phi_G = (1 - \mu)\Phi_E - \mu\Phi_V,$$

где $\mu \in [0, 1]$ является гиперпараметром, определяющим соотношение точности и разнородности прогнозов отдельных деревьев.

Поскольку каждое дерево в ансамбле строится отдельно от других, необходимо получить критерий для построения очередного дерева. Обозначим через D_E^k и D_V^k изменение функционалов Φ_E и Φ_V при включении в ансамбль дополнительного алгоритма A_{k+1} .

Предлагаемый метод

Критерий D_E^k

$$\begin{aligned} D_E^k &= \Phi_E(A_1(X), \dots, A_{k+1}(X)) - \Phi_E(A_1(X), \dots, A_k(X)) \\ &= \frac{1}{m(k+1)} \sum_{j=1}^m (y_j - A_{k+1}(X_j))^2 - C_E, \end{aligned}$$

где C_E не зависит от $A_{k+1}(X)$.

Критерий D_V^k

$$\begin{aligned} D_V^k &= \Phi_V(A_1(X), \dots, A_{k+1}(X)) - \Phi_V(A_1(X), \dots, A_k(X)) \\ &= \frac{k}{m(k+1)^2} \sum_{j=1}^m (A_{k+1}^2(X_j) - 2L_k(X_j)A_{k+1}(X_j)) + C_V, \end{aligned}$$

где C_V не зависит от $A_{k+1}(X)$.

Предлагаемый метод

Объединяя эти выражения, получаем функционал, который необходимо минимизировать при построении очередного дерева $A_{k+1}(X)$:

Критерий D_G^k

$$\begin{aligned} D_G^k &= (1 - \mu)D_E^k - \mu D_V^k = \\ &= \frac{1 - \mu}{m(k + 1)} \sum_{j=1}^m (y_j - A_{k+1}(X_j))^2 \\ &\quad - \frac{\mu k}{m(k + 1)^2} \sum_{j=1}^m (A_{k+1}^2(X_j) - 2L_k(X_j)A_{k+1}(X_j)) + C_G, \end{aligned} \tag{1}$$

где C_G не зависит от $A_{k+1}(X)$.

Предлагаемый метод

Теперь рассмотрим вопрос оптимального значения в листе дерева $A_{k+1}(X)$. Пусть в лист попали объекты $(X_{n_1}, y_{n_1}), \dots, (X_{n_p}, y_{n_p})$. В листе алгоритм предсказывает одно значение для всех объектов, попавших в этот лист: $A_{k+1}(X_{n_j}) \equiv \tilde{A}$, $j = \overline{1, p}$. Найдем производную функционала (1) относительно прогноза \tilde{A} :

$$\begin{aligned}\frac{\partial D_G^k}{\partial \tilde{A}} &= \frac{2(1-\mu)}{p(k+1)} \sum_{j=1}^p (\tilde{A} - y_{n_j}) - \frac{2\mu k}{p(k+1)^2} \sum_{j=1}^p (\tilde{A} - L_k(X_{n_j})) \\ &= \frac{2}{p(k+1)} \sum_{j=1}^p \left(\left(1 - \mu \frac{2k+1}{k+1}\right) \tilde{A} - (1-\mu)y_{n_j} + \frac{k\mu}{k+1} L_k(X_{n_j}) \right)\end{aligned}$$

Приравнивая производную к нулю, получаем оптимальный прогноз:

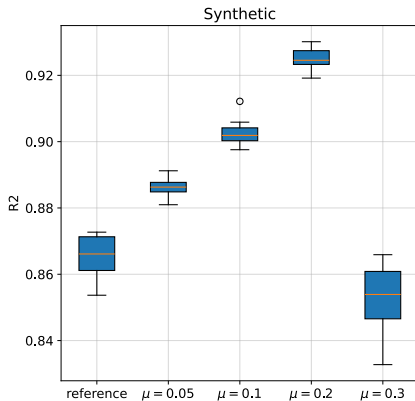
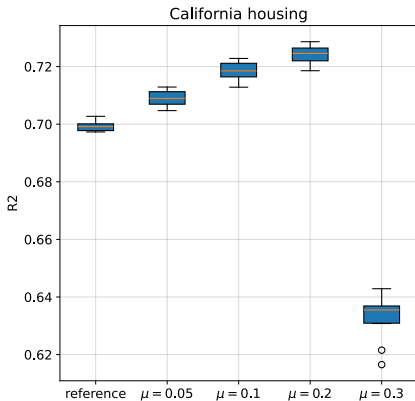
$$\tilde{A} = \sum_{j=1}^p \frac{(k+1)(1-\mu)y_{n_j} - \mu k L_k(X_{n_j})}{p(k+1 - \mu(2k+1))} \quad (2)$$

Предлагаемый метод

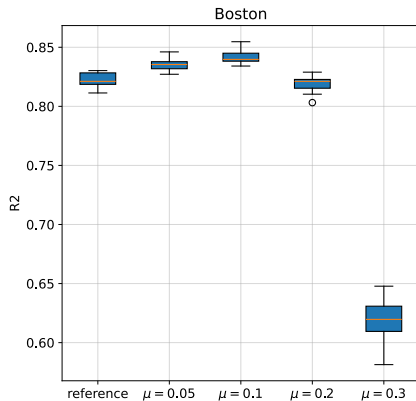
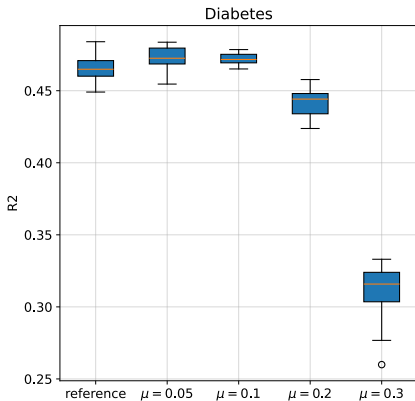
Algorithm Предложенный алгоритм

- 1: Сгенерировать выборку \tilde{X}_1 с помощью бутстрэпа
 - 2: Построить решающее дерево $A_1(x)$ по выборке \tilde{X}_1 , используя только средне-квадратичную ошибку
 - 3: Вычислить $L_1(X) = A_1(X)$ для всех X_1, \dots, X_m
 - 4: **for** $k = 2, \dots, N$ **do**
 - 5: Сгенерировать выборку \tilde{X}_k с помощью бутстрэпа
 - 6: Построить решающее дерево $A_k(x)$ по выборке \tilde{X}_k , используя $L_{k-1}(X)$:
 - В каждой вершине ищется оптимальное разбиение относительно функционала (1)
 - Для вычисления значений в листе используется выражение (2)
 - 7: Вычислить $L_k(X) = \frac{1}{k}((k-1)L_{k-1}(X) + A_k(X))$ для всех X_1, \dots, X_m
 - 8: **end for**
 - 9: Вернуть композицию $\mu_N(X) = \frac{1}{N} \sum_{k=1}^N A_k(X)$
-

Эксперименты



Эксперименты



Выводы

В работе был предложен новый метод ансамблирования деревьев, а также его теоретическое обоснование. Были проведены эксперименты на реальных и синтетических данных, которые показали, что метод достигает лучшего качества, чем обычный случайный лес.