УДК 004.89

НЕКОТОРЫЕ ПОДХОДЫ К ПОВЫШЕНИЮ ТОЧНОСТИ ПРОГНОЗИРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ АНСАМБЛЕВЫХ МЕТОДОВ

С. Ма[1] [0009-0004-0356-9996], О. В. Сенько[2] [0000-0002-5586-3503]

[1, 2]*Московский государственный университет имени М. В. Ломоносова, г. Москва, Россия*

[2] *Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва,Россия*

[1]xinyuema35@gmail.com, [2]OSenko@frccsc.ru

Аннотация

Представлены результаты экспериментального исследования эффективности использования сверхслучайных деревьев в моделях, основанных на градиентном бустинге, а также в новом ансамблевом методе, в котором лес генерируется, исходя из условия повышенной внутренней дивергенции. сследована эффективность сверхслучайных деревьев при использовании расширенных наборов признаков с включением новых признаков, вычисляемых как расстояния Идо набора описаний опорных объектов из обучающей выборки. Показано, что использование сверхслучайных деревьев в моделях градиентного бустинга и дивергентного леса позволяет улучшить обобщающую способность, а также, что к еще большему росту обобщающей способности приводит использование расширенных наборов признаков.

Ключевые слова*: регрессионное моделирование, ансамблевое обучение, метрическое пространство, метод сверхслучайных деревьев.*

ВВЕДЕНИЕ

Ансамблевые методы, основанные на использовании комбинаций более простых алгоритмов, получили широкое распространение при решении разнообразных прикладных задач прогнозирования числовых целевых переменных. Высокая эффективность ансамблевых методов, убедительно подтверждаемая результатами многочисленных экспериментов, делает актуальными дальнейшие

исследования по их совершенствованию. Методы ансамблевого обучения основаны на вычислении коллективного прогноза по набору прогнозов, вычисляемых базовыми алгоритмами, вошедшими в ансамбль, что позволяет повысить устойчивость и точность предсказаний. Формально предсказание ансамбля можно записать в виде

$$a(x) = m\big(b_1(x), \ldots, b_n(x)\big),$$

где $b_i(x)$, $i = 1, \ldots, n$, – предсказания отдельных базовых моделей, а $m(x)$ – мета-алгоритм, агрегирующий эти предсказания.

Построение ансамбля состоит из двух этапов: во-первых, обучение нескольких базовых моделей; во-вторых, применение стратегии объединения их выходов для получения финального результата. Однако ансамблевые методы обычно используют фиксированные способы объединения моделей. Теоретическое обоснование целесообразности использования ансамблей восходит к теореме Кондорсе о присяжных [1]. В соответствии с этой теоремой, если каждый голосующий высказывает независимое мнение и в среднем принимает верное решение чаще, чем ошибается, то вероятность правильного вердикта большинства стремится к единице по мере увеличения числа голосующих.

Ансамблевые методы имеют длительную историю. Одними из первых ансамблевых методов были тестовый алгоритм [2] и алгоритм Кора [3], предложенные еще в 1960-е годы. Идея использования ансамблей решающих и регрессионных деревьев возникла в 1993 г. (см. в [4]). В 2001 г. окончательно оформилась идея случайного леса [5], в котором ансамбль регрессионных или решающих деревьев генерируется с использованием метода бэггинга (bagging, [6]) и метода случайных подпространств [7]. В методе бэггинга деревья обучаются по выборкам, которые являются выборками с возвращением из исходной обучающей выборки. В методе случайных подпространств обучение производится по выборкам, получаемым из исходной выборки с помощью случайного выбора подмножества признаков. Необходимо отметить, что при построении случайного леса каждое новое дерево строится независимо от предыдущих деревьев, исходя из условия наилучшей аппроксимации исходной целевой переменной.

В отличие от случайного леса, метод бустинга (boosting, [8–11]) направлен на построение линейной комбинации «слабых» алгоритмов. На каждом шаге

в нее добавляется новое слагаемое согласно условию минимизации ошибки линейной комбинации. Одним из первых представителей данного подхода является алгоритм AdaBoost [12], в котором при обучении используются веса объектов обучающей выборки. На первом шаге веса объектов выбираются равными. На последующих шагах увеличиваются веса объектов, предсказания для которых были ошибочными. Более широкое распространение по сравнению с AdaBoost в последнее время получил градиентный бустинг, в котором минимизация потерь на каждом шаге производится с использованием градиентного спуска, а каждое новое дерево, добавляемое в линейную комбинацию, аппроксимирует антиградиент функции потерь. Подобные методы демонстрируют очевидные преимущества при моделировании сложных нелинейных зависимостей.

