© 2022 г. Н.С. КОРОЛЕВ (korolev.nikolay.s@gmail.com)
(Московский государственный университет им. М.В. Ломоносова),
О.В. СЕНЬКО, д-р физ.-мат. наук (senkoov@mail.ru)
(ФИЦ "Информатика и управление" РАН, Москва)

МЕТОД ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ОБУЧЕНИЯ ГРАДИЕНТНОГО БУСТИНГА, ОСНОВАННЫЙ НА МОДИФИЦИРОВАННЫХ ФУНКЦИЯХ ПОТЕРЬ

Рассматривается новый метод повышения качества обучения градиентного бустинга, а также увеличения его обобщающей способности, основанный на использовании модифицированных функций потерь. В ходе выполнения вычислительных экспериментов была показана возможная применимость данного метода для улучшения качества градиентного бустинга, решающего различные задачи классификации и регрессии на реальных данных.

Ключевые слова: градиентный бустинг, дерево решений, функции потерь, машинное обучение, анализ данных.

DOI: 10.31857/S0005231022120078, **EDN:** KSLBAE

1. Введение

Методы машинного обучения широко используются при решении разнообразных прикладных задач [1–3], в которых требуется предсказать значения некоторой неизвестной величины по значениям известных показателей объектов. В последние годы все большее распространение приобретают ансамблевые методы [4, 5], среди которых следует выделить методы, использующие ансамбли регрессионных или решающих деревьев. Известными способами генерации ансамблей являются метод бэггинга и метод случайных подпространств, используемые в случайных лесах, а также метод градиентного бустинга. Эффективность случайных лесов и ансамблевых методов, основанных на градиентном бустинге, подтверждается обширной практикой решения прикладных задач. Вместе с тем убедительные теоретические обоснования их оптимальности отсутствуют. Все это свидетельствует об актуальности исследований, направленных на поиск новых критериев оптимальности генерируемых ансамблей и методов вычисления коллективных решений, которые позволили бы увеличить обобщающую способность ансамблевых алгоритмов. В [6] рассматривается схема поиска ансамбля регрессионных деревьев с минимальной квадратичной ошибкой для среднего прогноза по ансамблю. При этом регрессионные деревья, входящие в ансамбль, генерируются с помощью процедуры бэггинга и дополнительного одношагового градиентного спуска. В настоящей работе анализируется связь схемы с минимальной ошибкой для среднего прогноза со стандартной процедурой градиентного бустинга. Показано, что данная схема по сути сводится к относительно небольшой моди-

фикации стандартной процедуры, которая тем не менее позволяет добиться заметного увеличения обобщающей способности.

2. Модификация градиентного бустинга

Обозначим: $x_1, \ldots, x_N$ — точки в некотором многомерном пространстве, соответствующие известным и легко измеряемым признакам реальных объектов; $y_1, \ldots, y_N$ — значения некоторых трудно измеряемых признаков объектов. Встает задача поиска некоторой функции $f(x)$ такой, что $y_i = f(x_i) + \varepsilon_i$, где ε_i — ошибка предсказания на i-м объекте, т.е. функция $f(x)$ должна приближать реальную зависимость между искомыми значениями y_i и известными признаками x_i. Для построения функции $f(x)$ используется информация лишь о некоторых $T < N$ объектах, а качество приближения проверяется по оставшимся $N - T$ объектам.

Данная задача хорошо решается методами ансамблирования, основанными на построении большого количества деревьев решений. Одним из таких методов является градиентный бустинг [7]. В этой работе стоит задача улучшения качества работы данного метода за счет изменения процедуры обучения градиентного бустинга с использованием модифицированных функций потерь.

2.1. Градиентный бустинг

Пусть:

X — матрица из T строк, i-я строка равна x_i;

Y — вектор из T элементов, i-й элемент равен y_i.

Градиентный бустинг [7] основан на итеративном построении функции $f(x)$ за счет использования большого количества деревьев решений, каждое из которых исправляет ошибки предыдущих. Изначально задается оптимизируемый функционал. Одним из стандартных оптимизируемых функционалов является среднеквадратичная ошибка $L(f(x), X, Y) = \dfrac{1}{T}\sum\limits_{i=1}^{T}(f(x_i) - y_i)^2$. Также вводится некоторое изначальное значение функции предсказания $f_0(x) = 0$ или $f_0(x) = \dfrac{1}{T}\sum\limits_{i=1}^{T} y_i$. Далее итеративно строятся функции $f_i(x) = f_{i-1}(x) + \beta_i h_i(x)$, где $h_i(x)$ — дерево решений, обученное по следующей схеме:

Алгоритм 1 (оригинальный алгоритм градиентного бустинга).

