Бэггинг над MARS со случайными поворотами признаков

A Preprint

Додонов В.О.
МГУ им. М.В. Ломоносова
Москва
s02200360@gse.cs.msu.ru

Китов В.В.
МГУ им. М.В. Ломоносова
Москва
v.v.kitov@yandex.ru

Abstract

Алгоритм multivariate adaptive regression splines (MARS) обеспечивает гибкий метод статистического моделирования, который использует прямой и обратный проходы, где происходит подбор порогов и переменных, для определения комбинации базовых функций, которые наилучшим образом приближают исходные данные. В области оптимизации MARS успешно использовался для оценки неизвестных функций в стохастическом динамическом программировании, стохастическом программировании и в других направлениях. MARS потенциально может быть полезен во многих реальных задачах оптимизации, где необходимо оценить целевую функцию на основе наблюдаемых данных. Однако использование MARS в ансамбле позволяет добиться даже большего качества на данных. Использование случайных ортогональных преобразований в ансамбле помогает сделать алгоритм менее чувствительным к расположению признаков в пространстве. Таким образом, получается найти более оптимальное приближение целевой функции в задаче регрессии.

Keywords MARS · Сплайны · Регрессия · Бэггинг · Ансамбли

1 Введение

В качестве популярного метода непараметрической регрессии алгоритм многомерных адаптивных регрессионных сплайнов (MARS) был впервые представлен Джеромом Фридманом в 1991 г. [5]. Благодаря своей гибкости и точности MARS использовался во многих исследованиях, где возникает классическая для машинного обучения задача регрессии, включая прогнозирование спроса на энергию, необходимую для транспортировки [11], анализ реакций роста, связанных с макропитательными веществами [1], построение системы принятия решений по борьбе с загрязнением озоном [15], оценку подверженности овражной эрозии [3], оценку тепловой нагрузки в зданиях [10], моделирование суточной концентрации растворенного кислорода [6] и др. Также MARS используется и в медицинских целях, например, для выявления влияния пола на факторы, оказывающих воздействие на нарушения опорно-двигательного аппарата плеч, шеи и верхних конечностей [12].

С момента появления классического алгоритма MARS [5] уже было произведено множество исследований, направленных на улучшение точности предсказания в задаче регрессии и ускорение обучения модели. Одними из первых стали модели с полиномиальными сплайнами [13], в которой удаление базисной функции из алгоритма влекло за собой последующее удаление всех произведенных от нее элементов, и байесовская модификация MARS [4], где использовались марковские цепи и метод Монте-Карло. Еще одним ответвлением стал CMARS [14], использующий непрерывные методы оптимизации и регуляризации.

Чтобы улучшить точность предсказаний алгоритма MARS и время обучения, были испробованы различные подходы. Например, ранее был предложен способ быстрой оптимизации узлов с использованием метода восхождения к вершине, где перебор порогов начинается с точки, на которой был достигнут максимум на предыдущем шаге [7]. Также в MARS можно использовать B-сплайны, что показало

свою эффективность в решении задачи регрессии [2], либо другие улучшения, как Robust MARS [9], который лучше справляется с менее надежными данными. Существую и другие подходы, расширяющие применение MARS, например, выпуклый вариант MARS, подходящий для решения специфических типов задач [8]

В данной работе предлагается использовать бэггинг над MARS (Реализация PyEarth[1]), а так же использовать модификацию - случайный поворот признаков, на которых в последствии будет обучаться алгоритм. Предыдущие подходы не учитывали случай, предполагающий, что базисная функция может функционально зависеть от, например, некоторой линейной комбинации признаков. В таком случае целевая переменная не представима в виде базисных функций, которые предлагают строить другие подходы. Однако использование ансамбля моделей, где каждая модель использует свое собственное признаковое пространство, позволяет нивелировать данную проблему.

2 Постановка задачи

В задачах регрессии целевая переменная y представляется в виде функции от p переменных с некоторым шумом ε:

$$y = f(x^1, x^2, ..., x^p) + \varepsilon,$$

где f неизвестная функция, и $\mathbb{E}\varepsilon = 0$.

По рассматриваемому набору данных $D = \{x_i, y_i\}_{i=1}^N$, функция $f(x^1, x^2, ..., x^p)$ аппроксимируется некоторой функцией $g(x^1, x^2, ..., x^p)$. Качество аппроксимации оценивается с помощью среднего квадрата ошибки (MSE):

$$MSE(D) = \frac{1}{N}\sum_{k=1}^{N}(g(x_k) - y_k)^2$$

либо схожей величины $RMSE(D) = \sqrt{MSE(D)}$. Считается, что модель с меньшим значением среднего квадрата ошибки лучше соответствует данным.

