Методы повышения обобщающей способности, основанные на различных способах построения ансамблей

Медведев Дмитрий Владимирович

Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математических методов прогнозирования

Выпускная квалификационная работа бакалавра

Научный руководитель — д.ф-м.н., профессор Сенько О. В.

Москва, 2019

Постановка задачи

1. Известно, что коллективные методы позволяют увеличивать обобщающую способность, поэтому построение ансамблей становится важным направлением для решения практических задач.
2. Необходимо с помощью композиции нескольких более слабых, в смысле качества, алгоритмов, построить один сильный. То есть создать ансамбль.
3. Пример: для задачи классификации с помощью специального метода обучить различные решающие деревья и объеденить их в один итоговый алгоритм.

Разложение ошибки ансамбля

$$
\begin{aligned}
H(x) &= \sum_{i=1}^{T} h_i(x) \\
err(H) &= \mathbb{E}\left[\int (H(x) - f(x))^2 dx\right] = \mathbb{E}\left[(H - f)^2\right] \\
err(H) &= \underbrace{\frac{1}{T}\sum_{i=1}^{T} (\mathbb{E}[h_i] - f)^2}_{\overline{bias}(H)^2} + \frac{1}{T} \cdot \underbrace{\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}\left[(h_i - \mathbb{E}[h_i])^2\right]}_{\overline{variance}(H)} + \\
&+ \left(1 - \frac{1}{T}\right) \cdot \underbrace{\frac{1}{T(T-1)} \sum_{i=1}^{T} \sum_{\substack{j=1 \\ j \neq i}}^{T} \mathbb{E}\left[(h_i - \mathbb{E}[h_i])(h_j - \mathbb{E}[h_j])\right]}_{\overline{covariance}(H)}
\end{aligned}
$$

Разложение ошибки ансамбля

$$
\begin{aligned}
err(H) &= \overline{err}(H) - \overline{ambi}(H) \\
\overline{err}(H) &= \mathbb{E}\left[\frac{1}{T}\sum_{i=1}^{T}(h_i - f)^2\right] = \overline{bias}(H)^2 + \overline{variance}(H) \\
\overline{ambi}(H) &= \mathbb{E}\left[\frac{1}{T}\sum_{i=1}^{T}(h_i - H)^2\right] = \overline{variance}(H) - variance(H) = \\
&= \overline{variance}(H) - \frac{1}{T}\overline{variance}(H) - \left(1 - \frac{1}{T}\right)\overline{covariance}(H) \\
err(H) &\to min \Leftrightarrow \frac{1}{T}\sum_{i=1}^{T} err(h_i) - \frac{1}{2T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}\mathbb{E}(h_i - h_j)^2 \to min
\end{aligned}
$$

Известные методы построения ансамблей

- Бэггинг
- Метод случайных подпространств
- Случайный лес
- Бустинг

Цель работы: разработка и исследование нового метода ансамблей решающих деревьев. Разработанный метод сравнивается с известными моделями основанными на ансамблях деревьев включая: случайный лес и адаптивный бустинг.

Предлагаемый метод: критерий разбиения в узле

$$
\begin{aligned}
(\tau, d) &= \operatorname*{argmin}_{\tau,d} \left(\frac{|S_l|}{|S_l| + |S_r|} H(S_l) + \frac{|S_r|}{|S_l| + |S_r|} H(S_r) \right) \\
H(S) &= -\sum_{k \in K} p_1 \cdot \ln p_1 + \boxed{\lambda \cdot \sum_{k \in K} p_2 \cdot \ln p_2} \\
p_1 &= \frac{1}{|S|} \sum_{x \in S} \mathbb{I}[y(x) = k] \ , \ p_2 = \frac{1}{|S|} \sum_{x \in S} \hat{p}(k | C^{m-1}, x)
\end{aligned}
$$

Основная идея: строить различные деревья, используя построенный на предыдущем шаге ансамбль, максимизировать его энтропию и минимизировать энтропию реальных откликов.

