Применение активного обучения для повышения качества системы предсказания алгоритмов классификации

Забашта А. С., программист факультета информационных технологий и программирования Университета ИТМО, azabashta@corp.ifmo.ru

Фильченков А. А., доцент факультета информационных технологий и программирования Университета ИТМО, afilchenkov@corp.ifmo.ru

Аннотация

Мета-обучение использует мета-признаки, чтобы формально описывать наборы данных и находить возможные зависимости производительности алгоритмов от них. Но различных наборов данных недостаточно, чтобы заполнить пространство мета-признаков с приемлемой плотностью для будущего прогнозирования производительности алгоритмов. Для решения этой проблемы мы можем использовать активное обучение. Но для этого требуется способность генерировать нетривиальные наборы данных, которые могут помочь улучшить качество системы мета-обучения. В этой статье мы экспериментально сравниваем несколько таких подходов, основанных на максимизации разнообразия и байесовской оптимизации.

Введение

На сегодняшний день не существует одного идеального алгоритма классификации [1]. Для разных наборов данных оказываются полезными различные алгоритмы. Для решения данной задачи применяется подход мета-обучения [2]. Суть этого подхода в описании наборов данных численными характеристиками, которые называются мета-признаками, и поиск зависимостей производительности алгоритмов от них [3].

Одна из проблем мета-обучения заключается в том, что реальных наборов данных недостаточно для заполнения пространства мета-признаков с приемлемой плотностью. Чтобы решить эту проблему, мы используем стратегию активного обучения [4] путем постепенной оцен-

ки неисследованных областей и заполнения их недостающими наборами данных.

Генерация наборов данных

Общим подходом к задаче генерации наборов данных является сведение её к задаче минимизации, в которой целевой функцией является расстояние между целевым и результирующим наборами данных в пространстве мета-признаков.

Метод DIRECT

Метод **DIRECT** основан на идее непосредственного отображения набора данных в вектор чисел и обратно [5]. Первоначальный подход был в состоянии генерировать только наборы данных фиксированного размера. Но мы изменили его, добавив несколько измерений в числовое пространство поиска, которые определяют количество классов, атрибутов и количество объектов каждого класса. Для этого мы рассчитываем максимальное количество атрибутов, классов и объектов на класс среди всей коллекции используемых наборов данных. Таким образом, большинство обработанных векторов в пространстве поиска содержат избыточную информацию, которая отбрасывается при преобразовании в набор данных.

Метод GMM

Подход **GMM** [6] аналогичен предыдущему подходу. Но он вместо объектов хранит ковариационную матрицу и вектор сдвига для многомерного гауссовского распределения, из которого затем отбираются объекты для соответствующего класса. Мы модифицировали этот подход так же, как и предыдущий.

Метод NDSE

Метод NDSE является модификацией аналогичного метода из нашего предыдущего исследования [7]. Первоначально этот метод мог генерировать только набор данных для задачи бинарной классификации, но мы обобщили его на любое количество классов.

Этот метод также основан на эволюционных алгоритмах. Однако он использует естественные для классификации операции изменения наборов данных, такие как добавление и удаление атрибутов или объектов. Этот метод может одновременно работать с наборами данных произвольного размера.

Активное обучение

Мы используем стратегии активного обучения для создания новых синтетических наборов данных, чтобы повысить производительность системы мета-обучения во время ее обучения.

Случайная генерация

Мы будем ссылаться на подход **RAND** как на базовый подход, который просто генерирует случайные наборы данных.

Максимизация разнообразия

Подход **DIV** [8] генерирует новые наборы данных в попытке увеличить разнообразие в пространстве мета-признаков. Разнообразие DIV наборов данных $\{D_i\}$ определяется как $DIV(\{D_i\}) = \sum_{D_x} \frac{\min_{D_y} \mu(D_x, D_y)}{\#\{D_i\}}$, где μ — функция расстояния в пространстве мета-признаков.

Таким образом, для этого метода стратегия генерации заключается в максимизации $\min_{D_x} \mu(D_x, D_{new})$.

Максимизация неопределенности

Стратегия генерации **VAR** основана на байесовской оптимизации [9]. Но поскольку у нас есть прямой доступ к мета-системе, мы можем попытаться сгенерировать наборы данных в тех областях, где такая система больше всего не уверена.

Как и в оригинальной работе, мы основывали нашу мета-систему на классификаторе случайных лесов, который состоит из нескольких деревьев решений, которые обучаются независимо. Таким образом, мы можем использовать дисперсию ответов нескольких классификаторов как неопределенность мета-системы.

Эксперимент с генерацией наборов данных

Мы экспериментально сравнили описанные подходы генерации наборов данных для задачи классификации: **DIRECT**, **GMM** и **NDSE**. Для этого эксперимента мы использовали 371 набор данных из OpenML [10]. Мы выполнили перекрестную проверку [11] для этой коллекции наборов данных. Для каждого контрольного набора данных мы попытались создать набор данных с аналогичными характеристиками, используя 10 эволюционных алгоритмов из библиотеки jMetal [12]. В качестве функции ошибки мы использовали расстояние Махаланобиса [13] между целевым и полученным наборами данных в метапризнаковом пространстве.

Результаты

Результаты эксперимента по генерации наборов данных показаны в Таблице 1. Как можно заметить, если эволюционный алгоритм начинает со случайных наборов данных в качестве начальной популяции, лучше использовать подход **NDSE**, в то время как **DIRECT** лучше при работе с реальными наборами данных.

Таблица 1: Результат генерации наборов данных разными подходами с разными эволюционными алгоритмами: среднее расстояние Махаланобиса. Первый и второй большие столбцы отвечают за тесты без и с использованием реальных наборов данных соответственно.

