

УДК 004.85

ДЕРЕВЬЯ РЕШЕНИЙ С ЛИНЕЙНОЙ КОМБИНАЦИЕЙ ПРИЗНАКОВ

А. А. Рогов*, К. А. Кулаков, Н. Д. Москин

*Институт математики и информационных технологий,
Петрозаводский государственный университет (пр. Ленина, 33,
Петрозаводск, Республика Карелия, Россия, 185910), *rogov@petrsu.ru*

Дерево решений — это один из алгоритмов машинного обучения, у которого есть ряд преимуществ (хорошая интерпретация, возможность работы с разными типами данных, автоматическое формирование правил и пр.). В некоторых задачах, где набор данных ограничен, а признаки связаны между собой (например, в задаче атрибуции текстов), можно при построении условий в узлах модели использовать не одиночные признаки, а их линейные комбинации. С учетом этого в статье изложено развитие метода «дерево решений», которое заключается в том, что к старым k признакам добавляются новые признаки как линейные комбинации двух исходных: $x_{ij} = \alpha x_i \pm (1 - \alpha)x_j$, где $i, j = 1, \dots, k$ и параметр $\alpha \in [0, 1]$. Для проведения численных экспериментов на примере задачи атрибуции текстов выполнена реализация построения линейных комбинаций признаков в информационной системе СМАЛТ («Статистические методы анализа литературного текста»). Результаты классификации дореволюционных текстов из журналов «Время», «Эпоха» и еженедельника «Гражданин» с использованием разных видов n -грамм показали, что данное улучшение повышает точность метода, при этом несущественно снижает интерпретацию полученного результата.

Ключевые слова: машинное обучение; дерево решений; линейная комбинация признаков; атрибуция текстов; n -граммы; информационная система СМАЛТ

Для цитирования: Рогов А. А., Кулаков К. А., Москин Н. Д. Деревья решений с линейной комбинацией признаков // Труды Карельского научного центра РАН. 2026. № 6. С. 100–105. doi: 10.17076/mat2314

A. A. Rogov*, K. A. Kulakov, N. D. Moskin. DECISION TREES WITH A LINEAR COMBINATION OF ATTRIBUTES

*Institute of Mathematics and Information Technologies, Petrozavodsk State University (33 Lenin St., 185910 Petrozavodsk, Karelia, Russia), *rogov@petrsu.ru*

A decision tree is a machine learning algorithm featuring a number of advantages (good interpretability, the ability to work with different types of data, automatic rule generation, etc.). In some tasks, where the data set is limited and attributes are interrelated (e.g., in the text attribution task), linear combinations of attributes can be used instead of single attributes when constructing conditions at the model's nodes. Taking this into account, the article presents an advancement of the decision

tree method, which consists in supplementing k old attributes with new attributes as linear combinations of two original ones: $x_{ij} = \alpha x_i \pm (1 - \alpha)x_j$, where $i, j = 1, \dots, k$ and the parameter $\alpha \in [0, 1]$. To conduct numerical experiments, the construction of linear combinations of attributes was implemented in the SMALT information system («Statistical Methods for the Analysis of Literary Text») for the case of the text attribution task. The results of classifying pre-1917 texts from the «Vremya» and «Epokha» magazines, and the weekly edition «Grazhdanin» using different types of n -grams showed that this improvement increases the accuracy of the method, but does not significantly reduce the interpretation of the outcome.

Key words: machine learning; decision tree; linear combination of attributes; text attribution; n -grams; SMALT information system

For citation: Rogov A. A., Kulakov K. A., Moskin N. D. Decision trees with a linear combination of attributes. *Trudy Karelskogo nauchnogo tsentra RAN = Transactions of the Karelian Research Centre RAS*. 2026. No. 6. P. 100–105. doi: 10.17076/mat2314