Предлагаемый метод: обозначения

- T^M — дерево, построенное на M-м шаге. $Leaf$ — множество объектов в листовом узле, в котором находится x_i.

$$\hat{p}(k|T^M, x) = \frac{1}{|Leaf|} \sum_{\hat{x} \in Leaf} \mathbb{I}[y(\hat{x}) = k].$$

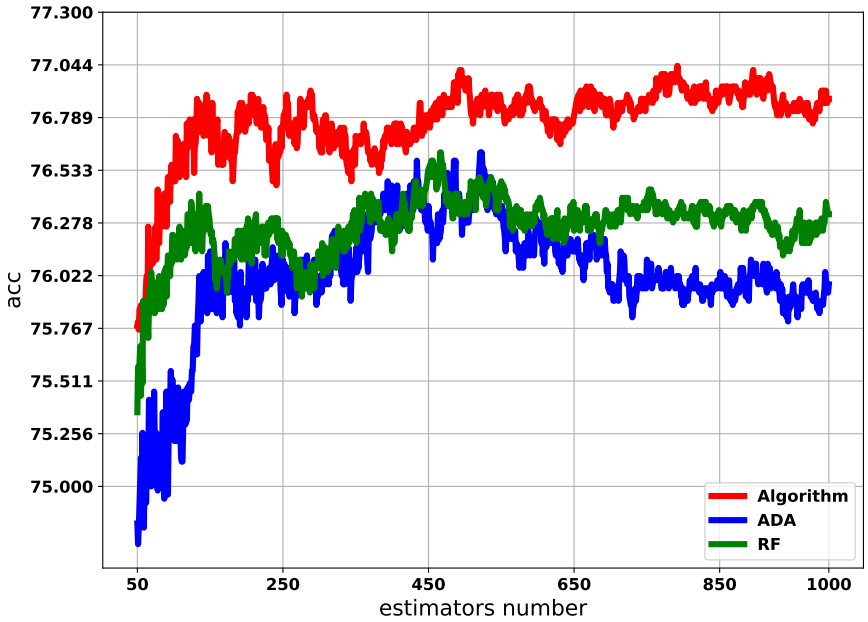
- C^M — ансамбль построенный на M-м шаге.

$$\hat{p}(k|C^M, x) = \frac{1}{M} \sum_{m=1}^{M} \hat{p}(k|T^m, x)$$

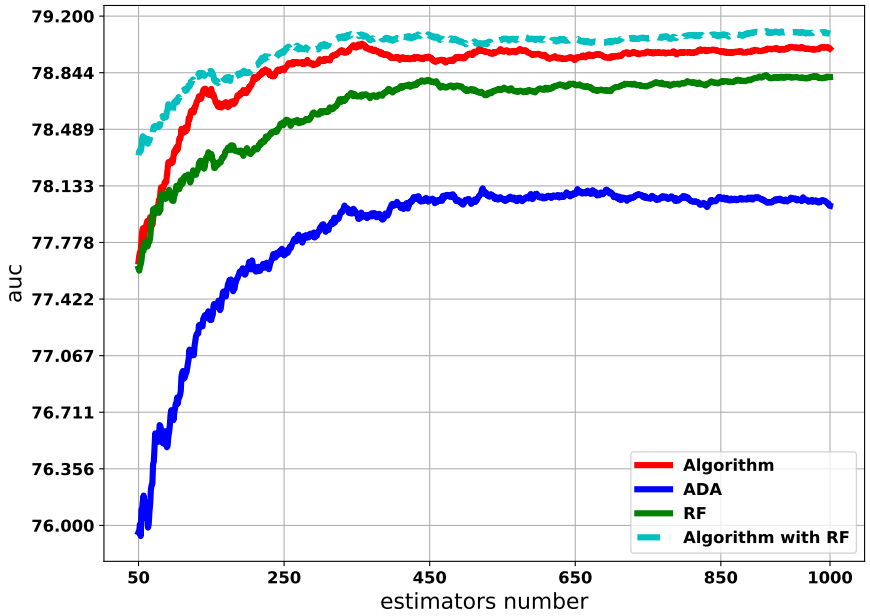
$$C^M(x) = \underset{k \in 1,\ldots,K}{\operatorname{argmax}} \hat{p}(k|C^M, x)$$

- λ — коэффициент "влияния" предыдущих деревьев на построение.

