# Повышение обобщающей способности бустинга в задачах с перекрывающимися классами

***Баринова О. В., Вежневец А. П., Вежневец В. П.***
olga.barinova@gmail.com, {avezhnevets, vvp}@graphics.cs.msu.ru
Москва, МГУ им. М. В. Ломоносова

Проблема переобучения является одной из центральных в машинном обучении. Известно, что алгоритмы бустинга (boosting) на одних задачах демонстрируют хорошую способность к обобщению, а на других склонны к переобучению. В работе [1] описан ряд методов для уменьшения переобучения в бустинге. В данной работе предлагаются два новых усовершенствования бустинга и проводится сравнительный анализ их обобщающей способности.

Рассмотрим задачу классификации на два класса. Обозначим обучающую выборку через $T = \{(x_i, y_i)\}_{i=1}^N$, где $x_i \in X$ — вектор признаков, $y_i \in \{-1, 1\}$ — метка класса. Пусть прецеденты $(x_i, y_i)$ взяты из неизвестного распределения $P(x, y)$. В основе алгоритмов бустинга лежит процедура минимизации эмпирического риска:

$$R_N(F) = \frac{1}{N} \sum_{i=1}^{N} C\left(y_i, F(x_i)\right) \to \min,$$

где $F(x)$ — выход бустинга, $C\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ — функция потерь.

### Удаление из обучающей выборки путающих прецедентов

Будем называть *путающими* те прецеденты из обучающей выборки, на которых ошибается идеальный байесовский классификатор:

$$\left\{(x_i, y_i) \in T \;\middle|\; P(-y_i \,|\, x_i) > 0.5 > P(y_i \,|\, x_i)\right\}.$$

*Путающие* прецеденты встречаются в задачах с перекрывающимися распределениями классов. Неправильная классификация этих прецедентов предпочтительнее, чем правильная, однако процедура минимизации эмпирического риска ведет к настройке на *путающих* прецедентах, что может приводить к переобучению.

Первый предлагаемый подход к повышению обобщающей способности бустинга состоит в удалении *путающих* прецедентов из обучающей выборки.

### Использование оценки математического ожидания функции потерь с меньшей дисперсией

Второй предлагаемый подход состоит в замене минимизации эмпирического риска при настройке бустинга минимизацией другой оценки

**Алгоритм 1.** Алгоритм оценивания апостериорной вероятности.

**Вход:**
Обучающая выборка $T = \{(x_i, y_i)\}_{i=1}^N$;
число итераций алгоритма $K$;

**Выход:**
Оценки апостериорной вероятности $\bar{p}(1 \mid x_i),\ i = 1, \ldots, N$;

1: **для всех** $k = 1, \ldots, K$
2: Разделить обучающую выборку случайным образом на три равные части $T_k^1 \cup T_k^2 \cup T_k^3 = T$: $T_k^i \cap T_k^j = \varnothing,\ i \neq j$;
3: Обучить бустинг на первой части выборки $T_k^1$, получить классификатор $F_k$;
4: Оценить параметры калибровки $A$, $B$ для $F_k$ (оптимальные параметры сигмоиды) на второй части выборки $T_k^2$ при помощи шкалирования Платта [3];
5: Вычислить апостериорные вероятности на третьей части выборки по формуле: $p^k(1 \mid x_i) = \dfrac{1}{1 + \exp\left(AF_k(x_i) + B\right)}$;
6: Вычислить среднее значение апостериорной вероятности для каждого прецедента: $\bar{p}(1 \mid x_i) = \frac{1}{K} \sum_{k=1}^K p^k(1 \mid x_i)$;

математического ожидания функции потерь:

$$R'_N(F) = \frac{1}{N} \sum_{i=1}^N \Big( P(1 \mid x_i) C\big(1, F(x_i)\big) + P(-1 \mid x_i) C\big(-1, F(x_i)\big) \Big).$$

Нетрудно показать, что для дисперсий справедливо соотношение $\mathsf{D}R'_N(F) < \mathsf{D}R_N(F)$, и скорость сходимости $R'_N(F)$ к математическому ожиданию функции потерь значительно выше, чем у $R_N(F)$.