С развитием алгоритмов и вычислительных ресурсов современные методы градиентного бустинга были значительно усовершенствованы с точки зрения производительности и эффективности. Широкое распространение получили модификации градиентного бустинга: XGBoost [13], LightGBM [14] и CatBoost [15]. В [16–18] предложен новый вариант регрессионного леса, основанный на анализе разложения квадратичной ошибки выпуклых комбинаций предикторов. Из разложения следует, что ошибка может быть значительно снижена при увеличении взаимного квадратичного отклонения прогнозов алгоритмов, входящих в ансамбль. В связи с этим было предложено при построении нового дерева, включаемого в ансамбль, не только минимизировать квадратичную ошибку прогноза, но и одновременно максимизировать квадратичное отклонение от текущего ансамбля. Эксперименты показали, что такой подход, который носит название дивергентного леса [18], во многих случаях позволяет снижать ошибку ансамбля.

В современных вариантах алгоритмов случайного регрессионного леса и различных вариантов градиентного бустинга обычно используются регрессионные деревья, в которых пороги для признаков выбираются по критериям минимизации ошибки прогноза. В последнее время растет интерес к так называемым сверхслучайным деревьям (Extra Randomized Trees [19]), в которых пороговые значения для признаков выбираются случайно. Эксперименты показали [19], что нередко леса, состоящие из сверхслучайных деревьев, превосходят по обобщаю-

щей способности стандартные регрессионные случайные леса, требуя значительно меньше времени на обучение. В связи с этим возникла идея исследовать эффективность использования сверхслучайных деревьев также в градиентном бустинге и дивергентном лесе.

Другим способом повышения обобщающей способности является трансформация признакового пространства [20]. Поэтому еще одной целью настоящей работы стало исследование эффективности ансамблевых методов с использованием в качестве признаков расстояний до опорных векторных описаний объектов из обучающей выборки. Обучающая выборка состоит из n объектов, каждый из которых описывается m признаками: x_i=(x_{i1}, x_{i2}, ..., x_{im}) ∈ R^m и включает соответствующее целевое значение $y_i \in R$. Тогда обучающая выборка записывается как $S = \{(x_i, y_i)\}, i = 1, \ldots, n$.

МЕТОДОЛОГИЯ

Опишем методологию исследования, включая характеристики набора данных, общую структуру ансамблевых моделей и принципы их построения.

Описание данных

В проведенном исследовании был использован набор данных, содержащий физико-химические характеристики различных химических соединений и соответствующие значения их температуры плавления. Каждый образец (соединение) описан множеством числовых признаков, соответствующих свойствам структуры, энергии и состава вещества. Целевой переменной является температура плавления соединения T_m, измеряемая в кельвинах (K).

Данные представлены в табличной форме:

- первая колонка содержит наименование соединения (Compound);
- во второй указано значение целевой переменной T_m;
- остальные столбцы соответствуют числовым признакам, всего представлено m характеристик;
- полный набор данных включает *n* наблюдений.

Перед началом обучения данные были разделены на обучающую и тестовую выборки в соотношении 80/20 для последующего обучения и проверки модели.

Используемые модели

Перед описанием предлагаемых ансамблей рассмотрим базовый алгоритм, на основе которого они построены, – метод сверхслучайных деревьев.

Краткое сравнение сверхслучайного дерева и стандартного регрессионного дерева приведено в табл. 1.

Метод снижает вычислительные затраты на построение модели и повышает разнообразие деревьев по сравнению с классическим методом [21], что способствует лучшей обобщающей способности ансамбля и снижает риск переобучения.

Табл. 1. Сравнение ансамблей на основе моделей сверхслучайного дерева и стандартного регрессионного дерева (Regression Tree).

Критерий сравнения	Сверхслучайное дерево	Регрессионное дерево
Выбор точки разделения	Случайная точка разделения (*random split*)	Лучшая точка разделения (*best split*)
Скорость обучения	Быстрее (нет поиска, выбирается случайно)	Медленнее (нужно искать оптимальное разделение)
Смещение/дисперсия	Более высокое смещение, низкая дисперсия → устойчивость	Низкое смещение, высокая дисперсия → переобучение

В качестве модели регрессии A мы реализовали два ансамблевых подхода на основе метода сверхслучайных деревьев, в которых одиночное дерево используется в качестве базового алгоритма, добавляемого в ансамбль на каждой итерации:

- градиентный бустинг с аппроксимацией на каждом шаге градиента функции потерь;
- дивергентный лес с итеративным обновлением целевой переменной.