Эксперименты: кредитный скоринг



а)

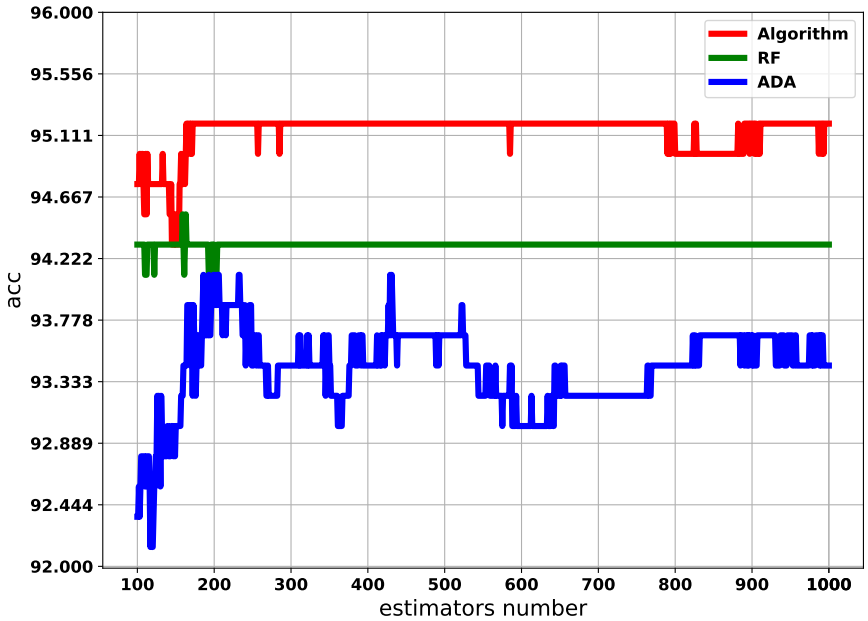


б)

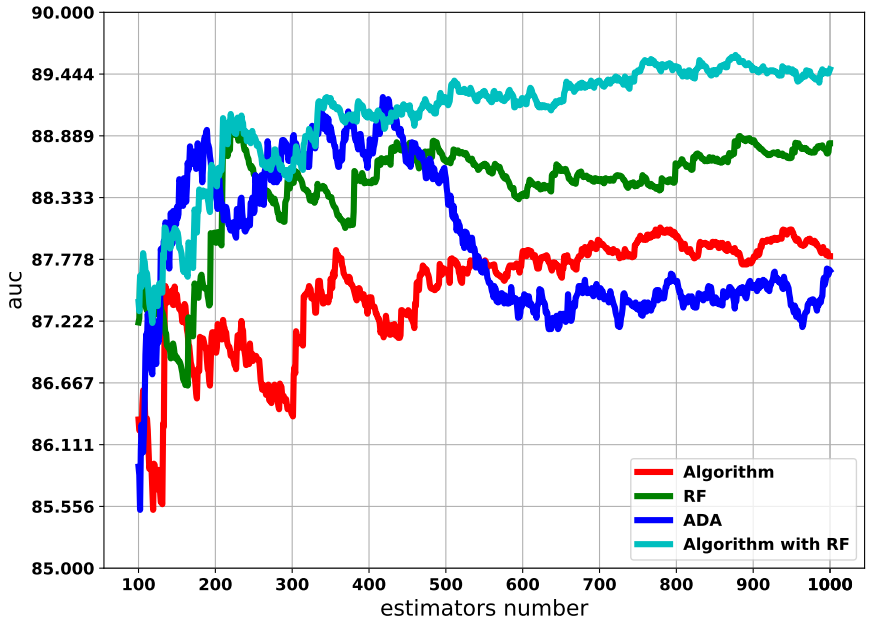
Рис.: Кредитный скоринг: а) точность (accuracy), б) AUC-ROC.

Здесь и далее красный: предлагаемый алгоритм, зелёный: случайный лес, синий: AdaBoost, голубой: алгоритм + случайный лес.