1. Из выборки X, Y при помощи метода бутстрапирования и случайной проекции на произвольное подпространство признаков получается новая выборка $\hat{X}_i, \hat{Y}_i$.

2. Находится антиградиент функции потерь по функции f в точке текущего ансамбля $\hat{h}(x) = -\nabla_f (L(f(x), \hat{X}_i, \hat{Y}_i))\Big|_{f(x)=f_{i-1}(x)}$.

3. Строится дерево решений $h_i(x)$ по выборке $\hat{X}_i, \hat{h}(\hat{X}_i)$.

4. $f_i(x) = f_{i-1}(x) + \beta_i h_i(x)$, где $\beta_i \in \mathbb{R}$ — некоторый коэффициент, с которым добавляется дерево $h_i(x)$ в уже существующий лес $f_{i-1}(x)$.

Спустя некоторое заранее оговоренное количество итераций процесс останавливается и очередное $f_i(x)$ считается искомым $f(x)$.

При этом возможно достаточно большое количество выбора стратегий β_i. Наиболее известными являются:

- $\beta_i = const$,
- $\beta_i = \dfrac{const}{\sqrt{i+1}}$,
- $\beta_i = \dfrac{const}{i+1}$,
- $\beta_i = \arg\min\limits_{\beta} L(f_{i-1}(x) + \beta h(x), \hat{X}_i, \hat{Y}_i)$.

Также существуют различные модификации градиентного бустинга, позволяющие улучшить его обобщающую способность и скорость работы, такие как CatBoost [8], XGBoost [9], LightGBM [10].

2.2. Измененный алгоритм градиентного бустинга

В данной статье придется изменить алгоритм градиентного бустинга. Изменим шаг 2. Вместо того, чтобы брать функцию, равную антиградиенту функции потерь, будем искать такую функцию $\hat{h}(x)$, добавка которой к исходной функции $f_{i-1}(x)$ приведет к минимизации функции потерь, т.е. $L(f_{i-1}(x) + \hat{h}(x), \hat{X}_i, \hat{Y}_i) \to \min\limits_{\hat{h}(x)}$.

Алгоритм 2 (измененный алгоритм градиентного бустинга).

1. Из выборки X, Y при помощи метода бутстрапирования и случайной проекции на произвольное подпространство признаков получается новая выборка $\hat{X}_i, \hat{Y}_i$.

2. Находится такая функция $\hat{h}(x)$, что функция потерь от функции $f_{i-1}(x) + \hat{h}(x)$ на выборке $\hat{X}_i, \hat{Y}_i$ будет минимальна: $L(f_{i-1}(x) + \hat{h}(x), \hat{X}_i, \hat{Y}_i) \to$ $\to \min$. Приравнивая градиент функции потерь к нулю, получим

$$\nabla_{\hat{h}(x)} L(f_{i-1}(x) + \hat{h}(x), \hat{X}_i, \hat{Y}_i) = \frac{2}{T} \sum_{k=1}^{T} (f_{i-1}(\hat{x}_k^i) + \hat{h}(\hat{x}_k^i) - \hat{y}_k^i) = 0,$$

$$\hat{h}(\hat{x}_k^i) = \hat{y}_k^i - f_{i-1}(\hat{x}_k^i),$$

где $\hat{x}_k^i, \hat{y}_k^i$ — признаки k-го объекта в выборке $\hat{X}_i, \hat{Y}_i$.

3. Строится дерево решений $h_i(x)$ по выборке $\hat{X}_i, \hat{h}(\hat{X}_i)$.

4. $f_i(x) = f_{i-1}(x) + \beta_i h_i(x)$, где $\beta_i \in \mathbb{R}$ — некоторый коэффициент, с которым добавляется дерево $h_i(x)$ в уже существующий лес $f_{i-1}(x)$.

Заметим, что в случае, когда функция ошибки L зависит квадратично от функции $f(x)$ и при этом коэффициент при $f^2(x)$ не зависит от значе-

ний выборки, предложенный измененный алгоритм градиентного бустинга полностью совпадает с оригинальным вариантом (с точностью до изменения learning rate'а, который в данном случае обозначается как β_i).

