УДК 004.8:64.011.3

УПРАВЛЕНИЕ РЕШЕНИЕМ ЗАДАЧ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ В УСЛОВИЯХ ОГРАНИЧЕНИЙ

В.А. Мулюха
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, Политехническая ул., 29
E-mail: vladimir.muliukha@spbstu.ru

А.В. Востров
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, Политехническая ул., 29
E-mail: vostrov_av@spbstu.ru

Д.Е. Моторин
Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, Санкт-Петербург, Политехническая ул., 29
E-mail: motorin_de@spbstu.ru

Ключевые слова: машинное обучение, ограниченный ресурс, малая выборка, время вычисления, предобработка.

Аннотация: Технологии искусственного интеллекта и в частности машинного обучения являются одними из самых динамично развивающихся и перспективных в современном мире. Они позволяют получить решение задач, которые до недавнего времени были исключительно прерогативой человека. Однако при решении многих практических задач приходится преодолевать ряд сложностей, многие из которых связаны с ограничениями на ресурсы, доступные при решении задачи, при этом, ресурсы могут быть как вычислительные и временные (т.е. задача должна быть решена за определенное время и с использованием определенного аппаратного обеспечения, чаще всего речь идет о различных мобильных платформах), так и информационные, когда речь идет о малых, цензурированных, неполных или зашумленных данных. В докладе рассматриваются теоретические основы и практические кейсы решения промышленных задач методами машинного обучения в условиях ограничений.

1. Введение

Машинное обучение – это одна из самых быстрорастущих областей в современной науке. Оно представляет собой процесс, при котором вычислительное устройство анализирует доступные данные и автоматически обучаются делать прогнозы и принимать решения в различных областях, таких как медицина, финансы, производство и т.д. Традиционно одним из главных преимуществ машинного обучения является его способность обрабатывать большие объемы данных, которые раньше были недоступны для анализа. Однако, как и любая другая технология, машинное обучение имеет свои ограничения и недостатки. Некоторые из них включают в себя проблему переобучения, когда модель начинает обучаться на случайных шумах и ошибках, а не на реальных закономерностях. Также проблемой является наличие малых данных для обучения, когда объемов выборки недостаточно, чтобы выявить статистически значимые закономерности. В связи с развитием больших моделей нейронных сетей, в частности

больших языковых моделей, на первый план выходит проблема производительности и стоимости работы моделей. Также для ряда областей критически важной остается проблема доверия к моделям машинного обучения, так как они базируются на обучающих данных и, таким образом, могут быть предвзятыми или неполными.

Существует ряд подходов к решению проблемы использования моделей машинного обучения в условиях ограничений на ресурсы. В частности, при наличии ограничений на объем обучающей выборки можно использовать методы регуляризации, которая позволяет уменьшить влияние шума и увеличить точность модели [1]. Другим подходом является использование ансамблевых методов, которые позволяют заменить одну сложную модель, имеющую большое число параметров и требующую большой объем обучающей выборки, на множество простых моделей, обученных на разных данных или на различных комбинациях исходных данных. При этом каждая из таких простых моделей имеет небольшое количество параметров и может быть обучена на сравнительно небольшой выборке. Ошибки решения задачи, совершаемые отдельными простыми моделями, нивелируются с помощью коллегиальных методов принятия решений, тем самым увеличивают точность и надежность модели. Также можно увеличивать объем выборки используя специализированные модели машинного обучения, например, сиамские сети, которые учатся не на самих примерах, а на их парах, при этом число пар растет пропорционально квадрату элементов выборки и позволяет осуществлять обучения даже на небольшом числе примеров.

2. Решение задачи сравнения объектов в условиях ограничения на объем исходных данных

2.1. Ограниченный размер обучающей выборки

Одной из практических задач, решаемой в условиях ограничений на объем обучающей выборки, была разработка системы сравнения формы трехмерных объектов для Роспатента [2]. Задача, решаемая при помощи нейронной сети в проекте, – вычисление функции семантической близости или определение схожести цифровых трехмерных моделей. Сравнение объектов с точки зрения их семантической близости в теории машинного обучения относится к классу задач, которые объединены названием distance metric learning [3]. Основная идея, лежащая в основе решения задачи сравнения объектов, заключается в том, что расстояние между семантически близкими объектами должно быть меньше, чем расстояние между семантически различными объектами.

