Секция «Вычислительная математика и кибернетика»

Оценки вероятности переобучения и комбинаторные отступы объектов в задачах классификации

Соколов Евгений Андреевич
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: sokolov.evg@gmail.com

Переобучением в задачах классификации называют ситуацию, когда частота ошибок алгоритма классификации на обучающей выборке существенно меньше частоты его ошибок на независимой контрольной выборке. В комбинаторном подходе [1] рассматривается множество объектов $\mathbb{X} = \{x_1, \ldots, x_L\}$ и семейство алгоритмов $\mathbb{A}$. Каждый алгоритм a порождает бинарный вектор ошибок $\vec{a} = (a_1, \ldots, a_L)$, где $a_i = 1$ означает, что алгоритм a ошибается на объекте x_i. Рассматривается граф Хассе естественного отношения порядка на множестве векторов ошибок. Истоки этого графа соответствуют лучшим алгоритмам. Для каждой вершины графа вводятся две количественные характеристики — *связность*, равная числу исходящих рёбер, и *расслоение*, характеризующее удалённость вершины от истоков графа. Эти характеристики используются в комбинаторной оценке вероятности переобучения, благодаря чему она оказывается на порядки точнее классических оценок Вапника–Червоненкиса. Для приближённого вычисления оценки расслоения–связности достаточно перебрать вершины нескольких нижних слоёв графа, что позволяет непосредственно применять её на практике.

В данной работе вводится понятие *комбинаторного отступа* $d(x_i)$ объекта x_i. Это минимальное число объектов, на которых необходимо «испортить» ответ одного из лучших алгоритмов, чтобы классификация объекта x_i изменилась. Комбинаторный отступ тем меньше, чем ближе x_i к границе классов. Классическое понятие отступа объекта, широко используемое в теории классификации, имеет схожую интерпретацию, но определяется из геометрических соображений.

Доказано, что если в графе из вершины a выходит ребро, соответствующее объекту x_i, то вклад алгоритма a в оценку расслоения–связности экспоненциально убывает с ростом $d(x_i)$. Если $d(x_i) = t$, то объекту x_i могут соответствовать только рёбра графа, выходящие из вершины слоя $(t-1)$ или выше. Отсюда следует, что если из графа удалить все ребра, соответствующие объектам с отступом не меньше t, то нижние t слоёв не будут нарушены. Таким образом, зная отступы всех объектов, можно упростить граф и эффективно вычислить приближённые оценки расслоения–связности.

Численные эксперименты проводились с семейством линейных классификаторов. На рис. 1 представлены точная оценка и её приближения при различных значениях t для выборки из $L = 100$ объектов. Вычисление точной оценки по всему графу заняло 1 час, тогда как для вычисления приближения с $t = 5$ потребовалось всего 4 секунды.

Литература

1. Воронцов К.В. Комбинаторная теория переобучения: результаты, приложения и открытые проблемы // 15-ая Всеросс. конф. Математические методы распознавания образов. М.: МАКС Пресс, 2011. С. 40–43.

Иллюстрации

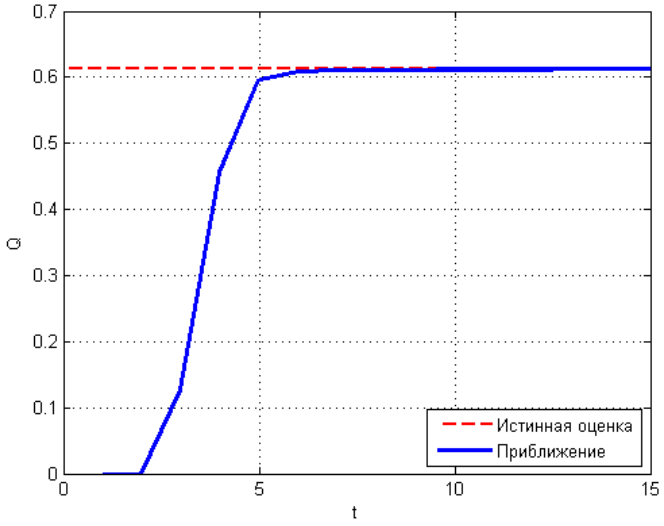

Рис. 1: Зависимость приближённой оценки вероятности переобучения от t.