	DIRECT	GMM	NDSE	DIRECT	GMM	NDSE
CMAES	—	4.3158	—	—	4.4002	—
DE	4.0883	3.7022	—	3.0721	3.7407	—
GDE3	3.8260	3.4788	—	3.3606	3.6214	—
MOCell	3.5173	3.2745	3.6275	**1.9398**	3.5640	2.9385
NSGAII	3.6567	3.2604	2.8400	2.1929	3.5337	2.1634
RS	4.2800	3.9901	3.9514	3.1065	4.2174	3.1673
SMSEMOA	3.7648	3.3472	2.9062	2.3113	3.6596	2.1772
SPEA2	3.7569	3.2677	**2.8160**	2.2488	3.5773	2.1107
SPSO11	3.3951	3.1893	—	1.9530	3.1141	—
gGA	3.8322	3.4008	3.0211	2.6363	3.7488	2.5554

Эксперимент с активным обучением

В этом эксперименте мы проверили подходы генерации наборов данных со стратегиями активного обучения. В соответствии с предыдущим экспериментом мы выбрали лучшие эволюционные алгоритмы для каждого подхода к генерации: **MOCell** для **DIRECT**, **SPSO11** для **GMM** и **SPEA2** для **NDSE**. Мы протестировали эти подходы с помощью описанных стратегий генерации наборов данных: **RAND**, **DIV** и **VAR**.

В каждом тесте мы обучали мета-систему, которая предсказывала F-меру классификатора KNN для набора данных после повторённой 10 раз перекрестной проверки.

Для каждой стратегии активного обучения и метода генерации было сгенерировано 300 новых наборов данных, на которых обучалась система мета-обучения.

Результаты

Результаты эксперимента с активным обучением показаны в Таблице 2. Для каждой пары из подхода и стратегии генерации представлено минимальное среднее RMSE мета-системы. Как можно видеть, подход **NDSE** был лучше для стратегий **DIV** и **VAR**. Комбинация **NDSE** и **VAR** превзошла другие подходы.

Таблица 2: Результаты эксперимента с активным обучением: минимальное среднее RMSE мета-системы.

	DIRECT	GMM	NDSE
RAND	0.1751	0.1790	0.1765
DIV	0.1723	0.1798	0.1633
VAR	0.1777	0.1799	0.1597

Заключение

В этой статье мы исследовали 3 подхода к генерации наборов данных для задачи классификации и 3 стратегии для создания активных наборов данных для повышения качества системы мета-обучения.

Мы экспериментально сравнили их в задаче создания существующих наборов данных. Было показано, что для генерации наборов дан-

ных без реальных наборов данных лучше использовать подход NDSE, тогда как подход DIRECT лучше при использовании с реальными наборами данных.

Мы использовали исследуемые подходы генерации наборов данных для создания системы активного обучения для улучшения мета-системы, которая предсказывает качество классификации. Эксперимент показал, что стратегия максимизации неопределенности оказалась лучше максимизации разнообразия.

Таким образом, было показано, что, применяя методы генерации наборов данных, мы можем использовать стратегии активного обучения для улучшения качества системы мета-обучения.

Литература

[1] Wolpert D. H. The supervised learning no-free-lunch theorems // Soft Computing and Industry. — Springer, 2002. — С. 25–42.

[2] Metalearning: Applications to data mining / P. Brazdil [и др.]. — Springer Science & Business Media, 2008.

[3] Giraud-Carrier C. Metalearning-a tutorial // Tutorial at the 7th international conference on machine learning and applications (ICMLA), San Diego, California, USA. — 2008.

[4] Settles B. Active learning // Synthesis Lectures on Artificial Intelligence and Machine Learning. — 2012. — Т. 6, No 1. — С. 1–114.

[5] Reif M., Shafait F., Dengel A. Dataset generation for meta-learning // Poster and Demo Track of the 35th German Conference on Artificial Intelligence (KI-2012). — 2012. — С. 69–73.

[6] Muñoz M. A., Smith-Miles K. Generating Custom Classification Datasets by Targeting the Instance Space // Proceedings of the Genetic and Evolutionary Computation Conference Companion. — Berlin, Germany : ACM, 2017. — С. 1582–1588. — (GECCO '17). — ISBN 978-1-4503-4939-0.

[7] Zabashta A., Filchenkov A. NDSE: Instance Generation for Classification by Given Meta-Feature Description // CEUR Workshop Proceedings. Т. 1998. — 2017. — С. 102–104.

[8] Abdrashitova Y., Zabashta A., Filchenkov A. Spanning of Meta-Feature Space for Travelling Salesman Problem // Procedia Computer Science. — 2018. — T. 136. — C. 174–182.

[9] Hutter F., Hoos H. H., Leyton-Brown K. Sequential model-based optimization for general algorithm configuration // International Conference on Learning and Intelligent Optimization. — Springer. 2011. — C. 507–523.

[10] OpenML: networked science in machine learning / J. Vanschoren [и др.] // ACM SIGKDD Explorations Newsletter. — 2014. — T. 15, No 2. — C. 49–60.

[11] A study of cross-validation and bootstrap for accuracy estimation and model selection / R. Kohavi [и др.] // Ijcai. T. 14. Вып. 2. — Montreal, Canada. 1995. — C. 1137–1145.

[12] Durillo J. J., Nebro A. J., Alba E. The jMetal framework for multi-objective optimization: Design and architecture // IEEE congress on evolutionary computation. — IEEE. 2010. — C. 1–8.

[13] Mahalanobis P. C. On the generalized distance in statistics // Proceedings of National Institute of Sciences (India). — 1936. — T. 2, No 1. — C. 49–55.