ВВЕДЕНИЕ

Дерево решений – это алгоритм машинного обучения, основанный на правиле «если <условие>, то <результат>». Представляет собой иерархическую структуру, состоящую из узлов и листьев, где каждый узел – это проверка условия, а листья – конечные результаты. Сама идея создания и дальнейшего развития моделей дерева решений появилась в середине XX века и была развита в работах К. Ховеленда «Компьютерное моделирование мышления» [10] и Е. Ханта «Эксперименты по индукции» [11]. Данным методом решаются следующие задачи: классификация данных, прогнозирование и выявление значимых факторов. Сегодня деревья решений используются в финансовой сфере, медицине, клиентском сервисе, промышленности и других областях [2, 5, 8]. Они помогают прогнозировать риски, диагностировать заболевания, оптимизировать процессы и принимать решения на основе больших объемов данных. Деревья решений можно применять в задачах атрибуции текстов (например, при решении вопросов их принадлежности тому или иному автору) [7]. Однако при этом не всегда удается добиться желаемого качества распознавания, так как объем текстов может быть весьма ограничен.

Этот метод обладает рядом преимуществ:

- простота интерпретации результатов (под интерпретацией результатов в машинном обучении понимают объяснение того, как модель пришла к своим прогнозам и какие факторы на них повлияли. Это нужно, чтобы сделать «черный ящик» модели понятным для людей: аналитиков, бизнеса, регуляторов);
- работа с разными типами данных;

- не требует нормализации данных;
- автоматическое формирование правил;
- быстрая скорость обучения (порядок $O(n \cdot \ln(n) \cdot d)$, где n – число объектов, d – число признаков).

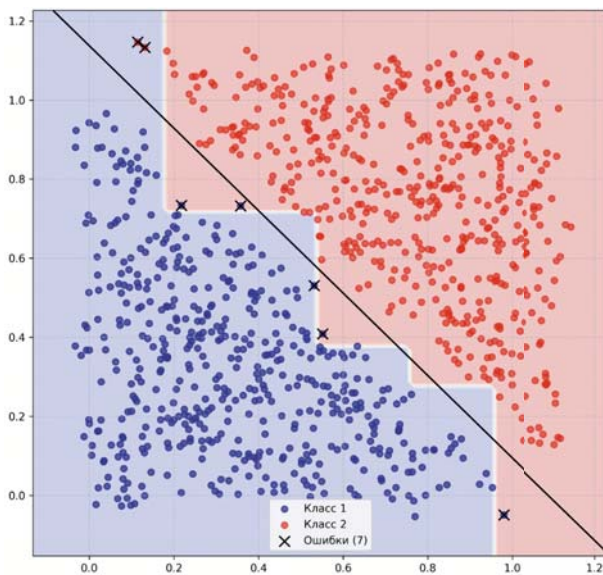
Как правило, при делении анализируемых данных в узле, содержащем <условие>, по двум веткам используется разделитель в виде простого правила, которое сравнивает некоторое значение признака с заданным порогом. Деревья решений с более сложными условиями рассмотрены в [1, 12]. Помимо одномерных (*univariate decision trees*) существует описание многомерных деревьев решений (*multivariate decision trees*). Во втором случае условие содержит более чем один признак, а именно лучшую линейную комбинацию признаков. Для ее поиска могут использоваться нейронная сеть и градиентный метод. В данной работе показано, что можно обойтись более простыми способами.

К основным существенным недостаткам указанного метода следует отнести проблему точности. Для борьбы с этой проблемой предлагают использовать набор деревьев (леса деревьев решений) [7]. В этом случае ухудшается интерпретация результатов или она становится вообще невозможной. В данной статье предлагается развитие метода «дерево решений», увеличивающее точность классификации за счет аугментации исходных данных.

МАТЕМАТИЧЕСКОЕ ОПИСАНИЕ ДЕРЕВА РЕШЕНИЙ

В качестве исходных данных для поставленной задачи используются размеченные пары $\{(z_l, y_l)\}_{l=1}^n$, где n – число объектов, z_l –

признаковое описание l -го объекта (точка в k -мерном пространстве $z_l = (x_1^l, x_2^l, \dots, x_k^l)$), а y_l – метка класса. На искусственно сгенерированном авторами демонстрационном примере (рис.) видно, что дерево решений в качестве классификатора строит прямоугольные границы решений (все расщепления перпендикулярны той или иной оси), при этом допущено 7 ошибок. На самом деле это простой линейно сепарабельный набор данных. То есть существует линейная разделяющая прямая, которая непараллельна осям координат (на рисунке пример такой прямой изображен сплошной линией). Если ее использовать в качестве классификатора, то ошибок не будет.



