

# Семинары по решающим деревьям

Евгений Соколов

10 октября 2013 г.

## 1 Методы стрижки деревьев

Описанный выше жадный метод строит дерево, идеально классифицирующее обучающую выборку, что зачастую приводит к низкой обобщающей способности алгоритма (то есть к его *переобучению*). Для повышения качества дерева производят его *стрижку* — удаление некоторых вершин. Опишем один из классических методов стрижки — *cost-complexity pruning* (см. [1], а также [2], гл. 7).

Выберем некоторый функционал качества  $R(t)$  для листовых вершин. Это могут быть:

- ошибка классификации  $R(t_m) = \frac{1}{N_m} \sum_{x_i \in R_m} [y_i \neq k_m]$ ;
- индекс Джини  $R(t_m) = \sum_{k \neq k'} p_{mk} p_{mk'}$ ;
- энтропия  $R(t_m) = - \sum_{k=1}^K p_{mk} \log_2 p_{mk}$ ;
- и т.д.

При стрижке, как правило, используется ошибка классификации [3]. Потребуем, чтобы функционал достигал своего минимума на вершинах, в которые попали объекты лишь одного класса (для указанных выше трех функционалов это верно). Определим качество  $R(T)$  поддерева  $T$  как сумму значений функционала качества во всех листьях этого поддерева.

Обозначим дерево, полученное в результате работы жадного алгоритма, через  $T_0$ . Поскольку в каждом из листьев находятся объекты только одного класса, значение функционала  $R(T)$  будет минимально на самом дереве  $T_0$  (среди всех поддеревьев). Однако данный функционал характеризует лишь качество дерева на обучающей выборке, и чрезмерная подгонка под нее может привести к переобучению. Чтобы преодолеть эту проблему, введем новый функционал  $R_\alpha(T)$ , представляющий собой сумму исходного функционала  $R(T)$  и штрафа за размер дерева:

$$R_\alpha(T) = R(T) + \alpha|T|, \tag{1.1}$$

где  $|T|$  — число листьев в поддерева  $T$ , а  $\alpha \geq 0$  — параметр. Это один из примеров *регуляризованных* критериев качества, которые ищут баланс между качеством классификации обучающей выборки и сложностью построенной модели. В дальнейшем мы много раз будем сталкиваться с такими критериями.

Мы покажем, что существует последовательность вложенных деревьев с одинаковыми корнями:

$$T_K \subset T_{K-1} \subset \dots \subset T_0,$$

(здесь  $T_K$  — тривиальное дерево, состоящее из корня дерева  $T_0$ ), в которой каждое дерево  $T_i$  минимизирует критерий (1.1) для  $\alpha$  из интервала  $\alpha \in [\alpha_i, \alpha_{i+1})$ , причем

$$0 = \alpha_0 < \alpha_1 < \dots < \alpha_K < \infty.$$

Возможно, существует несколько поддеревьев дерева  $T_0$  с одинаковым значением критерия  $R_\alpha(T)$ . В этом случае  $\alpha$ -оптимальным мы будем называть то из них, которое является поддеревом всех остальных (если такое имеется), и обозначать его  $T_0(\alpha)$ .

Рассмотрим любое бинарное дерево  $T$ , состоящее из более чем одной вершины, и для любой его внутренней вершины  $t$  обозначим с помощью  $T_t$  поддерево дерева  $T$  с корнем в  $t$ . *Левым поддеревом* вершины  $t$  будем называть поддерево дерева  $T$  с корнем в левой дочерней вершине вершины  $t$ , а *правым поддеревом* — с корнем в правой. Будем обозначать их  $T_{tl}$  и  $T_{tr}$  соответственно. Заметим, что функционал  $R_\alpha$  можно вычислять следующим образом:

$$R_\alpha(T_t) = R_\alpha(T_{tl}) + R_\alpha(T_{tr}). \quad (1.2)$$

Введем следующую величину:

$$g(t, T) = \frac{R(t) - R(T_t)}{|T_t| - |t|} = \frac{R(t) - R(T_t)}{|T_t| - 1}.$$

Сразу же отметим, что  $g(t, T) > \alpha$  тогда и только тогда, когда  $R_\alpha(t) > R_\alpha(T_t)$ . Отметим также (это нам пригодится позже), что функционал  $R_\alpha(T)$  можно представить в виде суммы  $R(t) + \alpha$  всех листовых вершин  $t$  дерева  $T$ .