В данной работе функция $g(x)$ строилась согласно алгоритму MARS (Multivariate Adaptive Regression Splines) [5], который аппроксимирует исходную функцию f в виде:

$$h(x) = \sum_{m=0}^{M} a_m \cdot B_m(x),$$

где $B_m(x) = \prod_{k=1}^{K_m} b_{k,m}$. $b_{k,m}$ - базисная функция от одной переменной, которая имеет вид либо $max\{+(x-t),\ 0\}$, либо $max\{-(x-t), 0\}$.

3 Ансамбли и случайные повороты признаков

В данной работе предлагается рассмотреть ансамблевые модели с использованием MARS с целью улучшения предсказательной способности классического алгоритма. В общем виде ансамблевые модели можно представить в виде:

$$M = R(a_1, a_2, ..., a_T),$$

где a_t - базовые алгоритмы из некоторого рассматриваемого семейства $\mathcal{D}$. В данном случае предлагается рассмотреть:

$$M(x) = \frac{1}{T}\sum_{t}^{T} a_t(x), \tag{1}$$

где M - полученная модель, T - Количество базовых алгоритмов, a_i - i-я базовая модель, построенная по алгоритму MARS.

Исходя из вида (1), в случае некоррелированности базовых моделей, разброс композиции (в силу разложения ошибки на смещение-разброс) должен убывать пропорционально $\frac{1}{T}$. Чтобы достичь меньшей корреляции между моделями, они обучаются на случайных подмножествах объектов (с повторениями) и случайных подмножествах признаков.

[1] `https://github.com/scikit-learn-contrib/py-earth`

Так как каждая базисная функция в MARS является функцией от одной переменной, получившаяся модель может по разному справляться с аппроксимацией функции от, например, линейной комбинации переменных. Чтобы разнообразить семейство алгоритмов и еще сильнее уменьшить корреляцию между моделями, предлагается рассматривать алгоритм MARS со случайными поворотами признаков. В общем случае матрицей поворота называется ортогональная матрица Q. Перед обучением алгоритма MARS можно производить преобразование признаков из набора данных D: $\overline{x} = Qx$, где Q - случайная ортогональная матрица, $QQ^T = I$, x - $(x^1, ..., x^p)^T$ вектор в исходном пространстве признаков.

Матрица Q генерируется из распределения Хаара. Матрицы из этого распределения можно получить следующим способом: если $A \in \mathbb{R}^{n \times n}$, где каждый элемент сгенерирован независимо из стандартного нормального распределения, тогда матрица Q, участвующая в QR -разложении A, - искомая матрица поворота.

4 Эксперименты

4.1 Исходные данные и условия эксперимента

Для сравнения различных алгоритмов и подходов производились эксперименты на четырех различных наборах данных в задаче регрессии. Соответственно, во всех задачах в качестве метрики использовался RMSE. Все рассматриваемые наборы данных представляют собой реальные выборки, в которых необходимо предсказывать вещественное число (Boston Housing (BH), California Housing (CH), Real estate price prediction (RE), Diabetes Dataset (DD)).

Во всех экспериментах исходный набор данных разбивался на 3 выборки: обучающую, валидационную и тестовую. Соотношение между этими частями составляет 8:1:1 соответственно.

Размерность пространства признаков варьировалась от 6 до 10 в зависимости от задачи. В задачах применялась стандартизация наборов данных по обучающей выборке.

Лучшие параметры модели подбирались по валидационной выборке, которые затем использовались для сравнения результатов на тесте. Параметры бэггинга включают в себя следующие величины:

- Количество базовых алгоритмов. Оптимальное значение зависит от задачи. Сотни базовых алгоритмов оказывалось достаточно.
- Доля признаков, которая выбиралась для обучения модели. В экспериментах наилучшее значение лежало в промежутке $[0.5, 1]$.
- Максимальное количество слагаемых в MARS. Использовались значения в диапазоне $[20, 40]$
- Использование случайных поворотов.

Для корректного анализа предсказательной способности моделей бэггинг над алгоритмами MARS сравнивался еще и с другими классическими моделями машинного обучения: случайный лес (RandomForest), градиентный бустинг (GradientBoosting), Ridge-регрессия. Аналогичным образом по валидационной выборке подбирались параметры максимальной глубины для случайного леса ($depth \in \{2, 4, 6, 8, 10\}$), шага обучения ($lr \in \{0.1, 0.3, 0.5\}$) и доли используемых объектов в обучении ($subsample \in \{0.7, 0.8, 0.9, 1\}$) для градиентного бустинга, коэффициента регуляризации ($C \in \{10^k : k = -3, ..., 4\}$) для Ridge-регрессии.