Градиентный бустинг

Градиентный бустинг формирует итоговую модель поэтапно, последовательно добавляя новые базовые алгоритмы, каждый из которых нацелен на устранение ошибок предыдущих шагов. В отличие от случайного леса, где деревья строятся независимо, бустинг организует обучение в последовательной форме, направляя внимание на трудные для предсказания наблюдения.

Пусть задана функция потерь $L\big(y, F(x)\big)$, измеряющая расхождение между истинным значением y и предсказанием модели $F(x)$.

На каждом шаге бустинг добавляет новую базовую модель $h_k(x)$, которая аппроксимирует направление наискорейшего спуска функции потерь – отрицательный градиент.

Алгоритм градиентного бустинга:

- Инициализация

$$F_0(x) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_\text{i}, c).$$

- Для каждой итерации $k = 1,2,\ldots,K$:

  – вычисляются антиградиенты

$$r_i^{(k)} = \frac{-\partial L\left(y_j, F_{k-1}(x_\text{i})\right)}{\partial F_{k-1}(x_\text{i})};$$

  – обучается базовая модель $h_k(x)$ по парам $(r_i^{(k)}, x_i)$;

  – обновляется ансамбль:

$F_k(x) = F_{k-1}(x) + \eta \cdot h_k(x)$, где $\eta \in (0,1]$ – шаг обучения (learning rate).

Если функция потерь имеет вид

$$L(y, F) = \frac{1}{2}(y - F)^2,$$

то отрицательный градиент совпадает с обычным остатком:

$$r_i^{(k)} = y_i - F_{k-1}(x).$$

Следовательно, на каждом шаге базовая модель $h_k(x)$ обучается на остатках предыдущего шага.

В рамках настоящего исследования градиентный бустинг был реализован с использованием сверхслучайных деревьев в качестве базовых моделей. В реализации для текущей задачи процесс обучения выглядит следующим образом.

- На первом шаге вычисляются остатки $r_i^{(1)} = y_i - \overline{Y}$, где $\overline{Y}$ – среднее значение Y на обучающей выборке.
- Строится сверхслучайное дерево T_1 по парам $\left(X_{\text{train}},\ r_{\text{train}}^{(1)}\right)$.
- Предсказание вычисляется по формуле

$$F_1(X_{\text{train}}) = \overline{Y} + \eta T_1(X_{\text{train}}).$$

- На шаге k вычисляются остатки

$$r_i^{(k)} = y_i - F_{k-1}(x_i).$$

- Строится сверхслучайное дерево T_k по парам $\left(X_{\text{train}},\ r_{\text{train}}^{(k)}\right)$.
- Предсказания обновляются по правилу

$$F_k(X_{\text{train}}) = F_{k-1}(X_{\text{train}}) + \eta \cdot T_k(X_{\text{train}}).$$

Дивергентный лес

В отличие от градиентного бустинга, который минимизирует функцию потерь на каждом шаге, метод дивергентного леса [17, 18] использует альтернативную схему обновления целевой переменной.

Разработанная нами модификация дивергентного леса представляет собой ансамблевый алгоритм, построенный на основе сверхслучайных деревьев. Основная идея метода заключается в итеративном изменении целевых переменных, что способствует увеличению разнообразия базовых моделей и улучшению общей точности предсказаний, такой подход объединяет концепции последовательного обучения, присущего бустингу, и стохастического характера сверхслучайных деревьев.

Каждая итерация ансамбля использует не исходные метки, а адаптированную целевую переменную $S_{\text{train}}^{(k)}$, скорректированную с учетом накопленных предсказаний. Это позволяет каждой новой модели обучаться на обновленных данных, что обеспечивает более устойчивое и сбалансированное поведение ансамбля.

Алгоритм метода можно представить следующим образом.

- Инициализация

$$Z_{\text{train}}^{(0)} = Y_{\text{train}}.$$

- Для каждой итерации k:

– обучается сверхслучайное дерево T_k на текущем векторе $Z_{\text{train}}^{(k-1)}$;

– вычисляется предсказание $P_k = T_k(X_{\text{train}})$;

– обновляется целевая переменная

$$\overline{P_k} = \frac{1}{k}\sum_{j=1}^{k} P_j, \quad Z_{\text{train}}^{(k-1)} = \frac{Y_{\text{train}} - \mu\bar{P}_k}{1-\mu};$$

где $\mu \in (0,1)$ – параметр, регулирующий баланс между дивергенцией и точностью аппроксимации.

- Финальное предсказание при количестве итераций r вычисляется по формуле $\frac{1}{\text{r}}\sum_{j=1}^{\text{r}} P_j$.