Сокращение (или стрижка) вершины  $t$  заключается в замене поддерева  $T_t$  на листовую вершину  $t$ .

Докажем ряд утверждений.

**Теорема 1.1.** *Пронумеруем вершины дерева  $T$  так, чтобы любая вершина имела номер меньший, чем ее родительская вершина. Если мы будем посещать вершины в порядке этой нумерации (то есть «снизу вверх») и сокращать текущую вершину  $t$ , если  $R_\alpha(t) \leq R_\alpha(T_t)$ , то в результате получим дерево  $T(\alpha)$  —  $\alpha$ -оптимальное поддерево дерева  $T$ .*

**Доказательство.**

Предположим, что мы сейчас находимся в вершине  $t$  дерева  $T$ , совершив перед этим определенное число шагов (то есть, возможно, сократив некоторое количество вершин). Текущее дерево обозначим с помощью  $T'$ . Докажем по индукции, что после завершения текущей итерации получившееся поддерево  $T'_t$  с корнем в  $t$  будет  $\alpha$ -оптимальным.

В качестве базы индукции рассмотрим листья дерева  $T$ , для которых данное утверждение очевидно.

На текущем шаге мы либо заменяем все поддерево  $T'_t$  на его корневую вершину  $t$ , либо сохраняем  $T'_t$  целиком (в зависимости от того, как соотносятся  $R_\alpha(t)$  и  $R_\alpha(T'_t)$ ). Предположим, что ни один из двух вариантов не является оптимальным, и существует такое поддерево  $T''_t$  с корнем в  $t$ , что  $R_\alpha(T''_t) < \min(R_\alpha(t), R_\alpha(T'_t))$ . Тогда из представления (1.2) для  $R_\alpha$  следует, что выполнено хотя бы одно из двух неравенств

$$\begin{aligned} R_\alpha(T''_{tl}) &< R_\alpha(T'_{tl}), \\ R_\alpha(T''_{tr}) &< R_\alpha(T'_{tr}). \end{aligned}$$

Значит, одно из текущих поддеревьев не является  $\alpha$ -оптимальным, что противоречит предположению индукции.

Предположим теперь, что существует некоторое нетривиальное поддерево  $T''_t$ , такое что  $R_\alpha(T''_t) = \min(R_\alpha(t), R_\alpha(T'_t))$ . Покажем, что полученное в результате текущей итерации поддерево с корнем в  $t$  будет вложено в  $T''_t$ . Если в результате итерации будет произведено сокращение поддерева  $T'_t$  до вершины  $t$ , то утверждение очевидно. Если же будет сохранено поддерево  $T'_t$ , то будут выполнены оба равенства

$$\begin{aligned} R_\alpha(T''_{tl}) &= R_\alpha(T'_{tl}), \\ R_\alpha(T''_{tr}) &= R_\alpha(T'_{tr}); \end{aligned}$$

(если хотя бы в одном из них имеет место неравенство, то получим, что одно из текущих поддеревьев не является  $\alpha$ -оптимальным). Поскольку поддерева  $T'_{tl}$  и  $T'_{tr}$  оптимальны, то по определению  $T''_{tl}$  является поддеревом дерева  $T'_{tl}$  и  $T''_{tr}$  является поддеревом дерева  $T'_{tr}$ .

Таким образом, после принятия решения о сокращении вершины  $t$  оставшееся поддерево с корнем в  $t$  будет  $\alpha$ -оптимальным. То же самое справедливо и для момента, когда мы доберемся до корня всего дерева  $T$ , а значит в результате описанной процедуры мы получим  $T(\alpha)$ . ■

Описанный в формулировке теоремы метод назовем  $\alpha$ -отсечением. Итак, с помощью  $\alpha$ -отсечения мы можем получить  $\alpha$ -оптимальное поддерево дерева  $T$ .

**Теорема 1.2.** Пусть  $\alpha_1$  — наименьшее значение  $g(t, T)$  среди всех **внутренних** вершин:  $\alpha_1 = \min_{t: t \text{ внутр.}} g(t, T)$ . Тогда:

1. Для всех  $\alpha < \alpha_1$   $\alpha$ -оптимальным поддеревом дерева  $T$  является оно само.
2. Дерево  $T_1 = T(\alpha_1)$  получается из дерева  $T$  сокращением всех вершин  $t$ , для которых  $g(t, T) = \alpha_1$ .
3. Для всех внутренних вершин  $t$  дерева  $T_1$  выполнено  $g(t, T_1) > \alpha_1$ .