3. Модифицированные функции потерь

3.1. Разложение ошибки

Для произвольного ансамбля алгоритмов известно [11] разложение ошибки ансамбля на смещение отдельных алгоритмов и дисперсию. Сформулируем это разложение в виде теоремы.

Теорема 1. Пусть x, y — произвольные случайные величины; даны K функций $h_k(x)$, и функция $f(x)$ задана как среднее арифметическое этих K функций:

$$f(x) = \frac{1}{K}\sum_{k=1}^{K} h_k(x),$$

тогда верно:

$$\mathbb{E}(y - f(x))^2 = \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}(h_k(x) - y)^2 - \mathbb{E}(h_k(x) - f(x))^2\right]. \tag{1}$$

Доказательство.

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(h_k(x) - y)^2 = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(h_k(x) - f(x) + f(x) - y)^2 =$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}(h_k(x) - f(x))^2 - 2\,\mathbb{E}\left[(h_k(x) - f(x))(y - f(x))\right] + \mathbb{E}(y - f(x))^2\right] =$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}(h_k(x) - f(x))^2 + \mathbb{E}(y - f(x))^2\right] -$$

$$-\frac{2}{K}\,\mathbb{E}\left[\sum_{k=1}^{K}(h_k(x) - f(x))(y - f(x))\right] =$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}(h_k(x) - f(x))^2 + \mathbb{E}(y - f(x))^2\right] -$$

$$-\frac{2}{K}\,\mathbb{E}\left[(y - f(x))\sum_{k=1}^{K}(h_k(x) - f(x))\right] =$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}(h_k(x) - f(x))^2 + \mathbb{E}(y - f(x))^2\right] -$$

$$-\frac{2}{K}\,\mathbb{E}\left[(y - f(x))(Kf(x) - Kf(x))\right] =$$

$$= \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}(h_k(x)-f(x))^2+\mathbb{E}(y-f(x))^2\right] =$$

$$= \mathbb{E}(y-f(x))^2+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(h_k(x)-f(x))^2.$$

Перенося слагаемое $\frac{1}{K}\sum\limits_{k=1}^{K}\mathbb{E}(h_k(x)-f(x))^2$ в левую часть, получим:

$$\mathbb{E}(y-f(x))^2 = \frac{1}{K}\sum_{k=1}^{K}\left[\mathbb{E}(h_k(x)-y)^2-\mathbb{E}(h_k(x)-f(x))^2\right].$$

Соответственно для уменьшения среднеквадратичной ошибки необходимо уменьшать среднеквадратичную ошибку каждого отдельного предиктора $h_k(x)$, а также увеличивать расхождение между прогнозами различных предикторов. Вдохновляясь данным разложением, в следующей главе предложим использовать для обучения градиентного бустинга функцию ошибки, которая будет очень сильно напоминать правую часть данного разложения.

3.2. Среднеквадратичная ошибка с удалением от обученного ансамбля

В модифицированном алгоритме градиентного бустинга, описанном в главе 2.2, на каждой итерации градиентного бустинга оптимизируется среднеквадратичная ошибка:

$$L(f(x)+h(x),X,Y) = \frac{1}{T}\sum_{i=1}^{T}(f(x_i)+h(x_i)-y_i)^2,$$

где $f(x_i)$ — предсказания обученного на текущий момент ансамбля, y_i — отклики, $h(x_i)$ — предсказания нового дерева решений (по ним происходит оптимизация функционала).

Исходя из выведенной ранее формулы 1 для повышения обобщающей способности леса деревьев решений разумно на каждом шаге оптимизировать функцию

$$\frac{1}{T}\sum_{i=1}^{T}\left[(h(x_i)+f(x_i)-y_i)^2-\gamma(f(x_i)-h(x_i))^2\right]. \tag{2}$$

Здесь часть $\frac{1}{T}\sum\limits_{i=1}^{T}(h(x_i)+f(x_i)-y_i)^2$ соответствует слагаемому $\mathbb{E}(h_k(x)-y)^2$ в формуле 1 (для текущей итерации k) за тем исключением, что здесь хотим, чтобы именно сумма функций $f(x)+h(x)$ предсказывала искомый отклик y. Данное изменение было сделано в силу того, что функция $\frac{1}{T}\sum\limits_{i=1}^{T}(h(x_i)+f(x_i)-y_i)^2$ является среднеквадратичной функцией потерь для

алгоритма градиентного бустинга, в то время как функция $\frac{1}{T}\sum_{i=1}^{T}(h(x_i)-y_i)^2$ соответствует процедуре обучения случайного леса. Стоит отметить, что, формально говоря, это не соответствует формуле 1, поэтому отметим, что формула 1 приводится как источник вдохновения и мотивации авторов использовать функцию потерь под формулой 2.