Графическое представление результата работы алгоритма «дерево решений»

Graphical representation of the result of the «decision tree» algorithm

Линейный классификатор даст всего одно условие (точка на плоскости выше или ниже этой линии), а классификатор, основанный на дереве решений, вынуждает проводить в худшем случае проверку 9 условий, что существенно удлиняет объяснение того, какая точка к какому классу относится, т. е. ухудшает интерпретацию.

Предлагаемый подход позволяет при той же точности уменьшить глубину построения решения классификатора, а значит, сохранить главное достоинство – простоту объяснения результатов разбиения (подробнее о проблемах, возникающих при построении моделей машинного обучения, см. в [3]). Для этого увеличим размерность обучающей выборки следующим образом. Добавим к старым k признакам новые признаки как линейные комби-

нации двух исходных: $x_{ij} = \alpha x_i \pm (1 - \alpha)x_j$, где $\alpha \in [0, 1]$. Этот подход был впервые применен в работе [9]. Если α равно 0 или 1, получаем старые признаки. При этом параметр α для каждой пары признаков будет свой.

В качестве значений α можно брать равномерные значения из интервала $(0, 1)$. Пусть их будет m . Тогда общее количество новых признаков будет равно $2mk(k - 1)$. Если добавить исходные признаки, то их будет не больше, чем $2mk(k - 1) + k$ с учетом различных значений α . Как отмечено выше, сложность данного алгоритма линейно зависит от числа признаков, значит, для данного метода увеличение размерности числа признаков не критично, так как существующие программные реализации построения деревьев решений способны обрабатывать массивы данных высокой размерности, а также выполнять параллельные вычисления.

ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ

Для оценки данного подхода рассмотрим задачу атрибуции текстов и фрагментов текстов из журналов «Время», «Эпоха» и еженедельника «Гражданин» под редакторством Ф. М. Достоевского [4]. Известно, что ряд анонимных публикаций принадлежат Ф. М. Достоевскому, а также он в ходе редактуры вносил правки в статьи других авторов, поэтому перед филологами стоит задача классификации текстов на две группы: «тексты, написанные Ф. М. Достоевским» и «тексты других авторов».

Определим анализируемый текст как последовательность из N фрагментов, каждый фрагмент состоит из M слов. Для унификации с каждым словом сопоставим признак части речи (атрибут), обозначим их число за k . В результате получаем N фрагментов, каждый из которых состоит из M значений атрибутов.

Рассмотрим следующие способы признакового описания фрагмента текста:

1. Униграммы частей речи. Построим вектор $z_l = (x_1^l, x_2^l, \dots, x_k^l)$, где x_i^l – частота встречаемости атрибута i в фрагменте l (признак i), $n = 1, \dots, N$. В результате получаем для текста набор из N векторов размерности k .

2. Линейные комбинации униграмм. Построим вектор $z_l^\alpha = (x_{11}^l, y_{11}^l, x_{12}^l, y_{12}^l, \dots, x_{kk}^l, y_{kk}^l)$, где $x_{ij}^l = \alpha x_i^l + (1 - \alpha)x_j^l$, $y_{ij}^l = \alpha x_i^l - (1 - \alpha)x_j^l$ – линейные комбинации частоты встречаемости униграмм i и j в фрагменте n . Таким образом получаем для текста набор из N векторов размерности $k \times (k - 1) \times 2$ для одного значения α .

В результате получаем набор сопоставимых векторов одинаковой размерности.

Для проведения численного эксперимента выполнена реализация построения линейных комбинаций признаков в информационной системе СМАЛТ («Статистические методы анализа литературного текста») [7]. Подсистема состоит из модуля управления деревьями решений (Tree app), интерфейса пользователя (Tree UI) и сервиса для их построения (Tree worker service).

