О выразительных возможностях ансамблей решающих деревьев

А. П. Соколов [1], Л. А. Прохоренкова [2]

Решающие деревья широко применяются в машинном обучении, статистике и анализе данных. Предиктивные модели, основанные на решающих деревьях, показывают отличные результаты в терминах точности и времени обучения, особенно на гетерогенных табличных датасетах. Производительность, простота и надежность делают это семейство алгоритмов одним из наиболее популярных в машинном обучении и науке о данных.

Одним из важных гиперпараметров алгоритмов, основанных на решающих деревьях, является максимальная глубина.

В данной работе получен теоретический результат, который показывает как ограничение на максимальную глубину решающих деревьев влияет на выразительные возможности всего ансамбля. Этот результат применим к таким алгоритмам, как одиночное решающее дерево (Decision Tree), случайный лес (Random Forest), градиентный бустинг (GBDT) и другие.

Ключевые слова: машинное обучение, наука о данных, решающее дерево, случайный лес, градиентный бустинг.

1. Введение

Решающие деревья широко применяются в машинном обучении, статистике и анализе данных. Предиктивные модели, основанные на решающих деревьях, показывают отличные результаты в терминах точности и времени обучения, особенно на гетерогенных табличных датасетах ([1]). Производительность, простота и надежность делают это семейство алгоритмов одним из наиболее популярных в машинном обучении и науке о данных ([2]).

Одним из важных гиперпараметров алгоритмов, основанных на решающих деревьях, является максимальная глубина.

В данной работе получен теоретический результат, который показывает как ограничение на максимальную глубину решающих деревьев

[1] *Соколов Андрей Павлович* — к.ф.-м.н., м.н.с. каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: sokolov@intsys.msu.ru

Sokolov Andrey Pavlovich — junior scientific researcher, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

[2] *Прохоренкова Людмила Александровна* — д.ф.-м.н., исследователь, Яндекс, e-mail: ostroumova-la@yandex.ru

Prokhorenkova Liudmila Alexandrovna — researcher, Yandex.

влияет на выразительные возможности всего ансамбля. Этот результат применим к таким алгоритмам, как одиночное решающее дерево (Decision Tree), случайный лес (Random Forest), градиентный бустинг (GBDT) и другие.

2. Определения и основной результат

Предположим задан датасет из n примеров по m признаков.

$$D = \{(X_i, y_i) \,|i \in \{1, ..., n\}\,, X_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}\,.$$

В задачах машинного обучения обычно X_i называется *вектором признаков*, а y_i – *значением*.

Будем использовать обозначение $D\,(X) = y$, где $(X, y) \in D$.

Задача регрессии в машинном обучении обычно ставится следующим образом: необходимо найти такую функцию $\Phi : \mathbb{R}^m \to \mathbb{R}$, которая бы минимизировала некоторую функцию потерь $\mathcal{L}$ на датасете D. Функция Φ называется *моделью*.

В данной работе для простоты мы намеренно не рассматриваем вопросы, связанные с обобщающей способностью модели. Поэтому мы не разбиваем датасет на обучающую и валидационную выборки, как это принято при решении прикладных задач машинного обучения. Единственный вопрос, который мы поднимаем, – насколько точно может описывать датасет модель, основанная на ансамбле решающих деревьев.

Обычно функция потерь (лосс-функция) $\mathcal{L}$ может быть представлена как сумма элементарных функций потерь на отдельных примерах $\mathcal{L}\,(D, \Phi) = \sum\limits_{i=1}^{n} l(D\,(X_i)\,, \Phi\,(X_i))$, где $l : \mathbb{R}^2 \to \mathbb{R}$ – элементарная функция потерь. В данной работе мы рассматриваем только такие лосс-функции, которые могут быть представлены как сумма элементарных функций потерь на индивидуальных примерах.

Различные элементарные функции потерь используются на практике: L_1, L_2 и другие. В данной статье мы будем рассматривать только такие элементарные функции потерь l, для которых $l\,(a, b) = 0 \iff a = b$. Для таких функций потерь $\mathcal{L}\,(D, \Phi) = 0$ тогда и только тогда, когда $D\,(X_i) = \Phi\,(X_i)$ для всех $i \in \{1, ..., n\}$, и $\mathcal{L}\,(D, \Phi) > 0$ если существует такое $i \in \{1, ..., n\}$, что $D\,(X_i) \neq \Phi\,(X_i)$.

Будем говорить, что модель Φ *идеально аппроксимирует* датасет D, тогда и только тогда, когда $\mathcal{L}\,(\Phi, D) = 0$. Идеальная аппроксимация означает, что выход модели $\Phi\,(X_i)$ равен истинному значению y_i на всех примерах X_i из датасета.

Ансамбль Φ из k деревьев имеет вид $\Phi(X) = \sum_{i=1}^{k} t_i(X)$, $t_i \in \mathbb{T}$, где t_i – индивидуальные решающие деревья из множества всех возможных решающих деревьев – $\mathbb{T}$.