4.2 Результаты эксперимента

В результате эксперимента классический алгоритм оказался хуже ансамблевого подхода, который показал результат на валидационной (Рис. 1) и тестовой выборках лучше (Табл. 1). Причем использование случайных поворотов позволило добиться меньшего значения RMSE на валидационной выборке, чем у градиентного бустинга. Также выяснилось, что с ростом числа базовых алгоритмов, использование случайных поворотов пространства признаков помогло улучшить предсказательную способность.

Рис. 2 иллюстрирует зависимость качества модель от используемых признаков. Использование части пространства признаков позволило уменьшить корреляцию между алгоритмами.

Результаты моделей, с наилучшими параметрами (с точки зрения качества на валидационной выборке) приведены в следующей Таблице (1) (MARS* соответствует модель, которая использовалась в качестве базового алгоритма в бэггинге):

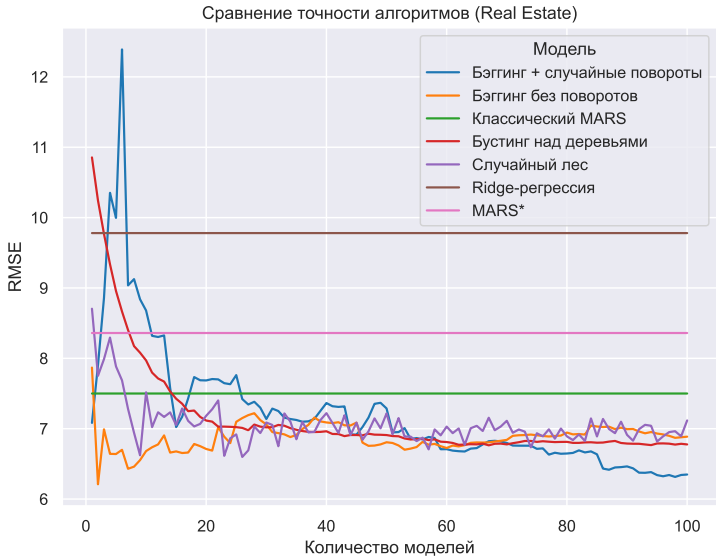

Рис. 1: Сравнение ансамблевого подходов с/без случайного поворота признаков со стандартным алгоритмом MARS и другими классическими алгоритмами машинного обучения (Набор данных - RE). Использовалось все признаковое пространство.

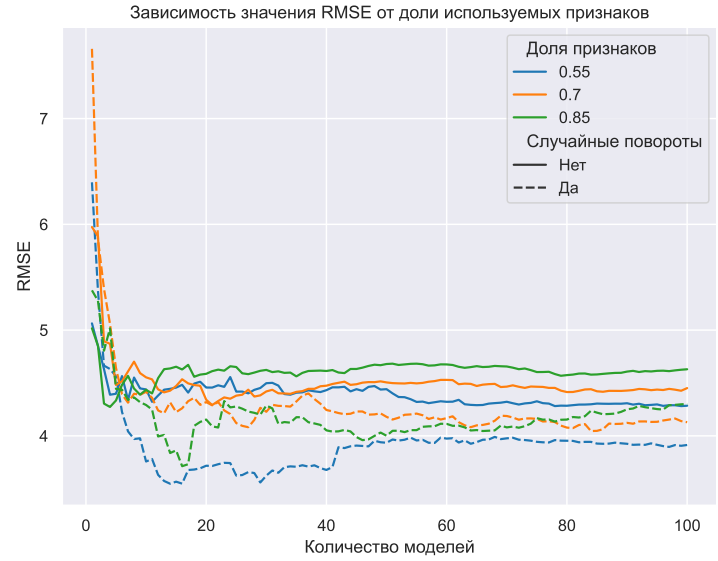

Рис. 2: Влияние выбора доли используемых признаков в каждом базовом алгоритме ансамбля (Набор данных - BH). Классический алгоритм MARS достиг $RMSE = 4.402$

Модель\Набор данных	BH	CH	RE	DD
Классический MARS	3.09	0.66	5.27	53.5
MARS*	3.32	0.67	6.08	57.2
Бэггинг без случайных поворотов	2.43	0.63	4.92	53.1
Бэггинг + случайные повороты	2.47	0.62	4.48	52.2
Ridge-регрессия	5.41	0.75	6.81	58.3
Случайный лес	2.77	0.63	5.06	54.1
Бустинг над деревьями	2.51	0.61	4.85	54.3

Таблица 1: Значение RMSE на тестовой выборке для каждой модели. Синим цветом выделен лучший результат.

