УДК 004

АВТОРЕГРЕССИОННАЯ АНАЛИТИКА КОМПАКТНЫХ И МЕТРИЧЕСКИ СЛОЖНЫХ ДАННЫХ

Эль-Иман Ф., студент группы 03-220, I курс
Арабская нефтегазовая академия, г. Абу-Даби

Одним из наиболее алгоритмически эффективных способов разрешить задачу прогнозирования целевого столбца по относительно большому множеству прецедентов ($> 10^2$), является построение модели прикладного искусственного интеллекта, основанной на типе жадного добавления [1].

Шаговая регрессия – это один из представителей семейства алгоритмов жадного добавления [2-7].

Также, от всего семейства алгоритмов полного перебора, шаговая регрессия выгодно отличается тем, что её применимость не является узкой. Например, когда в наборе данных количество признаков будет достигать 1000 (а оптимальным в алгоритмах полного перебора состав признаков равен ~100), то модели полного перебора будут очень долго проходить обучение, вне зависимости от типа предоставленного им аппаратного обеспечения и вычислительных мощностей [8, 9].

Неэффективность алгоритмов полного перебора вынудила искать новую эвристику для создания иной концептуальной модели машинного обучения. Первоначальной целью шаговой регрессии в формализованном математическом языке была идея поиска оптимума параметров обобщающей функции ещё быстрее, чем в алгоритмах полного перебора, но за счет некоторой дополнительной аппроксимации.

В общем виде эту задачу можно описать следующим образом (1).

$$J_0: \emptyset;\ Q^* \coloneqq Q(\emptyset); \quad (1)$$

В условиях (1) поставлена цель оптимизировать Q(J), где J – это подмножество из конечного множества. Эта задача NP-трудная, то есть в общем случае необходимо перебрать все 2^N вариантов, чтобы гарантированно получить решение.

Но на практике приходится создавать оптимизацию, чтобы алгоритмическая сложность решения не вызывала больших временных затрат, требуемых на решение задачи.

Программная (алгоритмическая) аппроксимация основана на следующих знаниях:

1. Данные имеют нижнюю огибающую, которая, в свою очередь, имеет минимум. Нам нужно как можно скорее (приближенно) выяснить форму нижней огибающей функции данных;

2. Предположение о том, что если добавить/исключить один признак из данных, то функция Q(J) изменится не очень сильно. Это и есть аппроксимация: если мы будем совершать локальные изменения множества J, то мы будем приближаться к некоторому оптимуму искомой функции.

Самый очевидный и простой способ реализовать концепт шаговой регрессии – это поочередное добавление признаков в цикле для всех $j = 1, \dots, n$, где $j -$ это сложность наборов:

Находим признак, который наиболее выгоден для добавления к исходному набору (то есть к приближению) согласно ограничениям (2).

$$f^*: arg\ min_{f\ \epsilon\ F \setminus J_{j-1}} Q\left(J_{j-1} \cup \{f\}\right) \quad (2)$$

После этого, согласно условиям (2) добавления (f проходит всё множество признаков F $f\ \epsilon\ F \setminus J_{j-1}$), добавляем признак в набор по новым условиям (3):

$$\begin{cases} J_j \coloneqq J_{j-1} \cup \{f^*\}; \\ \text{если } Q(J_j) < Q^*, \text{то } j^* \coloneqq j;\ Q^* \coloneqq Q(J_j); \\ \text{если } j - j^* \geq d, \text{то вернуть } J_{j^*}; \end{cases} \quad (3)$$

Из ограничений (3) получается, что, если в случае полного перебора, для каждого значения j строятся все возможные наборы признаков с данной мощностью, то здесь (2) совершается гораздо меньше математических операций: строятся не все наборы, а только те, которые появляются за счет добавления нового признака $J_{j-1} \cup \{f\}$. Таким образом, экономия ресурсов на каждом шаге составляет N – J.

Главной целью условий (1) является поиск самой нижней точки огибающей функции данных (глобальный минимум множества), при этом эвристики (1), (2) позволяют нам в программном виде добиваться того, чтобы в явном виде это множество никогда не строить [10, 11].

Список литературы:

1. Свидетельство о государственной регистрации программы для ЭВМ № 2023680124 Российская Федерация. BrainPower : № 2023669010 : заявл. 16.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QXBJIM.
2. Математические и программные методы построения моделей глубокого обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 176 с. – ISBN 978-5-9729-1484-5. – EDN PZLUAH.

3. Свидетельство о государственной регистрации программы для ЭВМ № 2023680229 Российская Федерация. Direct Computer Vision Model : № 2023669589 : заявл. 25.09.2023 : опубл. 27.09.2023 / Р. В. Майтак. – EDN DAKXRI.
4. Методы восстановления непараметрической регрессии в условиях несбалансированных данных / А. Д. Салычева и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 192 с. – ISBN 978-5-9729-1856-0. – EDN AAJATW.
5. Свидетельство о государственной регистрации программы для ЭВМ № 2023680070 Российская Федерация. Модернизированная модель DBSCAN для определения скрытых взаимосвязей : № 2023668841 : заявл. 13.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN KQUUKF.
6. Асимптотический анализ поведения прикладных моделей машинного обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.
7. Свидетельство о государственной регистрации программы для ЭВМ № 2023680103 Российская Федерация. Cognitive Solution : № 2023669189 : заявл. 19.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QEMFJA.
8. Свидетельство о государственной регистрации программы для ЭВМ № 2024610040 Российская Федерация. Программа для расчета значений необходимых критериев условия отсутствия конвекции : № 2023689238 : заявл. 25.12.2023 : опубл. 09.01.2024 / П. А. Пылов. – EDN AOTYDS.
9. Свидетельство о государственной регистрации программы для ЭВМ № 2024610076 Российская Федерация. Программа для выполнения атомно-абсорбционной спектроскопии тяжелых нефтей : № 2023688334 : заявл. 17.12.2023 : опубл. 09.01.2024 / П. А. Пылов. – EDN IGYYIO.
10. Свидетельство о государственной регистрации программы для ЭВМ № 2024610078 Российская Федерация. Программа для выполнения проявительного анализа тяжелой нефти : № 2023688309 : заявл. 16.12.2023 : опубл. 09.01.2024 / П. А. Пылов. – EDN LMNACM.
11. Свидетельство о государственной регистрации программы для ЭВМ № 2024610187 Российская Федерация. Программа для выполнения аналитики тяжелой нефти методом электронного парамагнитного

резонанса : № 2023688401 : заявл. 18.12.2023 : опубл. 09.01.2024 / П. А. Пылов. – EDN BBPQGQ.