**Доказательство.**

Заметим, что если

$$g(t, T_t) = \frac{R(t) - R(T_t)}{|T_t| - 1} = \alpha_1 > \alpha,$$

то

$$R_\alpha(t) > R_\alpha(T_t),$$

а значит для таких значений  $\alpha$  ни одно сокращение внутренней вершины произведено не будет. Отсюда получаем первое утверждение теоремы.

Пусть теперь  $\alpha = \alpha_1$ . В этом случае  $\alpha$ -отсечение соответствует удалению всех вершин, для которых  $R_{\alpha_1}(t) \leq R_{\alpha_1}(T_t)$ . Поскольку вершин  $t$  с  $R_{\alpha_1}(t) < R_{\alpha_1}(T_t)$  не существует (т.к.  $\alpha_1 = \min_{t \in T} g(t, T)$ ), то данное условие равносильно условию  $R_{\alpha_1}(t) = R_{\alpha_1}(T_t)$ , или, что то же самое,  $g(t, T) = \alpha_1$ . Второе утверждение доказано.

Перейдем к третьему пункту. Мы уже говорили, что не существует вершин с  $R_{\alpha_1}(t) < R_{\alpha_1}(T_t)$ . Поэтому для всех вершин  $t$  дерева  $T_{1t}$  выполнено  $R_{\alpha_1}(T_{1t}) = R_{\alpha_1}(T_t)$ . Мы хотим доказать, что

$$\frac{R(t) - R((T_1)_t)}{|(T_1)_t| - 1} > \alpha_1$$

для всех внутренних вершин  $t$  дерева  $T_1$ . Запишем:

$$\begin{aligned} R(t) - R(T_{1t}) &= R(t) + \alpha_1 - (R(T_{1t}) + \alpha_1|T_{1t}|) + \alpha_1(|T_{1t}| - 1) = \\ &= R_{\alpha_1}(t) - R_{\alpha_1}(T_{1t}) + \alpha_1(|T_{1t}| - 1) = \\ &= R_{\alpha_1}(t) - R_{\alpha_1}(T_t) + \alpha_1(|T_{1t}| - 1) > \\ &> \alpha_1(|T_{1t}| - 1), \end{aligned}$$

где мы воспользовались тем фактом, что раз вершина  $t$  осталась внутренней (то есть не была сокращена), то  $R_{\alpha_1}(t) > R_{\alpha_1}(T_t)$ . ■

**Теорема 1.3.** *Для любого  $\beta > \alpha$  дерево  $T(\beta)$  является поддеревом  $T(\alpha)$  и получается из него в результате  $\beta$ -отсечений.*

**Доказательство.**

Докажем по индукции, что дерево  $T_t(\beta)$  является поддеревом  $T_t(\alpha)$ , что будет означать, что  $T(\beta)$  является поддеревом  $T(\alpha)$ . Для листовых вершин дерева  $T_t(\beta)$  утверждение очевидно.

Рассмотрим внутреннюю вершину  $t$ . Имеем, что  $T_{t\ell}(\beta) \subseteq T_{t\ell}(\alpha)$  и  $T_{tr}(\beta) \subseteq T_{tr}(\alpha)$ . В данном случае правое и левое поддерева вершины  $t$ , полученные с использованием  $\alpha$ , равны  $T_{tr}(\alpha)$  и  $T_{t\ell}(\alpha)$  соответственно (мы доказали это по индукции в теореме 1.1). Поддерева с корнем в вершине  $t$ , полученные с использованием  $\alpha$  и  $\beta$ , обозначим  $T_t^\alpha$  и  $T_t^\beta$  соответственно (имеются в виду поддерева до принятия решения о сокращении вершины  $t$ ).

Наша цель — показать, что если мы сократим вершину  $t$  с использованием  $\alpha$ , то и при использовании  $\beta$  мы ее сократим. Иными словами, мы хотим показать, что если  $R_\alpha(t) \leq R_\alpha(T_t^\alpha)$ , то и  $R_\beta(t) \leq R_\beta(T_t^\beta)$ . Это будет означать, что дерево  $T_t(\beta)$  является поддеревом  $T_t(\alpha)$ .