4.3 Выводы

Таким образом, использование бэггинга над MARS на трех из четырех наборов данных позволило достичь лучшего результата на тестовой выборке. Однако стоит заметить, что в большинстве случаев именно подход со случайными поворотами признакового пространства помог добиться меньшего значения $RMSE$. То есть случайные повороты действительно смогли расширить семейство рассматриваемых базовых алгоритмов, уменьшив корреляцию между моделями.

Однако сложность построения ансамблевой модели составляет $\mathcal{O}(n)$, а если используется еще и матрица поворота, то появляются дополнительные затраты памяти порядка $\mathcal{O}(nd^2)$, где n - количество базовых алгоритмов в ансамбле, d - размерность пространства признаков. Бороться с этой проблемой можно, например, используя лишь некоторое подмножество матриц поворота.

5 Заключение

Таким образом, в данной работе была показана эффективность ансамблевого подхода над алгоритмом MARS для решения задач регрессии. Так же был предложен новый способ построения базовых алгоритмов - применение случайных поворотов пространства признаков для каждой отдельной модели в ансамбле, что помогло разнообразить семейство алгоритмов и добиться меньшего значения функции потерь на тестовой выборке у рассмотренных наборов данных.

Список литературы

[1] Meleksen Akin, Sadiye Peral Eyduran, Ecevit Eyduran, and Barbara M Reed. Analysis of macro nutrient related growth responses using multivariate adaptive regression splines. Plant Cell, Tissue and Organ Culture (PCTOC), 140:661–670, 2020.

[2] Sergey Bakin, Markus Hegland, and Michael R Osborne. Parallel mars algorithm based on b-splines. Computational Statistics, 15:463–484, 2000.

[3] Christian Conoscenti, Valerio Agnesi, Mariaelena Cama, Nathalie Alamaru Caraballo-Arias, and Edoardo Rotigliano. Assessment of gully erosion susceptibility using multivariate adaptive regression splines and accounting for terrain connectivity. Land degradation & development, 29(3):724–736, 2018.

[4] David GT Denison, Bani K Mallick, and Adrian FM Smith. Bayesian mars. Statistics and Computing, 8:337–346, 1998.

[5] Jerome H Friedman. Multivariate adaptive regression splines. The annals of statistics, 19(1):1–67, 1991.

[6] Salim Heddam and Ozgur Kisi. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and m5 model tree. Journal of Hydrology, 559:499–509, 2018.

[7] Xinglong Ju, Victoria CP Chen, Jay M Rosenberger, and Feng Liu. Fast knot optimization for multivariate adaptive regression splines using hill climbing methods. Expert Systems with Applications, 171:114565, 2021.

[8] Diana L Martinez, Dachuan T Shih, Victoria CP Chen, and Seoung Bum Kim. A convex version of multivariate adaptive regression splines. Computational statistics & data analysis, 81:89–106, 2015.

[9] Ayşe Özmen, Gerhard Wilhelm Weber, İnci Batmaz, and Erik Kropat. Rcmars: Robustification of cmars with different scenarios under polyhedral uncertainty set. Communications in Nonlinear Science and Numerical Simulation, 16(12):4780–4787, 2011.

[10] Sanjiban Sekhar Roy, Reetika Roy, and Valentina E Balas. Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of mars and elm. Renewable and Sustainable Energy Reviews, 82:4256–4268, 2018.

[11] Mohammad Ali Sahraei, Hakan Duman, Muhammed Yasin Çodur, and Ecevit Eyduran. Prediction of transportation energy demand: multivariate adaptive regression splines. Energy, 224:120090, 2021.

[12] N Busto Serrano, A Suárez Sánchez, F Sánchez Lasheras, Francisco Javier Iglesias-Rodríguez, and G Fidalgo Valverde. Identification of gender differences in the factors influencing shoulders, neck and upper limb msd by means of multivariate adaptive regression splines (mars). Applied Ergonomics, 82:102981, 2020.

[13] Charles J Stone, Mark H Hansen, Charles Kooperberg, and Young K Truong. Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. The Annals of statistics, 25(4):1371–1470, 1997.

[14] Gerhard-Wilhelm Weber, Inci Batmaz, Gülser Köksal, Pakize Taylan, and Fatma Yerlikaya-Özkurt. Cmars: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. Inverse Problems in Science and Engineering, 20(3):371–400, 2012.

[15] Zehua Yang, Victoria CP Chen, Michael E Chang, Melanie L Sattler, and Aihong Wen. A decision-making framework for ozone pollution control. Operations Research, 57(2):484–498, 2009.