Предложенный подход отличается от традиционного бустинга тем, что вместо оптимизации градиента ошибки применяется явное аналитическое обновление, основанное на сглаженном ансамбле предсказаний. Такой механизм позволяет контролировать вклад новых и предыдущих моделей посредством параметра μ, снижая чувствительность к шуму и нестабильности меток.

Таким образом, дивергентный лес можно рассматривать как обобщенную версию бустинга, ориентированную на повышение устойчивости и стабильности ансамбля при работе с реальными, зашумленными данными.

Метод опорных точек и расстояние Махаланобиса

Одним из важных элементов представленного подхода является использование метода опорных точек для построения нового метрического пространства. Основная идея состоит в том, чтобы расширить исходное пространство признаков новыми координатами, которые отражают расстояния от каждого объекта до специально выбранных репрезентативных (опорных) точек. Предполагается, что такое преобразование позволит точнее выделить структурные связи в данных и облегчит поиск скрытых закономерностей, которые не всегда хорошо улавливаются моделью регрессионных деревьев в исходных координатах при коррелированных признаках.

Пусть имеется обучающая выборка из *n* объектов, каждый из которых описывается *m* признаками. Введем множество из *k* опорных точек – базовых элементов, относительно которых будет формироваться новое пространство признаков. Эти точки можно рассматривать как «ориентиры», позволяющие измерять степень близости других объектов к сложным областям исходного пространства.

Для отбора опорных точек используем предварительно обученную модель A_0. После обучения для каждого объекта вычислим ошибку предсказания

$$e_i = (A_0(x) - y_i)^2.$$

В качестве опорных выберем объекты, у которых значение ошибки наибольшее. Интуитивно ясно, что это те наблюдения, которые труднее всего аппроксимировать базовой моделью. Если использовать их как эталоны, то можно улучшить способность модели учитывать сложные или слабо представленные области пространства признаков. Такой принцип близок идее адаптивного бустинга, где внимание последующих моделей сосредоточено на наблюдениях, для которых предыдущие модели допускают наибольшую ошибку. Для каждого объекта x_i построим новый вектор признаков z_i, компоненты которого представляют собой расстояния до всех выбранных опорных точек:

$$z_i = (\rho(x_i, x_1), \rho(x_i, x_2), \ldots, \rho(x_i, x_k))).$$

В результате получим новое расширенное описание как конкатенацию векторов x_i и z_i, которое далее обозначим как $[x_i, z_i]$.

В качестве меры расстояния используется расстояние Махаланобиса [22], определяемое формулой

$$\rho_M(x_i, x_k) = \sqrt{\sum_{p=1}^{m} \sum_{q=1}^{m} \left(x_{ip} - x_{kp}\right) \Sigma_{pq}^{-1} \left(x_{iq} - x_{kq}\right)},$$

где Σ – ковариационная матрица.

В отличие от евклидовой метрики, расстояние Махаланобиса учитывает масштаб признаков и их взаимную корреляцию. Это особенно важно, когда признаки сильно различаются по дисперсии или связаны между собой. В таких случаях евклидово расстояние может искажать реальную структуру данных, тогда как расстояние Махаланобиса дает более корректную оценку близости.

После добавления новых координат преобразуем исходную выборку: $S' = \{([x_i, z_i], y_i)\}$, $i = 1, \ldots, n$. Затем обучим новую модель $A[x_i, z_i] \approx y_i$.

Построение признаков на основе расстояний до опорных точек делает модель более гибкой и чувствительной к структуре данных. Метод особенно полезен при наличии значительной корреляции признаков, а также при сложной

форме распределения данных. Тем не менее чрезмерное увеличение числа новых признаков может привести к переобучению, поэтому количество опорных точек *k* следует подбирать эмпирически.

В ходе экспериментов было показано, что использование расширенного пространства признаков на основе расстояния Махаланобиса дает более устойчивые результаты по сравнению с другими метриками. Причина заключается в том, что эта метрика устраняет два эффекта:

- масштабный эффект: признаки с высокой дисперсией не доминируют над остальными;
- корреляционный эффект: сильно зависимые признаки не учитываются дважды.

ПРОВЕДЕННЫЕ ЭКСПЕРИМЕНТЫ

Основная цель экспериментов заключалась в проверке эффективности предложенного подхода. Эксперименты были направлены на сравнение точности и устойчивости различных ансамблевых моделей в задаче регрессии.

Сравниваемые модели

В ходе экспериментов были рассмотрены следующие методы регрессии.