В данной функции есть дополнительная добавка $-\gamma(h(x_i)-f(x_i))^2$, позволяющая добиться дополнительной регуляризации за счет различия между новым обучаемым деревом и уже обученным лесом решающих деревьев. Кроме того, это соответствует разложению среднеквадратичной ошибки ансамбля по формуле 1 с одним лишь исключением — коэффициентом γ перед вторым слагаемым. Коэффициент $\gamma < 1$ необходим для существования минимума по $h(x)$ для функции 2.

Соответственно обучение деревьев будет производиться на откликах, преобразованных следующей функцией:

$$\hat{h}(\hat{x}_k^i) = \frac{\hat{y}_k^i - (1+\gamma)f_{i-1}(\hat{x}_k^i)}{1-\gamma}.$$

Заметим, что для $\gamma = 1$ (как в оригинальной формуле разложения ошибок) минимум не находится, поэтому используется коэффициент γ.

Кроме того, поскольку итоговое дерево $h(x)$, обученное на выходах $\hat{h}(x)$, будет добавляться с некоторым коэффициентом β_i, то, если обучать дерево на выходах $(1-\gamma)\hat{h}(\hat{x}_k^i) = \hat{y}_k^i - (1+\gamma)f_{i-1}(\hat{x}_k^i)$, обученное дерево на самом деле будет соответствовать $(1-\gamma)h(x)$ и можно получить точно такой же лес, добавив полученное дерево с коэффициентом $\dfrac{\beta_i}{1-\gamma}$.

3.3. Смещенная среднеквадратичная ошибка

Будем использовать $L(f(x), X, Y) = \frac{1}{T}\sum_{i=1}^{T}(\alpha f(x_i) - y_i)^2$, где $\alpha \in \mathbb{R}_{++}$. В такой ситуации новые деревья решений будут обучаться на выборке, у которой отклики были преобразованы следующей функцией:

$$\hat{h}(\hat{x}_k^i) = \frac{1}{\alpha}\hat{y}_k^i - f_{i-1}(\hat{x}_k^i).$$

Заметим, что, как и в предыдущем случае, возможно обучение дерева решений на выходах $\alpha\hat{h}(\hat{x}_k^i) = \hat{y}_k^i - \alpha f_{i-1}(\hat{x}_k^i)$ и добавления дерева решений с коэффициентом $\dfrac{\beta_i}{\alpha}$.

Основная идея данного подхода заключается в том, чтобы добавить шума в обучаемые деревья решений, для того чтобы уменьшить корреляцию между выходами различных деревьев решений в итоговом лесе, что позволяет увеличить обобщающую способность обучаемой модели [12].

Отметим, в такой ситуации оказывается, что использование данной функции потерь с параметром $\alpha = 1+\gamma$ абсолютно эквивалентно использованию

предыдущей модификации функции потерь. Кроме того, логично использовать только лишь $0 \leqslant \gamma \leqslant 1$, так как в случае $\gamma < 0$ будет поощряться похожесть откликов нового дерева решений на отклики всего ансамбля, в то время как было решено уменьшать корреляцию между ними, а в случае $\gamma > 1$ функция $\frac{1}{T}\sum\limits_{i=1}^{T}\left[(h(x_i)+f(x_i)-y_i)^2-\gamma(h(x_i)-f(x_i))^2\right]$ будет иметь минимум в точках $h(x_i) = \pm\infty$. В соответствии с границами изменения γ, а также выведенной зависимостью $\alpha = 1 + \gamma$ получаем, что имеет смысл рассматривать лишь $\alpha \in [1; 2]$.

