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

A Preprint

Аристархов Данила Дмитриевич
ВМК МГУ
aristarkhov.danila@yandex.ru

Сенько Олег Валентинович
ВМК МГУ
senkoov@mail.ru

Abstract

Рассмотрен новый метод построения ансамбля деревьев при решении задачи регрессии. При этом оптимизация производится исходя из одновременного достижения расходимости алгоритмов в пространстве прогнозов и хорошей аппроксимации данных отдельными алгоритмами ансамбля.

Ключевые слова регрессия · коллективные методы · бэггинг · градиентный бустинг

1 Введение

Методы решения задачи регрессии, основанные на вычислении более точного предсказания с помощью набора менее точных и более простых базовых алгоритмов, получили самое широкое распространение в современном машинном обучении. К числу таких методов может быть отнесен регрессионный случайный лес, а также методы, основанные на использовании адаптивного или градиентного бустинга. Важную роль при построении таких алгоритмов играет способ получения ансамбля так называемых слабых алгоритмов. Теоретический анализ показывает, что увеличение обобщающей способности может быть достигнуто за счет выбора ансамбля алгоритмов, обладающих не только высокой точностью, но и максимально расходящимися прогнозами [Докунин and Сенько, 2015]. В случайном лесе расходимость прогнозов достигается за счет обучения алгоритмов ансамбля на различных выборках, генерируемых из исходной выборки с использованием процедуры бутстрэпа [Brieman, 1996]. В методе градиентного бустинга [Hastie et al., 2009] ансамбль генерируется последовательно. При этом на каждой итерации в ансамбль добавляются деревья, аппроксимирующие первые производные функции потерь в точке, соответствующей текущему прогнозу алгоритма на данном шаге.

Построение ансамбля, имеющего как высокую точность предсказаний, так и сильное расхождение прогнозов отдельных алгоритмов, продемонстрировало улучшение качества модели [Журавлев et al., 2021]. Целью данной работы является продолжение исследования данного подхода.

2 Постановка задачи

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

3 Дисперсия ансамбля

Рассмотрим известное разложение среднеквадратичного риска модели $\mu(x)$ (bias-variance decomposition):

$$L(\mu) = N(\mu) + B(\mu) + V(\mu),$$

где

$$N(\mu) = \mathbb{E}_{x,y}\Big[\big(y - \mathbb{E}[y|x]\big)^2\Big] \text{ — шум,}$$
$$B(\mu) = \mathbb{E}_{x}\Big[\big(\mathbb{E}_X\big[\mu(X)\big] - \mathbb{E}[y|x]\big)^2\Big] \text{ — смещение,}$$
$$V(\mu) = \mathbb{E}_{X}\Big[\big(\mu(X) - \mathbb{E}_X\big[\mu(X)\big]\big)^2\Big] \text{ — разброс.}$$

Проанализируем это разложение в случае ансамбля базовых алгоритмов $\mu(X) = \frac{1}{k}\sum_{i=1}^{k} A_i(X)$. Шум является свойством выборки и не зависит от алгоритма. Выражения для смещения принимает вид:

$$\begin{aligned} B(\mu) &= \mathbb{E}_{x,y}\Big[\Big(\mathbb{E}_X\Big[\frac{1}{k}\sum_{i=1}^{k} A_i(X)(x)\Big] - \mathbb{E}[y|x]\Big)^2\Big] = \\ &= \mathbb{E}_{x,y}\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_X[A_i(X)(x)] - \mathbb{E}[y|x]\Big)^2\Big] = \\ &= \mathbb{E}_{x,y}\Big[\big(\mathbb{E}_X\big[A_i(X)(x)\big] - \mathbb{E}[y|x]\big)^2\Big]. \end{aligned}$$

Таким образом, ансамблирование не изменяет смещенности относительно базового алгоритма.

Запишем выражения для разброса ансамбля:

$$\begin{aligned} V(\mu) &= \mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\Big(\frac{1}{k}\sum_{i=1}^{k} A_i(X)(x) - \mathbb{E}_X\Big[\frac{1}{k}\sum_{i=1}^{k} A_i(X)(x)\Big]\Big)^2\Big]\Big] \\ &= \mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\frac{1}{k^2}\Big(\sum_{i=1}^{k}\Big[A_i(X)(x) - \mathbb{E}_X\big[A_i(X)(x)\big]\Big]\Big)^2\Big]\Big] = \\ &= \mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\frac{1}{k^2}\sum_{i=1}^{k}\Big(A_i(X)(x) - \mathbb{E}_X\big[A_i(X)(x)\big]\Big)^2 + \\ &\quad + \frac{1}{k^2}\sum_{i\neq j}\Big(A_i(X)(x) - \mathbb{E}_X\big[A_i(X)(x)\big]\Big)\Big(A_j(X)(x) - \mathbb{E}_X\big[A_j(X)(x)\big]\Big)\Big]\Big] \\ &= \frac{1}{k^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\sum_{i=1}^{k}\Big(A_i(X)(x) - \mathbb{E}_X\big[A_i(X)(x)\big]\Big)^2\Big]\Big] + \\ &\quad + \frac{1}{k^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\sum_{i\neq j}\Big(A_i(X)(x) - \mathbb{E}_X\big[A_i(X)(x)\big]\Big)\Big(A_j(X)(x) - \mathbb{E}_X\big[A_j(X)(x)\big]\Big)\Big]\Big] \\ &= \frac{1}{k}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\Big(A_i(X)(x) - \mathbb{E}_X\big[A_i(X)(x)\big]\Big)^2\Big]\Big] + \\ &\quad + \frac{k(k-1)}{k^2}\mathbb{E}_{x,y}\Big[\mathbb{E}_X\Big[\Big(A_i(X)(x) - \mathbb{E}_X\big[A_i(X)(x)\big]\Big)\Big(A_j(X)(x) - \mathbb{E}_X\big[A_j(X)(x)\big]\Big)\Big]\Big] \end{aligned}$$

В данном разложении первое слагаемое — это дисперсия одного базового алгоритма, деленная на длину композиции k. Второе — ковариация между двумя базовыми алгоритмами. При малой корреллляции алгоритмов происходит существенное снижение разброса ансамбля по сравнению с базовым алгоритмом.

4 Случайный лес

В качестве базового алгоритма в случайных лесах используются решающие деревья. Они имеют достаточную сложность, и, как следствие, низкую смещенность (при неограниченной глубине дерево

может идеально подстроиться под выборку), но переобучаются и имеют высокий разброс, который можно уменьшить с помощью ансамблирования. В случайных лесах коррелляция между деревьями понижается двумя способами:

- 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.

Так удается добавить случайность в построение деревьев и уменьшить коррелированность их прогнозов.

Список литературы

А. А. Докунин and О. В. Сенько. Регрессионная модель, основанная на выпуклых комбинациях, максимально коррелирующих с откликом. Ж. вычисл. матем и матем. физ., 53:530–544, 2015.

L. Brieman. Bagging predictors. Machine Learning., 24:123–140, 1996.

T. Hastie, R. Tibshirani, and J. H. Friedman. The Elements of Statistical Learning (2nd ed.). 2009.

Ю. И. Журавлев, О. В. Сенько, А. А. Докунин, Н. Н. Киселева, and И. А. Саенко. Двухуровневый метод регрессионного анализа, использующий ансамбли деревьев с оптимальной дивергенцией. Докл. РАН. Матем., информ., проц. упр., 499:63–66, 2021.