- Случайный лес, построенный с использованием сверхслучайных деревьев в качестве базового алгоритма (ET in RF).
- Дивергентный лес, построенный с использованием стандартных регрессионных деревьев в качестве базового алгоритма (RT in DF).
- Дивергентный лес, построенный с использованием сверхслучайных деревьев в качестве базового алгоритма (ET in DF).
- Дивергентный лес, построенный с использованием расширенного описания и сверхслучайных деревьев в качестве базового алгоритма (ET in DF (расш.)).
- Классический градиентный бустинг на деревьях решений (GBRT).
- Градиентный бустинг с использованием сверхслучайных деревьев в качестве базового алгоритма (GBET).

- Градиентный бустинг, построенный с использованием расширенного описания и сверхслучайных деревьев в качестве базового алгоритма (GBET (расш.))
- LightGBM – реализация градиентного бустинга с построением деревьев по принципу leaf-wise и оптимизациями по скорости и памяти.
- CatBoost – реализация бустинга, использующая упорядоченные статистики и устойчивую обработку категориальных признаков.

Параметры оценки качества

Для оценки качества регрессионных моделей был использован коэффициент детерминации R^2, который показывает, какая доля дисперсии целевой переменной объясняется моделью.

Формально R^2 определяется как

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \mathrm{f}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2},$$

где y_i – истинные значения целевой переменной, f_i – предсказанные моделью значения, y_i – среднее значение по всем наблюдениям. Числитель $\sum_{i=1}^{n}(y_i - \mathrm{f}_i)^2$ представляет собой остаточную сумму квадратов ошибок, а знаменатель $\sum_{i=1}^{n}(y_i - \overline{y_i})^2$ – полную сумму квадратов. Таким образом, R^2 измеряет, насколько модель уменьшает ошибку по сравнению с простейшей моделью, которая всегда предсказывает среднее значение.

В табл. 2 представлены значения коэффициента детерминации R^2 для десяти задач. По строкам указаны задачи, по столбцам – использованные методы, указанные выше; обозначения Rel. diff (%) – относительная разница между наилучшим и наихудшим результатом; *N* – число объектов; *n* – число признаков.

Таблица 2. Результаты эксперимента

	RT in DF	ET in RF	ET in DF	ET in DF (расш.)	CatBoost	LightGBM	GBRT	GBET	GBET (расш.)	Rel.diff (%)	Size (N x n)
Задача 1	0.894	0.902	0.904	0.904	0.881	0.845	0.902	0.917	**0.917**	8.5	439*86
Задача 2	0.906	0.905	0.924	**0.939**	0.903	0.921	0.912	0.932	0.934	4.1	451*98
Задача 3	0.918	0.94	0.942	0.943	0.935	0.932	0.923	0.945	**0.947**	3.1	439*86
Задача 4	0.850	0.892	0.914	0.915	0.901	0.917	0.870	0.908	**0.920**	8.2	196*88
Задача 5	0.851	0.892	0.916	0.919	0.908	0.912	0.870	0.908	**0.929**	9.2	447*90
Задача 6	0.797	0.860	0.896	**0.919**	0.890	0.886	0.796	0.897	0.906	15.4	234*98
Задача 7	0.902	0.908	0.927	**0.933**	0.873	0.927	0.899	0.928	0.928	7.0	231*98
Задача 8	0.902	0.889	0.926	**0.935**	0.908	0.911	0.919	0.926	0.926	5.2	235*86
Задача 9	0.864	0.833	0.869	**0.877**	0.848	0.877	0.851	0.870	**0.877**	5.3	195*88
Задача 10	0.844	0.768	0.863	0.863	0.869	0.841	0.768	0.863	**0.879**	14.5	173*68

Анализ результатов

Модели, основанные на сверхслучайных деревьях, демонстрируют более высокие результаты по сравнению с моделями на деревьях решений. Их использование в структуре дивергентного леса обеспечивает лучшие показатели, чем в случайном лесе, что подтверждает большую значимость дивергентного ансамблирования для данного базового алгоритма.

Сравнение градиентного бустинга на деревьях решений и градиентного бустинга на сверхслучайных деревьях показало преимущество второго: за счет большего разнообразия базовых моделей и снижения корреляции их ошибок ансамбль на сверхслучайных деревьях демонстрирует более высокую обобщающую способность и меньшую склонность к переобучению на рассмотренных данных.

Применение нового метрического пространства стабильно улучшает результаты. Так, сверхслучайные деревья в дивергентном лесе с новой метрикой дают более высокие значения R^2 по сравнению с исходным пространством во всех задачах. Аналогично, градиентный бустинг на сверхслучайных деревьях в новом метрическом пространстве также демонстрирует улучшения относительно базового варианта.