4. Вычислительные эксперименты

Для проверки качества работы представленного метода будем решать различные задачи классификации и регрессии, используя обычный градиентный бустинг, сравнивая результаты работы с градиентным бустингом с использованием смещенной квадратичной ошибки[1]. Кроме того, проводились вычислительные эксперименты с использованием среднеквадратичной ошибки с удалением от обученного ансамбля, но их результаты полностью совпадают с использованием обычной смещенной квадратичной ошибки, что соответствует теории.

4.1. Описание данных

4.1.1. Обнаружение аритмии. Данные[2] состоят из 452 записей ЭКГ. Каждая запись представлена в виде набора из 279 признаков, 206 из которых — линейные, остальные 73 являются перечислимыми. Кроме того, в признаках встречаются пропуски. В данных присутствует информация о том, какой вид аритмии присутствует у пациента, или отмечается факт, что аритмия отсутствует [13]. По этим данным строится дополнительный признак — «наличие аритмии», который равен 0, если аритмия отсутствует, или 1 при наличии любого вида аритмии. Ставится задача классификации по целевому бинарному признаку «наличие аритмии». Целевой метрикой качества является ROC AUC.

4.1.2. Таяние ледников в Арктике. Данные[3] состоят из 170 записей о ледниках в Арктике. Необходимо предсказать, растаял ли ледник или нет, имея информацию о 96 параметрах исследуемого ледника. Целевая метрика качества — ROC AUC.

4.1.3. Предсказание продаж. Данные состоят из 869 записей о количестве продаж некоторого товара вместе с описанием товара. Описание состоит из 286 бинарных признаков и двух линейных численных признаков. Решается задача регрессии по целевому признаку «количество продаж». Целевой метрикой качества является коэффициент детерминации R^2.

[1] https://drive.google.com/file/d/1EyiNNQ_u0CzQ7qYEdZEeEwFTwjkkv2xL/view
[2] https://archive.ics.uci.edu/ml/datasets/Arrhythmia
[3] https://drive.google.com/file/d/1ADa975pas6WPm5SDmPCRF4oPrAyoBkx4/view?usp=sharing

4.1.4. Предсказание систолического давления. Данные представляют из себя 837 записей о 160 параметрах пациентов. Ставится задача регрессии по целевому признаку «систолическое давление» выборки. Целевой метрикой качества является коэффициент детерминации R^2.

4.2. Процедура обучения

Каждая из выборок данных делится на три подвыборки: обучающую, проверочную и тестовую в соотношении 65%/10%/25% соответственно. Затем обучение разбивается на два этапа:

1) поиск гиперпараметра α,

2) проверка эффективности обучения модели.

4.2.1. Поиск гиперпараметра α. На каждом наборе данных производится поиск гиперпараметра α в модели градиентного бустинга со среднеквадратичной ошибкой. Для этого производится обучение нескольких моделей на обучающей выборке с различными значениями гиперпараметра α. Затем качество каждой модели проверяется на проверочной выборке. Параметр, соответствующий модели с наилучшим качеством на данной выборке, используется для следующего этапа обучения.

4.2.2. Проверка эффективности обучения модели. На этом этапе обучающая и проверочная выборки по каждому из наборов данных объединяются, и на соответствующей выборке производится обучение всех последующих моделей.

Модель градиентного бустинга со смещенной среднеквадратичной ошибкой обучается на полученном наборе данных с параметром α, выведенным с предыдущего этапа, затем качество модели оценивается на тестовой части выборки. Аналогичная процедура обучения производится с моделью обычного градиентного бустинга.

4.3. Результаты экспериментов

Наиболее хороших результатов на различных задачах удалось достичь для $\alpha = 1{,}1$. Использование модифицированной функции потерь позволило улучшить предсказательную способность градиентного бустинга как на задачах классификации, так и на задачах регрессии.

Целевая метрика на тестовой выборке для лесов, обученных стандартной процедурой градиентного бустинга (столбец «Среднекв. ошибка») и с использованием смещенной среднеквадратичной ошибки (столбец «Смещенная среднекв. ошибка»)

Набор данных	Среднекв. ошибка	Смещенная среднекв. ошибка	Параметр α
Аритмия	0,89 (ROC AUC)	**0,90** (ROC AUC)	1,7
Ледники	0,72 (ROC AUC)	**0,75** (ROC AUC)	1,1
Продажи	0,21 (R^2)	**0,26** (R^2)	1,1
Сист. давл.	0,41 (R^2)	**0,46** (R^2)	1,1