Модуль управления реализован в виде Django-приложения и предоставляет функциональность по добавлению, изменению, удалению и просмотру деревьев решений и алгоритм проверки текста в дереве решений. Для построения дерева решений необходимо указать следующие параметры:

1. название исследования;
2. перечень текстов первого автора(ов);
3. перечень текстов второго автора(ов);
4. размер блока слов;
5. глубина дерева решений;
6. используемые признаки, описание которых изложено ранее.

Сервис построения деревьев решений реализован в виде самостоятельного Python-приложения, запускается в виде демона и реализует алгоритм подготовки данных и построения деревьев решений.

Алгоритм построения деревьев решений состоит из следующих шагов:

1. Определение перечня признаков. В системе реализована разметка текстов с использованием 23 атрибутов [6]. При построении дерева решений пользователь может исключить некоторые из них из рассмотрения.
2. Построение частотных матриц для перечней текстов. Последовательность текстов разбивается на блоки слов заданной длины. Для каждого блока вычисляется распределение частот встречаемости признаков в виде строки матрицы. Число столбцов определяется числом признаков.
3. Расчет частотных характеристик линейных комбинаций атрибутов.
4. Построение деревьев решений. Для построения используется модуль RandomForestClassifier пакета sklearn. Для получения наилучших решений в настройках метода указывается параметр числа попыток ($n_estimators$) не

менее 100 или числа строк матриц, если их больше; глубина дерева решений (max_depth) и отключение построения попыток на меньшем наборе данных ($bootstrap = False$). После построения выполняется оценка точности работы дерева решений на исходных данных (методы $predict()$ и $accuracy_score()$). Так как построение деревьев решений выполняется случайным образом, то выполняется 10 запусков с выбором наилучшего результата по точности.

5. Сохранение результата алгоритма. В базу данных информационной системы сохраняется в бинарном виде дерево решений и статистики.

Для классификации текста выполняется разбиение текста на блоки, расчет частотных характеристик каждого блока и для каждого полученного вектора выполняется оценка принадлежности блока к первой или второй группе текстов с помощью выбранного дерева решений.

Ключевой особенностью расчета частотных характеристик является построение блока текстов. На вход подаются тексты в виде потока слов. Слово пропускается, если его атрибут не входит в перечень требуемых. Блок считается завершенным, если число обработанных слов совпадает с размером блока.

Для проведения эксперимента использовались следующие параметры:

- размер блока – 200 слов;
- глубина дерева решений – 5 узлов;
- первая группа текстов – 50 текстов, подписанных Ф. М. Достоевским, из журналов «Время», «Эпоха» и еженедельника «Гражданин» [6];
- вторая группа текстов – 60 текстов, подписанных другими авторами (А. А. Григорьев, Я. П. Полонский, М. М. Достоевский, Н. Н. Прокопович, К. К. Сунгуров, Н. Н. Страхов, В. П. Мещерский, К. П. Победоносцев, В. Ф. Пуцкович), из журналов «Время», «Эпоха» и еженедельника «Гражданин» [6];
- число блоков – 873;
- используемые атрибуты – существительное, прилагательное, числительное, местоимение, глагол, причастие, деепричастие, наречие, категория состояния, частица, предлог, союз, модальное слово, междометие, звукоподражательное слово, иностранное слово, вводное слово.

В ходе эксперимента выполнялась оценка точности деревьев решений при заданных параметрах (подробнее о методике деления текстов на обучающую и тестовую выборку можно ознакомиться в [7]). Результаты экспериментов представлены в таблице. Показатели «точность» (*precision*) (или «точность положительных предсказаний», эта метрика отвечает на вопрос: «Из всех случаев, когда модель

сказала “это положительный класс”, сколько раз она была права?») и «F1-мера» (гармоническое среднее между точностью и полнотой). Полнота показывает, какую долю реальных положительных случаев модель сумела найти. Они широко используются в сфере искусственного интеллекта (подробнее о метриках качества моделей машинного обучения см. в [3]).