Наиболее высокие результаты наблюдаются в следующих случаях:

- в задачах 2, 6, 7 и 8 – сверхслучайные деревья в дивергентном лесе с расширенным набором признаков;
- в задачах 1, 3, 4, 5 и 10 – градиентный бустинг на сверхслучайных деревьях с расширенным набором признаков;
- в задаче 9 оба подхода показали сопоставимые и максимально высокие значения R^2.

Таким образом, использование нового метрического пространства повышает устойчивость и эффективность моделей, однако выбор оптимального метода зависит от специфики конкретной задачи.

ЗАКЛЮЧЕНИЕ

Исследованы различные ансамблевые методы регрессии, включая случайный лес, градиентный бустинг, LightGBM, CatBoost, а также предложена модификация метода дивергентного леса. Особое внимание уделено выбору базовых алгоритмов и анализу влияния метрических методов расширения признакового пространства на качество предсказаний.

Результаты экспериментов показали, что использование метода сверхслучайных деревьев в качестве базового алгоритма обеспечивает наилучшее соотношение между точностью и устойчивостью модели. По сравнению с классическими

деревьями решений, метод сверхслучайных деревьев демонстрирует более высокие значения R^2 во всех вариантах ансамблей – как в случайном лесе, так и в дивергентном лесе и градиентном бустинге.

Метод дивергентного леса продемонстрировал преимущество по сравнению со случайным лесом, особенно при увеличении параметра баланса μ. Это подтверждает эффективность итеративного обновления целевой переменной, которое позволяет модели лучше адаптироваться к сложной структуре данных.

Переход к расширенному признаковому пространству с помощью метода на основе метрики Махаланобиса дал дополнительное улучшение качества для всех моделей. Данная метрика оказалась особенно полезной при наличии коррелированных признаков, так как она учитывает взаимные зависимости и масштаб признаков, формируя более адекватное представление о расстояниях между объектами.

В целом можно сделать следующие выводы.

- Метод сверхслучайных деревьев как базовый алгоритм обеспечивает стабильное и высокое качество прогнозов как в дивергентном лесе, так и в градиентном бустинге.
- Использование расширенного признакового пространства с помощью метода на основе метрики Махаланобиса позволяет улучшить обобщающую способность моделей и повысить точность на сложных данных.

Таким образом, предложен подход, основанный на сочетании двух принципов: использование сверхслучайных деревьев в дивергентном лесе и градиентном бустинге; расширение признакового пространства с использованием метрики Махаланобиса. Показано, что подход повышает эффективность регрессионного моделирования и улучшает адаптацию моделей к структурно сложным наборам данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Хабр.* Открытый курс машинного обучения. Тема 5. Композиции: бэггинг, случайный лес [Электронный ресурс].
URL: https://habr.com/ru/companies/ods/articles/324402/ (дата обращения: 06.11.2025).

2. *Дмитриев А.И., Журавлев Ю.И., Кренделев Ф.П.* О математических принципах классификации предметов или явлений // Дискретный анализ. 1967. Вып. 7. С. 3–17.

3. *Вайнцвайг М.Н.* Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов. М.: Сов. радио, 1973. С. 8–12.

4. *Heath D., Kasif S., Salzberg S.* k-DT: A multi-tree learning method // Proceedings of the Second International Workshop on Multistrategy Learning. 1993. P. 138–149. https://doi.org/10.1007/0-387-34296-6_10

5. *Breiman L.* Random Forests // Machine Learning. 2001. Vol. 45, No. 1. P. 5–32. https://doi.org/10.1023/A:1010933404324

6. *Breiman L.* Bagging predictors // Machine Learning. 1996. Vol. 24, No. 2. P. 123–140. https://doi.org/10.1007/BF00058655

7. *Ho T.K.* The Random Subspace Method for Constructing Decision Forests // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. Vol. 20, No. 8. P. 832–844. https://doi.org/10.1109/34.709601

8. *Freund Y., Schapire R.E.* A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting // Journal of Computer and System Sciences. 1997. Vol. 55. P. 119–139. https://doi.org/10.1006/jcss.1997.1504

9. *Friedman J.H.* Stochastic Gradient Boosting // Computational Statistics & Data Analysis. 2002. Vol. 38, No. 4. P. 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

10. *Zhou Z.H.* Ensemble Methods: Foundations and Algorithms. New York: Chapman and Hall/CRC, 2012. 446 p.

11. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer, 2009. 745 p. https://doi.org/10.1007/978-0-387-84858-7

12. *Beja-Battais P.* Overview of AdaBoost: Reconciling its Views to Better Understand its Dynamics // arXiv preprint arXiv:2310.18323 [cs.LG]. 2023. https://doi.org/10.48550/arXiv.2310.18323

13. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794. https://doi.org/10.48550/arXiv.1603.02754

14. *Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.