Соответствующее изменение легко встраивается в бустинг. Составляется расширенная обучающая выборка $T' = \big\{(x_i, y_i)\big\}_{i=1}^N \cup \big\{(x_i, -y_i)\big\}_{i=1}^N$ и изменяется инициализация весов бустинга [2]: вместо начальных весов $D_1(i) = \frac{1}{N},\ i = 1, \ldots, N$, берутся веса $D'_1(i) = p(y_i \mid x_i),\ i = 1, \ldots, 2N$.

### Оценивание апостериорной вероятности

В обоих методах используются апостериорные вероятности прецедентов из обучающей выборки. Апостериорные вероятности неизвестны, однако их можно оценить. Для этого предлагается Алгоритм 1.

В первом подходе полученная оценка апостериорной вероятности $\bar{p}(1 \mid x_i)$ используется для нахождения *путающих* прецедентов. Прецеденты, для которых $\bar{p}(y_i \mid x_i) < 0,5 < \bar{p}(-y_i \mid x_i)$, удаляются из обучающей

| Задача | Оценка ошибки | Исход. | Сокращ. | Расшир. |
|---|---|---|---|---|
| Breast | 3.73 | 4.6 | 3.65 | 4.58 |
| Australian | 13.06 | 15.2 | 13.8 | 12.75 |
| German | 24.66 | 25.72 | 25.35 | 24.8 |
| Heart | 18.34 | 21.41 | 18.36 | 16.67 |
| Pima | 24.03 | 25.58 | 23.99 | 23.31 |
| Spam | 6.49 | 6.19 | 6.02 | 6.06 |
| Vote | 4.51 | 4.75 | 4.61 | 4.37 |

**Таблица 1.** Ошибка кроссвалидации (%) бустинга, построенного по полной, сокращенной и расширенной обучающей выборке

выборки. Доля *путающих* прецедентов, обнаруженных в обучающей выборке, дает оценку ошибки обобщения.

Во втором подходе полученная оценка апостериорной вероятности $\bar{p}(1 \mid x_i)$ используется для инициализации весов бустинга: $D'_1(i) = \bar{p}(y_i \mid x_i),\ i = 1, \ldots, 2N$.

## Эксперименты

В экспериментах использовался алгоритм бустинга [2] с решающим деревом глубины 1 в качестве базового классификатора и числом итераций 100. Использовалась $50 \times 2$ кроссвалидация: данные 50 раз делились случайным образом на 2 равные части — на одной части настраивались алгоритмы, другая часть использовалась для вычисления ошибки.

В таблице 1 приведены результаты экспериментов на данных из репозитория UCI. В первом столбце приведена доля *путающих* прецедентов. Во втором столбце — ошибка кроссвалидации для бустинга, настроенного на исходной обучающей выборке. В третьем — ошибка кроссвалидации для бустинга, настроенного по сокращенной обучающей выборке (первый метод). В четвертом — ошибка кроссвалидации для бустинга, настроенного на расширенной обучающей выборке (второй метод).

Эксперименты показывают, что предлагаемые методы повышают обобщающую способность бустинга и помогают избежать переобучения.

## Литература

[1] *Friedman J.* Greedy function approximation: a gradient boosting machine // Annals of Statistics. — 2001. — Vol. 29, № 5. — Pp. 1189–1232

[2] *Schapire R., Singer Y.* Improved boosting algorithms using confidence-rated predictions // Machine Learning. — 1999. — Vol. 37, № 3. — Pp. 297–336

[3] *Niculescu-Mizil A., Caruana R.* Obtaining calibrated probabilities from boosting // 21st Conf. on Uncertainty in Artificial Intelligence, Edinburgh, Scotland, 2005.