

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

ММРО 2019

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

Москва, 2019

- 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 объединить их в один итоговый алгоритм.

Разложение ошибки ансамбля

$$H(x) = \sum_{i=1}^T h_i(x)$$

$$\text{err}(H) = \mathbb{E} \left[\int (H(x) - f(x))^2 dx \right] = \mathbb{E} [(H - f)^2]$$

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

Разложение ошибки ансамбля

$$err(H) = \overline{err}(H) - \overline{ambi}(H)$$

$$\overline{err}(H) = \mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T (h_i - f)^2 \right] = \overline{bias}(H)^2 + \overline{variance}(H)$$

$$\begin{aligned} \overline{ambi}(H) &= \mathbb{E} \left[\frac{1}{T} \sum_{i=1}^T (h_i - H)^2 \right] = \overline{variance}(H) - \overline{variance}(H) = \\ &= \overline{variance}(H) - \frac{1}{T} \overline{variance}(H) - \left(1 - \frac{1}{T} \right) \overline{covariance}(H) \end{aligned}$$

$$err(H) \rightarrow \min \Leftrightarrow \frac{1}{T} \sum_{i=1}^T err(h_i) - \frac{1}{2T^2} \sum_{i=1}^T \sum_{j=1}^T \mathbb{E}(h_i - h_j)^2 \rightarrow \min$$

- Бэггинг
- Метод случайных подпространств
- Случайный лес
- Бустинг (Адаптивный, Градиентный)

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

$$(\tau, d) = \operatorname{argmin}_{\tau, d} \left(\frac{|S_l|}{|S_l| + |S_r|} H(S_l) + \frac{|S_r|}{|S_l| + |S_r|} H(S_r) \right)$$

$$H(S) = - \sum_{k \in K} p_1 \cdot \ln p_1 + \lambda \cdot \sum_{k \in K} p_2 \cdot \ln p_2$$

$$p_1 = \frac{1}{|S|} \sum_{x \in S} \mathbb{I}[y(x) = k], \quad p_2 = \frac{1}{|S|} \sum_{x \in S} \hat{p}(k | C^{m-1}, x)$$

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

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

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

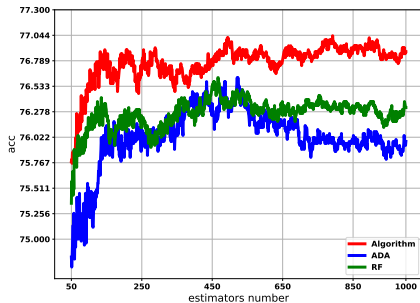
- C^M — ансамбль построенный на M -м шаге.

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

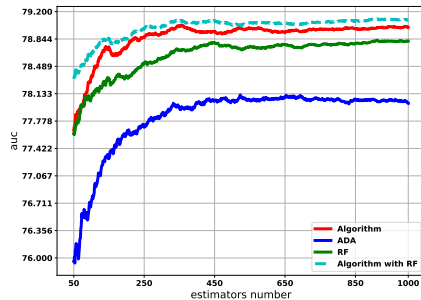
$$C^M(x) = \operatorname{argmax}_{k \in 1, \dots, K} \hat{p}(k|C^M, x)$$

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

Эксперименты: кредитный скоринг



а)

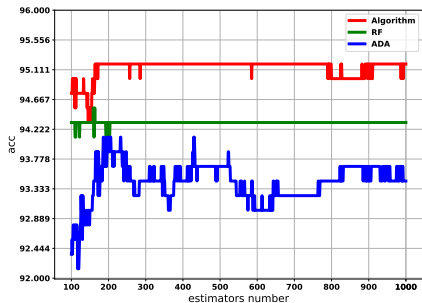


б)

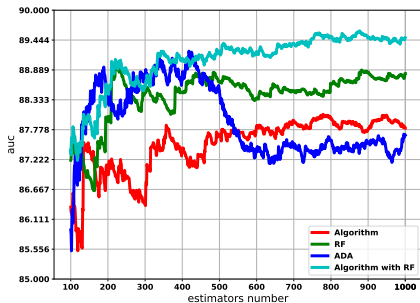
Рис.: Кредитный скоринг: а) точность (accuracy), б) AUC-ROC.

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

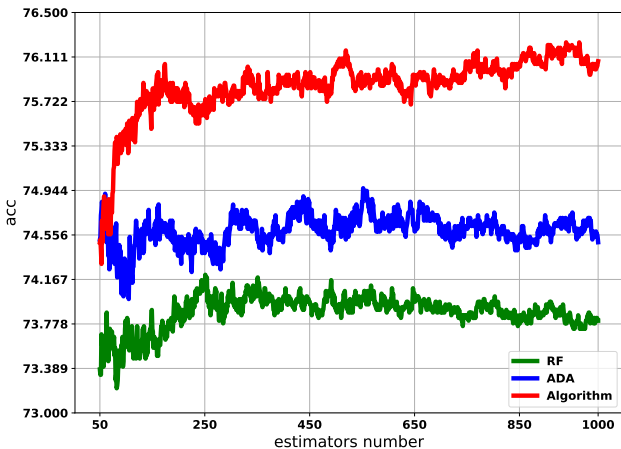


Рис.: Классификация силуэта машины. Точность (accuracy)

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

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

Спасибо за внимание!