Результаты экспериментов (оценка точности деревьев решений при заданных параметрах)
Experimental results (evaluation of the accuracy of decision trees for given parameters)

Признаки Attributes	размер вектора vector size	точность (сред.) accuracy (avg.)	точность (макс.) accuracy (max.)	F1-мера F1-score
Униграммы частей речи Unigrams of parts of speech	17	0,816	0,821	0,824
Линейные комбинации униграмм с $\alpha = 0; 0,25; 0,5$ Linear combinations of unigrams with $\alpha = 0; 0,25; 0,5$	1632	0,853	0,856	0,863

Анализируя таблицу, можно сделать вывод, что использование классификатора на основе линейных комбинаций униграмм показывает более высокую точность по сравнению с классификатором на основе униграмм. Если построить доверительные интервалы для средних значений точности на уровне значимости 0,001, то они будут соответственно равны (0,815; 0,817) и (0,852; 0,854). Можно считать, что различия между показателями в таблице являются статистически значимыми.

Полученные деревья решений можно посмотреть по ссылке на информационную систему СМАЛТ [6].

ЗАКЛЮЧЕНИЕ

В работе предлагается развитие метода «дерево решений», которое можно использовать в том числе для задачи атрибуции текстов. При построении условий в узлах модели предлагается использовать не одиночные признаки, а линейные комбинации признаков. Данное усовершенствование при фиксированной глубине позволяет добиться большей точности, а при заданной точности уменьшает число узлов в классификационном дереве решений, а значит, улучшает интерпретацию полученного результата. Заметим, что предлагаемая аугментация признаков для построения дерева решений для примера, показанного на рисунке, приведет к линейной классификации.

ЛИТЕРАТУРА

1. Девяткин Д. А. Построение ансамблей деревьев решений с использованием линейных и нели-

нейных разделителей: Дис... канд. физ.-мат. наук. М.; 2022. 115 с.

2. Жаркова О. С., Шаропин К. А., Сеидова А. С., Берестнева Е. В., Осадчая И. А. Построение систем поддержки принятия решений в медицине на основе деревьев решений // Современные наукоемкие технологии. 2016. № 6. С. 33–37.

3. Жерон О. Прикладное машинное обучение с помощью Scikit-Learn, Keras и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем. СПб.: Диалектика, 2020. 1040 с.

4. Захаров В. Н. Задачи атрибуции анонимных статей в изданиях Достоевского // Известный Достоевский. 2022. № 9(3). С. 5–24. doi: 10.15393/j10.art.2022.6221

5. Иванов В. В., Акишина Е. П., Приказчикова А. С. Деревья и леса решений в задаче классификации кредитных организаций // Письма в журнал «Физика элементарных частиц и атомного ядра». 2025. № 22(1). С. 62–74.

6. Информационная система СМАЛТ («Статистические методы анализа литературного текста») [Электронный ресурс]. URL: http://smalt.karelia.ru/pysmalt/research/r_tree/list (дата обращения: 10.02.2026).

7. Rogov A. A., Abramov P. V., Burchneva D. D., Zakharova O. V., Kulakov K. A., Lebedev A. A., Moskin N. D., Otlivanichik A. V., Savinov E. D., Sidоров Ю. В. Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин». Петрозаводск: Острова, 2021. 391 с.

8. Франчук В. В., Михайличенко Б. В., Франчук М. В. Применение метода дерева решений в судебно-медицинской экспертной практи-

ке при анализе «врачебных дел» // Судебно-медицинская экспертиза. 2020. № 63(1). С. 9–14. doi: 10.17116/sudmed2020630119

9. Friedman J. H. A recursive partitioning decision rule for nonparametric classification // IEEE Transactions on Computers. 1977. No. 26(4). P. 404–408. doi: 10.1109/TC.1977.1674849

10. Hovland C. I. Computer simulation of thinking // Am. Psychol. 1960. No. 15(11). P. 687–693. doi: 10.1037/h0044165

11. Hunt E. B., Marin J., Stone P. J. Experiments in induction. New York: Academic Press, 1966. 247 p.

12. Kotsiantis S. B. Decision trees: a recent overview // Artif. Intell. Rev. 2013. No. 39(4). P. 261–283. doi: 10.1007/s10462-011-9272-4

REFERENCES

1. Devyatkin D. A. Constructing ensembles of decision trees using linear and nonlinear separators: PhD (Cand. of Phys.-Math.) thesis. Moscow; 2022. 115 p. (In Russ.)