Эксперименты: систолическое давление



а)



б)

Рис.: Систолическое давление: а) точность (accuracy), б) AUC-ROC

Эксперименты: классификация силуэта машины

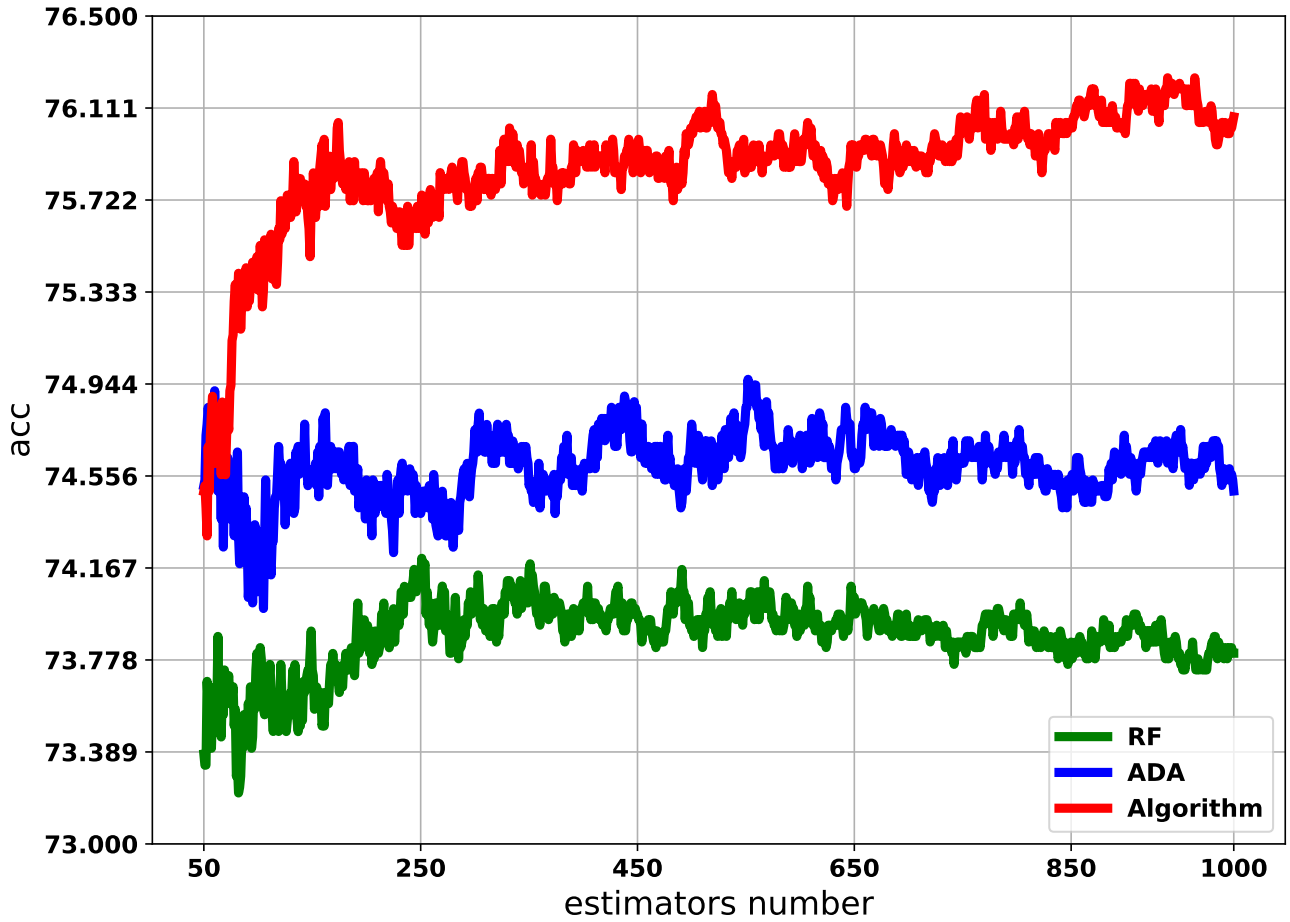


Рис.: Классификация силуэта машины. Точность (accuracy)

Эксперименты: таблицы результатов

Метод	Accuracy	Max accuracy	AUC	Max AUC
RF	0.763	0.766	0.788	0.788
AdaBoost	0.759	0.766	0.780	0.781
Algorithm	**0.768**	**0.770**	0.789	0.790
Algorithm + RF	0.764	0.767	**0.790**	**0.791**
Algorithm + AdaBoost	0.766	0.769	0.789	0.790