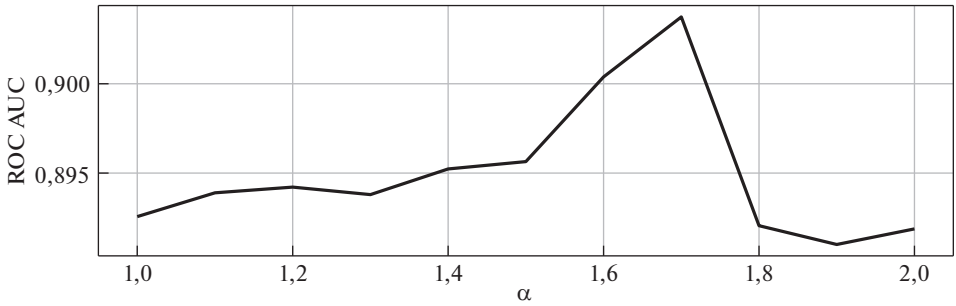

Рис. 1. ROC AUC при обучении классификатора с использованием смещенной среднеквадратичной ошибки для различных параметров α в задаче предсказания аритмии.

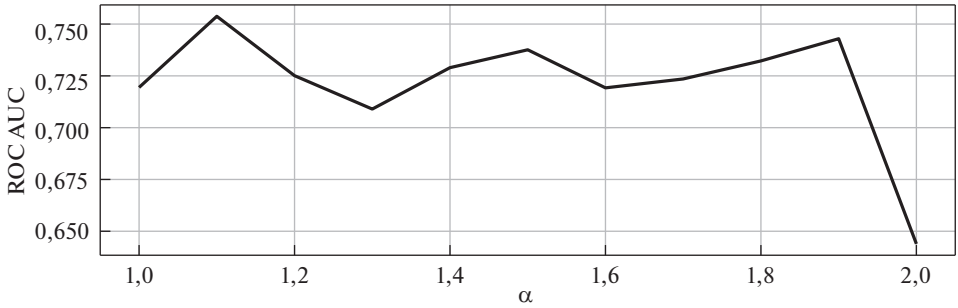

Рис. 2. ROC AUC при обучении классификатора с использованием смещенной среднеквадратичной ошибки для различных параметров α в задаче предсказания таяния ледников.

Итоговые результаты экспериментов представлены в таблице.

Помимо этого приведем графики качества в зависимости от параметра α для каждой задачи (рис. 1–4).

4.4. Анализ полученных результатов

Полученные результаты показывают, что использование модифицированных функций потерь в процедуре обучения решающего дерева при помощи градиентного бустинга способно достаточно серьезно увеличивать обобщающую способность предсказания в сравнении с обычной процедурой обучения градиентного бустинга.

На графиках зависимости качества модели от параметра α явно видно, что при больших значениях параметра α алгоритм имеет слишком низкую обобщающую способность и начинает терять в качестве, помимо этого в трех из четырех задачах наилучшим выбором коэффициента α было 1,1, поэтому считаем, что именно таким стоит выбирать значения гиперпараметра α.

Кроме того, наблюдается эффект недообучения при использовании модифицированных функций потерь в случае, когда объектов достаточно много.

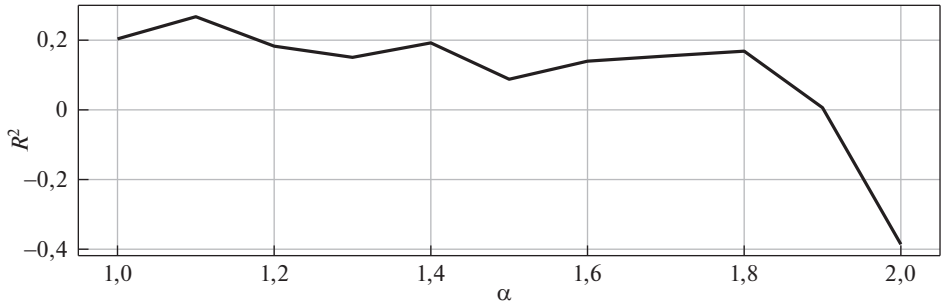

Рис. 3. R^2 при обучении регрессора с использованием смещенной среднеквадратичной ошибки для различных параметров α в задаче предсказания продаж.

