УДК 004.8

ФААИД ЭЛЬ-ИМАН, ст. группы 03-220
(Арабская нефтегазовая академия, КФУ)
г. Абу-Даби
ШИЯНОВ К. Е., ст. группы ИИм-231 (КузГТУ)
Научный руководитель: БАУМГАРТЭН М.И. (КузГТУ)
г. Кемерово

СТАНДАРТИЗИРОВАННАЯ МЕТОДОЛОГИЯ РЕГРЕССИОННОЙ АНАЛИТИКИ

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].

От остального семейства алгоритмов полного перебора шаговая регрессия выгодно отличается тем, что её применимость не является слишком узкой. Например, когда в наборе данных количество признаков будет достигать 1000 (а оптимальный состав признаков в алгоритмах полного перебора равен ∼100), то модели полного перебора будут очень долго проходить обучение вне зависимости от типа предоставленного им аппаратного обеспечения и вычислительных мощностей [3].

Неэффективность алгоритмов полного перебора вынуждает искать новую эвристику для создания иной концептуальной модели машинного обучения. Первоначальной целью шаговой регрессии в формализованном математическом языке был поиск оптимума параметров обобщающей функции, причём ещё более быстрый, 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. Однако 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, хотя и 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 вместе с этим 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. Для этого делаем ряд следующих шагов:

— 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 [4, 5].

Список литературы:

1. Яцевич, М. Ю. Формирование модели сильного искусственного интеллекта на основе принципа "Congruit universa" для решения геомеханической задачи методом межскважинного сейсмоакустического просвечивания / М. Ю. Яцевич, П. А. Пылов, А. В. Дягилева // Вестник научного центра по безопасности работ в угольной промышленности. – 2022. – № 4. – С. 14-19. – EDN JOZUTB.
2. Свидетельство о государственной регистрации программы для ЭВМ № 2024611038 Российская Федерация. Sliding Window Predictor LLM : № 2023689243 : заявл. 25.12.2023 : опубл. 17.01.2024 / П. А. Пылов. – EDN AGNEAO.
3. Свидетельство о государственной регистрации программы для ЭВМ № 2023669862 Российская Федерация. Insight IQ : № 2023669164 : заявл. 19.09.2023 : опубл. 21.09.2023 / Р. В. Майтак. – EDN YCLTHP.
4. Свидетельство о государственной регистрации программы для ЭВМ № 2024611019 Российская Федерация. Multi-view generation of 3D objects based on diffusion model : № 2024610024 : заявл. 02.01.2024 : опубл. 17.01.2024 / П. А. Пылов. – EDN CQQJRO.
5. 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.