Таким образом, если имеются два вектора обучающей выборки $x_i \in \mathbf{R}^m$ и $x_j \in \mathbf{R}^m$, то расстояние $d(x_i, x_j)$ должно минимизироваться, если x_i и x_j являются семантически близкими объектами, и это расстояние должно максимизироваться, если x_i и x_j являются семантически далекими. Наиболее общей и популярной функцией расстояния является квадратичное расстояние Махаланобиса $d_M^2(x_i, x_j)$, которое определяется для двух векторов как $d_M^2(x_i, x_j) = (x_i - x_j)^T M(x_i - x_j)$. Здесь $M \in \mathbf{R}^{m \times m}$ – симметричная положительно полуопределенная матрица (ее собственные числа являются неотрицательными). Расстояние Махаланобиса, в отличие от расстояния Евклида, позволяет учесть корреляции между входными данными и инвариантно к масштабу.

Матрица расстояний является симметричной и положительно определенной, что позволяет представить ее в виде:

$$S^{-1} = W^T W, W \in \mathbf{R}^{p \times M}, p \leq M.$$

Тогда расстояние Махаланобиса можно переписать как

$$d_M^2(x_i, x_j) = \left\| W_{x_i} - W_{x_j} \right\|_2.$$

Приведенное равенство означает, что вычисление расстояние Махаланобиса эквивалентно поиску такого линейного преобразования W, что каждый вектор отображается в пространство меньшей размерности, в котором евклидово расстояние равно расстоянию Махаланобиса в исходном пространстве. Благодаря ему может быть построена нейронная сеть, которая функционирует аналогично вычислению расстояния Махаланобиса, когда функции активации нейронов являются линейными. Фактически нейронная сеть позволяет получить нелинейный аналог расстояния Махаланобиса посредством использования комбинации линейных и нелинейных функций активации.

Однако, для эффективного функционирования объем обучающей выборки для нейронных сетей должен превышать количество весов соединений (обучающих параметров) в несколько раз. Для решения данной проблемы в проекте использовалась сиамская нейронная сеть, которая обучается на парах данных. Таким образом, наличие всего 600 исходных трехмерных моделей позволило обучить сиамскую нейронную сеть, состоящую из двух шестислойных полносвязных нейронных сетей, содержащих 96, 150, 120, 80, 40, 32 и 16 нейронов соответственно. Второй особенностью сиамской сети является возможность ее использования в так называемых one-shot learning (обучение на одном примере) или few-shot learning (обучение на нескольких примерах). В данном случае имеется в виду не обучение в действительности на одном примере, а обучение в ситуации, когда в одном классе имеется только один или несколько примеров. Такая возможность достигается за счет того, что сиамская сеть обучается не на отдельных примерах, а на парах примеров. Отсюда не количество самих объектов, а количество возможных пар объектов являются элементами для обучения.

2.2. Ограниченные ресурсы для решения задачи

Второй проблемой, решаемой в рамках данного проекта, было обеспечение требуемого уровня производительности при сравнении формы трехмерных моделей. Сами модели поступали в систему в форме каркасных моделей произвольного размера (вплоть до 100 МБ на одну модель). При этом сравнение вновь поступившей модели со всеми существующими в базе данных не должно, по требованиям заказчика, длиться более 15 секунд. В связи с ограничениями на использование аппаратной части и требованием обеспечения функционирования системы в контуре заказчика был предложен переход от прямого сравнения трехмерных объектов к сравнению дескрипторов их формы, обеспечивающих выявление существенных особенностей внешнего вида трехмерных объектов.