Решающее дерево t работает следующим образом. Входной вектор подается на корневую вершину дерева. Каждая внутренняя вершина s содержит предикат вида $p_s(X) = (x_{i_s} \geq r_s)$, где x_{i_s} – компонента вектора признаков с индексом i_s, а r_s – некоторое пороговое значение, ассоциированное с вершиной. Предикаты внутренних вершин дерева последовательно решают, куда должен спускаться вектор признаков X. В конце концов X спускается до листа дерева с индексом q и весом w_q. Этот вес определяет выходное значение дерева на данном векторе признаков, то есть $t(X) = w_q$.

Максимальная глубина дерева d определяет максимальную длину пути (число внутренних вершин с предикатами), который ведет от корневой вершины к листу. Заметим, что максимальное число листьев N решающего дерева и максимальная глубина d связаны соотношением $N \leq 2^d$.

Видно, что каждое решающее дерево t максимальной глубины d задает кусочно-константную функцию, которая отображает $\mathbb{R}^m$ в $\mathbb{R}$. При этом число константных областей в пространстве признаков ограничено сверху как 2^d. Следовательно, каждое решающее дерево t может быть представлено следующим образом $t(X) = \sum_{q=1}^{N_t} I_q(X) \cdot w_q$, где q пробегает N_t листьев дерева t, $I_q(X)$ – индикаторная функция листа с индексом q

$$I_q(X) = \begin{cases} 1, \text{ если } X \text{ принадлежит листу } q; \\ 0, \text{ иначе;} \end{cases}$$

$w_q \in \mathbb{R}$ – вещественные веса, ассоциированные с листьями дерева.

Каждая индикаторная функция $I_q(X)$ принимает значение 1 в области пространства признаков, которая определена набором не более, чем из d ограничений вида $a_i \leq x_i \leq b_i$, где $a_i, b_i \in \mathbb{R} \cup \{-\infty, +\infty\}$. Обозначим множество таких индикаторных функций как $\mathbb{I}_d$.

Имеет место следующее утверждение.

Теорема. *Для всякого $d > 1$ существует датасет D_d такой, что:*

1) *он не может быть идеально аппроксимирован никаким ансамблем решающих деревьев Φ с максимальной глубиной $(d-1)$;*

2) *он может быть идеально аппроксимирован одним решающим деревом t глубины d.*

Доказательство. Для начала рассмотрим представление решающего дерева t с помощью его индикаторных функций $t(X) = \sum_{q=1}^{N_t} I_q(X) \cdot w_q$.

Следовательно, всякий ансамбль Φ может быть представлен следующим образом $\Phi(X) = \sum\limits_{q=1}^{N_\Phi} I_q(X) \cdot w_q$, где N_Φ – общее число листьев в ансамбле Φ. Обозначим $\Phi(X) = \sum\limits_{q=1}^{N_\Phi} \Phi_q(X)$, где $\Phi_q(X) = I_q(X) \cdot w_q$ – элементарная предиктивная функция, реализуемая листом q одного из деревьев ансамбля.

Заметим, что каждая индикаторная функция решающего дерева с максимальной глубиной $(d-1)$ может быть представлена следующим образом $I_q(X) = \mathop{\&}\limits_{i=1}^{d-1} p_{s_i}^{\sigma_{s_i}}(X)$, где & – операция конъюнкции (логическое «И»), вершины $s_1, ..., s_{d-1}$ соответствуют пути в дереве от корня до листа q и

$$p_{s_i}^{\sigma_{s_i}} = \begin{cases} p_{s_i},\ \sigma_{s_i} = 1; \\ \bar{p}_{s_i},\ \text{иначе.} \end{cases}$$

Здесь $\bar{p}_{s_i}$ означает логическое отрицание p_{s_i}. Также $\sigma_{s_i} = 1$ если соответствующий предикат в вершине s_i равен 1 и $\sigma_{s_i} = 0$ иначе.

Теперь мы можем построить датасет, который не может быть аппроксимирован никаким ансамблем с деревьями максимальной глубины $(d-1)$. Рассмотрим булеву функцию $f(x_1, ..., x_d) = \sum\limits_{i=1}^{d} x_i$, где используется сложение по модулю 2. Положим

$$D_d = \left\{(x_1, ..., x_d, -1) \,|\, (x_1, ..., x_d) \in \{0,1\}^d, f(x_1, ..., x_d) = 0\right\} \cup$$

$$\left\{(x_1, ..., x_d, +1) \,|\, (x_1, ..., x_d) \in \{0,1\}^d, f(x_1, ..., x_d) = 1\right\}.$$

Занумеруем все значения в датасете D_d в единый вектор Y длины 2^d. Обозначим $Y(x_1, ..., x_d) = y$ для всякого вектора $(x_1, ..., x_d, y) \in D_d$. Далее выпишем в таком же порядке значения, предсказанные ансамблем Φ на примерах из датасета, и получим вектор Φ. Аналогичным образом введем в рассмотрение Φ_q – вектор предсказаний q-го листа ансамбля на всех примерах датасета D. Заметим, что D_d содержит элементы для всех возможных наборов признаков $(x_1, ..., x_d)$ и, таким образом, его размер равен 2^d. Далее, $Y(x_1, ..., x_j, ...x_d) = -Y(x_1, ..., \bar{x}_j, ...x_d)$ для всякого $j \in \{1, ..., d\}$. Заметим, что мы можем применить логическое отрицание к признакам, потому что они принадлежат множеству $\{0,1\}$. Рассмотрим скалярное произведение $\langle Y, \Phi \rangle$ в евклидовом пространстве размерности 2^d. Если ансамбль Φ идеально аппроксимирует D_d, тогда $\langle Y, \Phi \rangle = \|Y\|^2$. Следовательно, если $\langle Y, \Phi \rangle \neq \|Y\|^2$,

