Применение поворотов признакового пространства в контексте бустинга

Гой Антон Сергеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: g.a.s.s@mail.ru

Бустинг является одним из самых успешных методов объединения большого числа базовых алгоритмов в один предиктор, который способен давать более точные предсказания, чем его составляющие. Успех бустинга определён его ключевой идеей — добавлять в ансамбль такой базовый алгоритм, который сможет компенсировать ошибки на предыдущих шагах. Широкое распространение бустинга напрямую связано с алгоритмом The AdaBoost [1], который остаётся одним из самых точных алгоритмов в машинном обучении. Обычно AdaBoost строится над решающими деревьями небольшой глубины, что отражает другую идею бустинга — объединение «слабых» алгоритмов в один точный.

Существует множество других способов построения ансамблей, многие из которых, например бэггинг, независимо строят несколько алгоритмов, ответы которых усредняются. Основная задача в таких подходах — сделать каждый из алгоритмов как можно больше непохожим на другие. Как и в бустинге, такие методы в большинстве случаев строятся над деревьями решений, которые способны существенно видоизменять свою структуру даже при малейших изменениях в данных.

Серьёзный недостаток деревьев решений — трудность при работе с линейными зависимостями в данных.

В данном исследовании рассматривается AdaBoost над деревьями решений, а также ансамблевый алгоритм Rotation Forest [2]. Rotation Forest строит множество независимых деревьев решений, причём каждое дерево обучается на преобразованном признаковом пространстве. Данное преобразование представляет собой случайное разбиение всех признаков на подмножества одинаковой длины. Далее в каждом подпространстве, соответствующем конкретному подмножеству, осуществляется поворот за счёт применения метода PCA.

В ходе исследования были реализованы методы AdaBoost [3] и Rotation Forest, а также произведена оценка качества этих алгоритмов на десяти обучающих выборках из UCI Machine Learning

Данные	AdaBoost	RotationForest	Предложенный алгоритм
Zoo	0.960	0.970	0.990
Car	0.809	0.804	0.831
Wine	0.972	0.972	0.977
Diabetes	0.779	0.772	0.783
Iris	0.966	0.980	0.986
Voting	0.969	0.974	0.974
Digits	0.959	0.954	0.961
Breast Cancer	0.982	0.971	0.991
Ionosphere	0.925	0.954	0.960
Mamographic	0.810	0.746	0.814

Таблица 1: Результаты алгоритмов, Accuracy

Repository [4]. Из выборок были удалены объекты с пропущенными значениями, к категориальным признакам было применено one-hot кодирование.

Основная задача исследования заключалась в разработке и реализации нового алгоритма, который смог бы учесть лучшие стороны AdaBoost и Rotation Forest. Данный алгоритм представляет собой процедуру «жадного» построения ансамбля AdaBoost, на отдельном шаге которого обучающие данные преобразуются за счет поворотов случайных подмножеств признаков (как в Rotation Forest). Но вместо использования обычного метода PCA применяется его взвешенный вариант, где в качестве весов выступают веса объектов, которые пересчитываются на каждом шаге AdaBoost.

Результаты исследований представлены в Таблице 1.

Литература

1. Freund Y., Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences, Vol. 55, № 1, P. 119–139, Aug. 1997
2. Rodriguez J., Kuncheva L., Alonso C. Rotation Forest: A New Classifier Ensemble Method // IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 28, № 10, P. 1619–1630, Oct. 2006
3. Zhu J., Zou H., Rosset S., Hastie T. Multi-class AdaBoost // Statistics and Its. Interface, Vol. 2, P. 349–360, 2009
4. UCI Machine Learning Repository `https://archive.ics.uci.edu/ml/index.html`