Для формирования дескриптора формы трехмерного объекта использовалась трехмерная модификация метода хорд, который позволял перейти от трехмерной модели, состоящей из произвольного количества точек, к диаграмме длин отрезков, соединяющих точки на поверхности модели. Сами дескрипторы формировались в асинхронном режиме до момента поступления модели в систему, а в процессе работы эксперта осуществлялось только сравнение рассчитанного дескриптора новой модели с ранее рассчитанными дескрипторами других моделей в базе данных. В системе предусмотрена возможность смены дескриптора для сравнения. В случае появления нового дескриптора для каждой поступающей модели рассчитываются как старый, так и новый дескрипторы пока все имеющиеся в базе модели не будут пересчитаны. При этом, до полного перерасчета, сравнение моделей производится с использованием старых дескрипторов, о чем выводится соответствующая информация пользователю, а после перерасчета используются только новые и старые больше не рассчитываются.

Перенос процедур обработки данных в фоновый режим позволил существенно сократить требования к аппаратному обеспечению разработанной системы, а также обеспечить при этом требуемый уровень производительности.

В другом проекте при разработке системы интеллектуальной диспетчеризации задач на суперкомпьютере использовался обратный подход. Спецификой задачи там являлась необходимость предсказания времени решения задач на суперкомпьютере методами машинного обучения с использованием информации о выполнении близких по метаданным задач, а также всех других задач данного пользователя [4]. Особенностью постановки задач в суперкомпьютерном центе является возможность их пакетного запуска. В таком случае использование методов предварительной обработки данных вызывали ощутимую задержку при единовременной постановке в очередь большого количества задач. Часто эти задачи не большие и при наличии свободных ресурсов могут быть решены за время сопоставимое со временем их обработки в интеллектуальном модуле. В данном случае мы использовали постобработку данных, то есть задачи от пользователя сразу попадают в базу данных очереди диспетчера суперкомпьютерного центра, а наша система в асинхронном режиме корректирует ожидаемое время выполнения этих задач и при очередном переупорядочивании задач диспетчер учитывает изменившиеся параметры и «уплотняет» или меняет порядок выполнения задач. При этом в самой модели машинного обучения, которая осуществляет предсказание времени выполнения задачи используется ансамблевое обучение, при котором несколько простых моделей, например, небольших случайных лесов, обучаются на случайных подвыборках исходной выборки данных. Использование ансамблевых моделей позволяет существенно снизить требования к ресурсам, необходимым на стадии обучения модели, а также ускорить саму процедуру обучения.

3. Заключение

Использование машинного обучения в условиях малых выборок является сложной задачей, требующей тщательного анализа данных и выбора оптимальных методов. Однако, благодаря развитию новых методов и технологий, возможности использования машинного обучения в таких случаях постоянно расширяются.

Существует ряд методов, позволяющих использовать модели машинного обучения даже в условиях наличия ограничений на используемые ресурсы, будь то объемы данных, используемых для обучения модели, или ресурсы, необходимые для ее обучения или функционирования. Наиболее типовым приемом является использование специализированных моделей, поддерживающих методы few-shot learning или даже one-shot learning, когда процедура обучения происходит по экстремально небольшому количеству данных. Также, при наличии соответствующей возможности можно использовать предобработку исходных данных или постобработку полученных результатов, зачастую при помощи других моделей машинного обучения. Использование методов постобработки, помимо повышения эффективности работы модели машинного обучения, может использоваться для реализации механизмов интерпретации (или объяснения) результатов работы модели [5].

Список литературы

1. Boyd S., Vandenberghe L. Convex Optimization. Cambridge University Press, 2004. 716 p.
2. Научная Россия: 3D-система для Роспатента. https://rospatent.gov.ru/ru/news/nr-3d-sistema-dlya-rospatenta-280121 (дата обращения 20.01.2024).

3. Suárez J., García S., Herrera F. A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms and Software. 2018. https://arxiv.org/abs/1812.05944 (дата обращения 20.01.2024).
4. Zaborovsky V.S., Utkin L.V., Muliukha V.A., Lukashin A.A. Improving Efficiency of Hybrid HPC Systems Using a Multi-agent Scheduler and Machine Learning Methods // Supercomputing Frontiers and Innovations. 2023. Vol. 10, No. 2. P. 104-126. DOI: 10.14529/jsfi230207.
5. Kovalev M.S., Utkin L.V., Kasimov E.M. SurvLIME: A method for explaining machine learning survival models // Knowledge-Based Systems. 2020. Vol. 203. P. 106164. DOI: 10.1016/j.knosys.2020.106164.