2. Zharkova O. S., Sharopin K. A., Seidova K. S., Berestneva E. V., Osadchaya I. A. Construction of decision support systems in medicine based on decision tree. *Sovremennye naukoemkie tekhnologii = Modern Science-Intensive Technologies*. 2016;6:33–37. (In Russ.)

3. Geron A. Hands-on machine learning with Scikit-Learn, Keras and Tensor Flow: concepts, tools, and techniques to build intelligent systems. St. Petersburg: Dialektika; 2020. 1040 p. (In Russ.)

4. Zakharov V. N. Problems of attribution of anonymous articles in Dostoevsky's publications. *Neizvestnyi Dostoevskii = The Unknown Dostoevsky*. 2022; 9(3):5–24. (In Russ.). doi: 10.15393/j10.art.2022.6221

5. Ivanov V. V., Akishina E. P., Prikazchikova A. S. Decision trees and forests in the problem of classification of credit institutions. *Pis'ma v zhurnal «Fizika elementarnykh chastits i atomnogo yadra» = Letters to the journal 'Physics of Elementary Particles and Atomic Nuclei'*. 2025;22(1):62–74. (In Russ.)

6. SMALT information system («Statistical Methods for the Analysis of Literary Text») (In Russ.). URL: http://smalt.karelia.ru/pysmalt/research/r_tree/list (accessed: 10.02.2026).

7. Rogov A. A., Abramov R. V., Buchneva D. D., Zaharova O. V., Kulakov K. A., Lebedev A. A., Moskin N. D., Otlivanchik A. V., Savinov E. D., Sidorov Yu. V. The problem of attribution in the magazines 'Time', 'Epoch' and the weekly 'Citizen'. Petrozavodsk: Ostrova; 2021. 391 p. (In Russ.)

8. Franchuk V. V., Mikhaylichenko B. V., Franchuk M. V. Application of the decision tree method in forensic medical practice in the analysis of medical cases. *Sudebno-meditsinskaya ekspertiza = Forensic Medical Expertise*. 2020;63(1):9–14. (In Russ.). doi: 10.17116/sudmed2020630119

9. Friedman J. H. A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*. 1977;26(4):404–408. doi: 10.1109/TC.1977.1674849

10. Hovland C. I. Computer simulation of thinking. *Am. Psychol.* 1960;15(11):687–693. doi: 10.1037/h0044165

11. Hunt E. B., Marin J., Stone P. J. Experiments in induction. New York: Academic Press; 1966. 247 p.

12. Kotsiantis S. B. Decision trees: a recent overview. *Artif. Intell. Rev.* 2013;39(4):261–283. doi: 10.1007/s10462-011-9272-4

Поступила в редакцию / received: 23.02.2026; принята к публикации / accepted: 17.06.2026.

Авторы заявляют об отсутствии конфликта интересов / The authors declare no conflict of interest.

СВЕДЕНИЯ ОБ АВТОРАХ:

Рогов Александр Александрович
д-р техн. наук, заведующий кафедрой
e-mail: rogov@petrsu.ru

Кулаков Кирилл Александрович
канд. физ.-мат. наук, доцент
e-mail: kulakov@cs.petrsu.ru

Москин Николай Дмитриевич
д-р техн. наук, профессор
e-mail: moskin@petrsu.ru

CONTRIBUTORS:

Rogov, Aleksandr
Dr. Sci. (Tech.), Head of Department

Kulakov, Kirill
Cand. Sci. (Phys.-Math.), Associate Professor

Moskin, Nikolai
Dr. Sci. (Tech.), Professor