то Φ аппроксимирует D_d не идеальным образом. Обратим внимание, что $\Phi = \sum_{q=1}^{N_\Phi} \Phi_q$, где Φ_q соответствует вектору длины 2^d, который равен w_q на тех позициях, где индикаторная функция I_q принимает значение 1. На других позициях вектор Φ_q равен нулю. Следовательно, $\langle Y, \Phi \rangle = \sum_{q=1}^{N_\Phi} \langle Y, \Phi_q \rangle$. Заметим, что индикаторная функция $I_q \in \mathbb{I}_{d-1}$ может быть представлена как конъюнкция предикатов размера $(d-1)$, следовательно, она существенно зависит не более, чем от $(d-1)$ признака. Таким образом, для каждой индикаторной функции I_q найдется хотя бы одна переменная x_i такая, что I_q не зависит от нее существенным образом. То есть $I_q(x_1, ..., x_j, ...x_d) = I_q(x_1, ..., \bar{x}_j, ...x_d)$. Следовательно, для каждого листа $q \in \{1, ..., N_\Phi\}$ найдется такое $j \in \{1, ..., d\}$, что $\Phi_q(x_1, ..., x_j, ...x_d) = \Phi_q(x_1, ..., \bar{x}_j, ...x_d)$. В то же время, как было показано ранее, $Y(x_1, ..., x_j, ...x_d) = -Y(x_1, ..., \bar{x}_j, ...x_d)$. Из этого наблюдения следует, что $\langle Y, \Phi_q \rangle = 0$ для всякого листа ансамбля q. Следовательно $\langle Y, \Phi \rangle = 0$ для каждого ансамбля максимальной глубины $(d-1)$.

Теперь построим одно решающее дерево глубины d, которое будет идеально аппроксимировать D_d. В корневую вершину s_1 поместим предикат $p_{s_1}(X) = (x_1 \geq 0.5)$. Во всех вершинах, находящихся на глубине i в дереве, разместим предикаты $p_i(X) = (x_i \geq 0.5)$. Очевидно, что 2^d листьев построенного таким образом решающего дерева будут содержать все 2^d примеров датасета D_d. Теперь осталось лишь присвоить веса листьям в соответствии со значениями из датасета D_d. Таким образом построенное решающее дерево идеально аппроксимирует датасет D_d. Теорема доказана.

3. Заключение

В данной работе мы доказали, что произвольный ансамбль решающих деревьев максимальной глубины $(d-1)$ имеет меньшие выразительные возможности, чем одно решающее дерево глубины d.

Одним из следствий этого результата, которое имеет важное значение для прикладных задач машинного обучения, является то, что мы не можем компенсировать недостаток глубины решающих деревьев увеличением их количества в ансамбле.

4. Благодарности

Авторы выражают благодарность Сергею Иванову, Леониду Литовченко и Станиславу Моисееву за ряд ценных замечаний по данной работе.

Список литературы

[1] Vadim Borisov and Tobias Leemann and Kathrin Seßler and Johannes Haug and Martin Pawelczyk and Gjergji Kasneci, "Deep Neural Networks and Tabular Data: A Survey", *CoRR*, **abs/2110.01889** (2021), https://arxiv.org/abs/2110.01889.

[2] Iqbal H Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Comput Sci.*, **2(3)** (2021), 10.1007/s42979-021-00592-x.

On The Expressive Power of Decision Tree Ensembles
Sokolov A.P., Prokhorenkova L.A.

Decision trees are widely used in machine learning, statistics and data mining.

Predictive models based on decision trees show outstanding results in terms of accuracy and training time. Especially on heterogeneous tabular datasets. Performance, simplicity and integrity make this family of algorithms one of the most popular in data science.

One important hyper-parameter of decision tree training algorithms is maximum depth of the trees.

This paper proves theoretical result that shows how maximum depth of decision trees limits the expressive power of ensemble. This result is applicable to such tree based algorithms as plain Decision Tree, Random Forest, GBDT and others.

Keywords: machine learning, data science, decision tree, random forest, gradient boosting.

References

[1] Vadim Borisov and Tobias Leemann and Kathrin Seßler and Johannes Haug and Martin Pawelczyk and Gjergji Kasneci, "Deep Neural Networks and Tabular Data: A Survey", *CoRR*, **abs/2110.01889** (2021), https://arxiv.org/abs/2110.01889.

[2] Iqbal H Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions", *SN Comput Sci.*, **2(3)** (2021), 10.1007/s42979-021-00592-x.