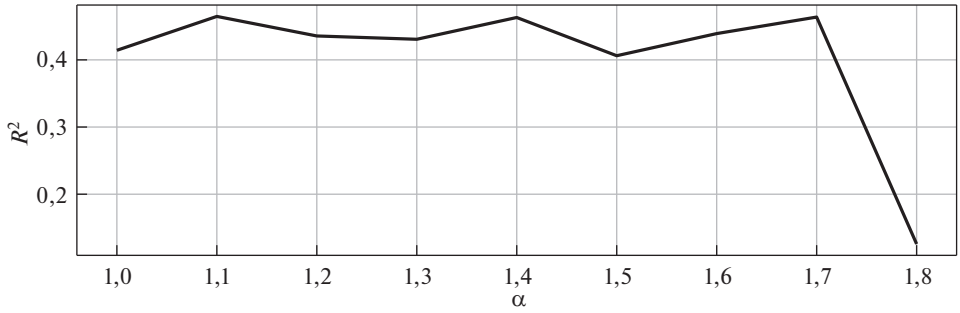

Рис. 4. R^2 при обучении регрессора с использованием смещенной среднеквадратичной ошибки для различных параметров α в задаче предсказания систолического давления.

В такой ситуации применялась смешанная стратегия обучения — первая половина итераций обучалась с использованием смещенной среднеквадратичной ошибки, а вторая половина обучалась в соответствии со стандартным алгоритмом градиентного бустинга.

5. Заключение

В процессе выполнения работы были получены следующие результаты.

— Разработан метод повышения эффективности обучения градиентного бустинга, основанный на использовании модифицированных функций потерь.

— Получено теоретическое доказательство эквивалентности метода, использующего смещенную среднеквадратичную ошибку, и метода, использующего среднеквадратичную ошибку с удалением от обученного ансамбля.

— Проведены вычислительные эксперименты, которые показали возможную применимость данного метода для улучшения обобщающей способности лесов решений, построенных алгоритмом градиентного бустинга, на различных реальных задачах регрессии и классификации.

— Выявлено значение параметра $\alpha = 1{,}1$ в смещенной среднеквадратичной ошибке, при котором достигается наиболее стабильный эффект прироста качества работы модели.

СПИСОК ЛИТЕРАТУРЫ

1. *Friedman Jerome H.* Multiple Additive Regression Trees with Application in Epidemiology // Statist. in Medicine. 2003. V. 22. No. 9. P. 1365–1381.
2. *Elith J.* Boosted Regression Trees for ecological modeling // CRAN. 2018. V. 77(4). P. 802–813.
3. *Lalchand V.* Extracting more from boosted decision trees: A high energy physics case study // arXiv:2001.06033. 2020.
4. *Breiman L.* Random forests // Machine Learning. 2001. V. 45. No. 1.
5. *Zhi-Hua Z.* Ensemble Methods: Foundations and Algorithms // Chapman and Hall/CRC. 2012.
6. *Журавлев Ю.И., Сенько О.В., Докукин А.А., Киселева Н.Н., Саенко И.А.* Двухуровневый метод регрессионного анализа, использующий ансамбли деревьев с оптимальной дивергенцией // ДАН. Математика, информатика, процессы управления. 2021. Т. 499. № 1. С. 63–66.
7. *Friedman Jerome H.* Stochastic gradient boosting // Comput. Statist. & Data Anal. 2002. V. 38. No. 4. P. 367–378.
8. *Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A.* CatBoost: unbiased boosting with categorical features // arXiv:1706.09516. 2017.
9. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System // arXiv:1603.02754. 2016.
10. *Ke G. et al.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems. 2017. V.30. P. 3146–3154.
11. *Gavin Brown, Jeremy Wyatt, Rachel Harris, Xin Yao* Diversity creation methods: A survey and categorisation // Information Fusion. 2005. V. 6. P. 367–378.
12. *Докукин А.А., Сенько О.В.* Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности // Ж. вычисл. матем. и матем. физ. 2011. Т. 51. С. 1751–1760.
13. *Guvenir H. Altay, Acar B., Muderrisoglu H.* Arrhythmia Data Set // https://archive.ics.uci.edu/ml/datasets/Arrhythmia

Статья представлена к публикации членом редколлегии А.А. Лазаревым.

Поступила в редакцию 31.01.2022

После доработки 21.06.2022

Принята к публикации 29.06.2022