Запишем:

$$\begin{aligned}
R_\beta(t) &= \\
&= R_\alpha(t) + (\beta - \alpha) \leq \\
&\leq R_\alpha(T_t^\alpha) + (\beta - \alpha) = \\
&= R_\alpha(T_{tl}(\alpha)) + R_\alpha(T_{tr}(\alpha)) + (\beta - \alpha) \leq \\
&\leq R_\alpha(T_{tl}(\beta)) + R_\alpha(T_{tr}(\beta)) + (\beta - \alpha) = \\
&= R_\beta(T_{tl}(\beta)) + R_\beta(T_{tr}(\beta)) - (\beta - \alpha)(|T_t^\beta| - 1) = \\
&= R_\beta(T_t^\beta) - (\beta - \alpha)(|T_t^\beta| - 1) \leq \\
&\leq R_\beta(T_t^\beta).
\end{aligned}$$

Поскольку по теореме 1.1 дерево  $T(\beta)$  минимизирует  $R_\beta(T')$  по всем поддеревьям  $T' \subseteq T$  с тем же корнем, и при этом оно является поддеревом  $T(\alpha)$ , то оно также минимизирует  $R_\beta(T')$  по всем поддеревьям  $T' \subseteq T(\alpha)$  с тем же корнем, значит его мы получим в результате  $\beta$ -отсечений дерева  $T(\alpha)$ . ■

Доказанные теоремы дают нам возможность сформулировать алгоритм построения описанных в начале последовательностей  $T_K \subset \dots, T_0$  и  $\alpha_K > \dots > \alpha_0$ .

С помощью процедуры, описанной в теореме 1.2, найдем значение  $\alpha_1$  и дерево  $T_1 = T_0(\alpha_1)$ . Затем повторим для него процедуру, находим  $\alpha_2 > \alpha_1$  (по теореме 1.2) и  $T_2 = T_0(\alpha_2)$ . Будем повторять это до тех пор, пока не получим тривиальное дерево  $T_K$ , состоящее лишь из корня дерева  $T_0$ . Теоремы 1.2 и 1.3 говорят, что для любого  $i$  дерево  $T_i$  является  $\alpha$ -оптимальным поддеревом дерева  $T_0$  для любого  $\alpha \in [\alpha_i, \alpha_{i+1})$ .

Мы получили следующий алгоритм стрижки дерева  $T$ :

1. Положим  $k = 0$ ,  $T_0 = T$ .
2. Положим  $\alpha = \infty$ .
3. Обходим **внутренние** вершины  $t$  дерева снизу вверх и вычисляем  $R(T_t)$ ,  $|T_t|$  и  $g(t, T)$ . Полагаем  $\alpha = \min(\alpha, g(t, T))$ .
4. Обходим вершины сверху вниз и сокращаем вершины  $t$ , для которых  $g(t, T) = \alpha$ .
5. Полагаем  $k = k + 1$ ,  $\alpha_k = \alpha$ ,  $T_k = T$ .
6. Если  $T$  имеет больше одной вершины — возвращаемся к шагу 2.

**Выбор параметра  $\alpha$ .** Как выбрать одно из построенных деревьев  $T_i$ ? Обычно используют один из двух вариантов:

1. Перед построением дерева обучающая выборка разбивается на две части, и для построения используется лишь первая часть. При стрижке выбирается то из деревьев  $T_i$ , которое дает наилучшее качество на второй, скрытой части выборки.

2. Применяется скользящий контроль. Выборка многократно разбивается на две части, первая используется для построения дерева и последовательности  $\{T_i\}$ , по второй вычисляется кусочно-постоянная функция  $R(T(\alpha))$ . Значения функции усредняются по всем разбиениям, и выбирается то  $\alpha$ , которое минимизирует эту усредненную функцию.

## Список литературы

- [1] *Breiman, L., Friedman, J., Olshen, R. and Stone, C.* Classification and regression trees. // Wadsworth, New York, 1984.
- [2] *Ripley, B. D.* Pattern Recognition and Neural Networks. // Cambridge University Press, 1996.
- [3] *Hastie, T., Tibshirani, R., Friedman, J.* The Elements of Statistical Learning. // Springer, 2003.