15. *Hancock J.T., Khoshgoftaar T.M.* CatBoost for big data: an interdisciplinary review // Journal of Big Data. 2020. Vol. 7, No. 94. 45 p. https://doi.org/10.1186/s40537-020-00369-8

16. *Zhuravlev Yu.I., Senko O.V., Dokukin A.A., Kiselyova N.N., Saenko I.A.* Two-Level Regression Method Using Ensembles of Trees with Optimal Divergence // Doklady Mathematics. 2021. Vol. 103, No. 1. P. 1–4. https://doi.org/10.1134/S1064562421040177

17. *Dokukin A.A., Sen'ko O.V.* A New Two-Level Machine Learning Method for Evaluating the Real Characteristics of Objects // Journal of Computer and Systems Sciences International. 2023. Vol. 62, No. 4. P. 607–614. https://doi.org/10.1134/S1064230723040020

18. *Senko O.V., Dokukin A.A., Kiselyova N.N., Dudarev V.A., Kuznetsova Yu.O.* New Two-Level Ensemble Method and Its Application to Chemical Compounds Properties Prediction // Lobachevskii Journal of Mathematics. 2023. Vol. 44, No. 1. P. 188–197. https://doi.org/10.1134/S1995080223010341

19. *Geurts P., Ernst D., Wehenkel L.* Extremely Randomized Trees // Machine Learning. 2006. Vol. 63, No. 1. P. 3–42. https://doi.org/10.1007/s10994-006-6226-1

20. *López-Iñesta E., Grimaldo F., Arevalillo-Herráez M.* Combining feature extraction and expansion to improve classification-based similarity learning // Pattern Recognition Letters. 2016. Vol. 85. P. 84–90. https://doi.org/10.1016/j.patrec.2016.11.005

21. *Breiman L., Friedman J., Olshen R.A., Stone C.J.* Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole, 1984. 358 p. ISBN 978-0-412-04841-8. https://doi.org/10.1201/9781315139470

22. *Mahalanobis P.C.* On the Generalised Distance in Statistics (reprint of 1936) // Sankhya A. 2018. Vol. 80, Suppl. 1. P. 1–7. https://doi.org/10.1007/s13171-019-00164-5

SOME APPROACHES TO IMPROVING PREDICTION ACCURACY USING ENSEMBLE METHODS

X. Ma[1] [0009-0004-0356-9996], O. V. Senko[2] [0000-0002-5586-3503]

[1, 2]Lomonosov Moscow State University, Moscow, Russia

[2]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

[1]xinyuema35@gmail.com, [2]OSenko@frccsc.ru

Abstract

This study presents the results of an experimental analysis evaluating the effectiveness of Extra Trees within gradient boosting models, as well as in a newly proposed ensemble framework where the forest is generated under conditions of enhanced internal divergence. Additionally, the paper explores the performance of Extra Trees when applied to novel feature representations computed as distances to a selected set of reference examples. It has been shown that the use of Extra Randomized Trees in gradient boosting and divergent forest models improves generalization ability. The use of expanded feature sets leads to even greater generalization ability.

Keywords*: regression modeling, ensemble learning, metric space, extremely randomized trees method.*

REFERENCES

1. *Habr***.** *Open Machine Learning Course. Topic 5. Ensembles: Bagging, Random Forest.* Available at: https://habr.com/ru/companies/ods/articles/324402/ (accessed 6 November 2025). (In Russ.).

2. *Dmitriev A.I., Zhuravlev Yu.I., Krendelev F.P.* O matematicheskikh printsipakh klassifikatsii predmetov ili yavlenii [On the Mathematical Principles of the Classification of Objects and Phenomena] // Diskretnyi analiz [Discrete Analysis]. 1967. No. 7. P. 3–17 (In Russ.).

3. *Vaintsvaig M.N.* Algoritm obucheniya raspoznavaniyu obrazov “Kora” [Algorithm for pattern recognition learning “Kora”] // Algoritmy obucheniya raspoznavaniyu obrazov [Algorithms for pattern recognition learning]. Moscow: Sovetskoe radio, 1973. P. 8–12 (In Russ.).