Таблица: Кредитный скоринг

Метод	Accuracy	Max accuracy	AUC	Max AUC
RF	0.888	0.943	0.891	0.958
ADA	0.876	0.934	0.892	0.943
Algorithm	0.878	0.951	0.880	0.951
Algorithm + RF	**0.894**	0.945	**0.896**	**0.958**
Algorithm + ADA	0.876	**0.954**	0.886	0.954

Таблица: Систолическое давление, задача 1

Эксперименты: таблицы результатов

Метод	Accuracy	Max accuracy
RF	0.738	0.742
AdaBoost	0.744	0.749
Algorithm	**0.760**	**0.762**
Algorithm + RF	0.750	0.751
Algorithm + AdaBoost	0.759	0.762

Таблица: Классификация силуэта машины

Положения, выносимые на защиту

На защиту выносится:

1. Новый метод построения различных решающих деревьев и объеденения их в ансамбль.
2. Экспериментальное сравнение предложенного метода с известными аналогичными подходами.

Предлагаемый метод: обозначения

- $y(x)$ — реальная метка, соответствующая объекту x в выборке.
- K — число классов в задаче классификации.
- $Node$ — множество объектов в текущем узле, для которого идёт поиск признака и порога по нему.
- T^M — дерево, построенное на M-м шаге. $Leaf$ — множество объектов в листовом узле, в котором находится x_i.
  $\hat{p}(k|T^M, x) = \frac{1}{|Leaf|} \sum_{\hat{x} \in Leaf} \mathbb{I}[y(\hat{x}) = k]$.
- C^M — ансамбль построенный на M-м шаге.
  $\hat{p}(k|C^M, x) = \frac{1}{M} \sum_{m=1}^{M} \hat{p}(k|T^m, x)$
  $C^M(x) = \underset{k \in 1,\ldots,K}{\operatorname{argmax}} \hat{p}(k|C^M, x)$
- λ — коэффициент "влияния" предыдущих деревьев на построение.

Предлагаемый метод: критерий разбиения в узле

$$\begin{cases} S_l = \{x \in Node | x^d <= \tau\} \\ S_r = \{x \in Node | x^d > \tau\} \\ F(\tau, d) = \frac{|S_l|}{|S_l|+|S_r|} H(S_l) + \frac{|S_r|}{|S_l|+|S_r|} H(S_r) \\ (\tau, d) = \underset{\tau,d}{\text{argmin}}(F(\tau, d)) \end{cases}$$

$$\begin{cases} p(k|S, Y) = \frac{1}{|S|} \sum\limits_{x \in S} \mathbb{I}[y(x) = k] \\ p(k|S, C^{m-1}) = \frac{1}{|S|} \sum\limits_{x \in S} \hat{p}(k|C^{m-1}, x) \\ H(S) = -\sum_{k \in K} p(k|S, Y) \cdot \ln p(k|S, Y) + \\ + \lambda \cdot \sum_{k \in K} p(k|S, C^{m-1}) \cdot \ln p(k|S, C^{m-1}) \end{cases}$$

Основная идея: строить различные деревья, используя построенный на предыдущем шаге ансамбль, максимизировать его энтропию и минимизировать энтропию реальных откликов.

Дополнительная информация

Используемые в эксперементах данные: University of California,Irvine) Machine Learning Repository

Энтропия: $-\sum p \log p$

FPR: $a(x, thr) = sign(f(x) - thr)$, $FPR = \frac{\sum_{i=1}^{N}[a(x_i,thr)=+1][y_i=-1]}{\sum_{i=1}^{N}[y_i=-1]}$

TPR: $TPR = \frac{\sum_{i=1}^{N}[a(x_i,thr)=+1][y_i=+1]}{\sum_{i=1}^{N}[y_i=+1]}$

ROC:$(0, 0)$ Соответсвует наибольшему thr: a всегда выдаёт отрицательный результат

Дополнительная информация

Задача	Train volume	Test volume	Количество признаков	Количество классов
Кредитный скоринг	1000	–	24	2
Классификация силуэта машины	846	–	18	4
Систолическое давление	718	458	116	2

Таблица: Информация о выборках