4. *Heath D., Kasif S., Salzberg S.* k-DT: A multi-tree learning method // Proceedings of the Second International Workshop on Multistrategy Learning. 1993. P. 138–149. https://doi.org/10.1007/0-387-34296-6_10

5. *Breiman L.* Random Forests // Machine Learning. 2001. Vol. 45, No. 1. P. 5–32. https://doi.org/10.1023/A:1010933404324

6. *Breiman L.* Bagging predictors // Machine Learning. 1996. Vol. 24, No. 2. P. 123–140. https://doi.org/10.1007/BF00058655

7. *Ho T.K.* The Random Subspace Method for Constructing Decision Forests // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998. Vol. 20, No. 8. P. 832–844. https://doi.org/10.1109/34.709601

8. *Freund Y., Schapire R.E.* A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting // Journal of Computer and System Sciences. 1997. Vol. 55. P. 119–139. https://doi.org/10.1006/jcss.1997.1504

9. *Friedman J.H.* Stochastic Gradient Boosting // Computational Statistics & Data Analysis. 2002. Vol. 38, No. 4. P. 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

10. *Zhou Z.H.* Ensemble Methods: Foundations and Algorithms. New York: Chapman and Hall/CRC, 2012. 446 p. ISBN 978-1-4398-3003-1.

11. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer, 2009. 745 p. https://doi.org/10.1007/978-0-387-84858-7

12. *Beja-Battais P.* Overview of AdaBoost: Reconciling its Views to Better Understand its Dynamics // arXiv preprint arXiv:2310.18323 [cs.LG]. 2023. https://doi.org/10.48550/arXiv.2310.18323

13. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794. https://doi.org/10.48550/arXiv.1603.02754

14. *Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y.* LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems (NeurIPS). 2017. Vol. 30.

15. *Hancock J.T., Khoshgoftaar T.M.* CatBoost for big data: an interdisciplinary review // Journal of Big Data. 2020. Vol. 7, No. 94. 45 p. https://doi.org/10.1186/s40537-020-00369-8

16. *Zhuravlev Yu.I., Senko O.V., Dokukin A.A., Kiselyova N.N., Saenko I.A.* Two-Level Regression Method Using Ensembles of Trees with Optimal Divergence // Doklady Mathematics. 2021. Vol. 103, No. 1. P. 1–4.
https://doi.org/10.1134/S1064562421040177

17. *Dokukin A.A., Sen'ko O.V.* A New Two-Level Machine Learning Method for Evaluating the Real Characteristics of Objects // Journal of Computer and Systems Sciences International. 2023. Vol. 62, No. 4. P. 607–614.
https://doi.org/10.1134/S1064230723040020

18. *Senko O.V., Dokukin A.A., Kiselyova N.N., Dudarev V.A., Kuznetsova Yu.O.* New Two-Level Ensemble Method and Its Application to Chemical Compounds Properties Prediction // Lobachevskii Journal of Mathematics. 2023. Vol. 44, No. 1. P. 188–197. https://doi.org/10.1134/S1995080223010341

19. *Geurts P., Ernst D., Wehenkel L.* Extremely Randomized Trees // Machine Learning. 2006. Vol. 63, No. 1. P. 3–42. https://doi.org/10.1007/s10994-006-6226-1

20. *López-Iñesta E., Grimaldo F., Arevalillo-Herráez M.* Combining feature extraction and expansion to improve classification-based similarity learning // Pattern Recognition Letters. 2016. Vol. 85. P. 84–90.
https://doi.org/10.1016/j.patrec.2016.11.005

21. *Breiman L., Friedman J., Olshen R.A., Stone C.J.* Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole, 1984. 358 p.
https://doi.org/10.1201/9781315139470

22. *Mahalanobis P.C.* On the Generalised Distance in Statistics (reprint of 1936) // Sankhya A. 2018. Vol. 80, Suppl. 1. P. 1–7.
https://doi.org/10.1007/s13171-019-00164-5

СВЕДЕНИЯ ОБ АВТОРАХ

МА Синьюэ – студентка 2 курса магистратуры кафедры «Математические методы прогнозирования» факультета ВМК, МГУ имени М.В. Ломоносова. Область научных интересов – машинное обучение и интеллектуальный анализ данных. Число научных публикаций – 2.

Xinyue MA – second-year Master's student of Computational Mathematics and Cybernetics, Lomonosov Moscow State University. Research interests: machine learning and data analysis. Number of publications – 2.

email: xinyuema35@gmail.com

ORCID*:* 0009-0004-0356-9996

СЕНЬКО Олег Валентинович – ведущий научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук. Область научных интересов – машинное обучение и интеллектуальный анализ данных. Число научных публикаций – более 200.

Oleg Valentinovich SEN'KO – leading Researcher at the Federal Research Center "Informatics and Control" of the Russian Academy of Sciences. Research interests: machine learning and data mining. The number of publications – over 200.

email: OSenko@frccsc.ru

ORCID*:* 0000-0002-5586-3503

Материал поступил в редакцию 2 ноября 2025 года