Общие и ансамблевые модели машинного обучения в задачах спектроскопии

А.О. Ефиторов[1], О.Э. Сарманова[1,2], К.А. Лаптинский[1,2], С.А. Буриков[1,2], Т.А. Доленко[1,2], С.А. Доленко[1]

[1]*Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына, Ленинские горы 1, стр. 2, Москва, Россия, 119991*
[2]*Московский государственный университет им. М.В. Ломоносова, Ленинские горы 1, стр. 2, Москва, Россия, 119991*

Аннотация
В статье рассматривается вопрос о влиянии на итоговый результат выбора как методов физических измерений, так и математической модели машинного обучения, решающей обратную задачу оптической спектроскопии биологической среды. Основным выводом проделанной работы является критическая зависимость результатов работы методов от исследуемого объекта в биологической среде (урина) и в конечном счете от статистики собранных данных. Продемонстрировано, что стационарность фоновых свойств среды (спектры поглощения) оказывается важнее наличия интенсивных собственных полос исследуемых объектов (спектры флуоресценции), особенно при решении задачи единой аппроксимационной моделью (нейронные сети, хемометрические модели). Ситуация может быть обратной лишь при решении задачи комитетом кусочных аппроксимационных моделей (деревья решений).

Ключевые слова
Спектры поглощения, спектры флуоресценции, углеродные точки, нейронные сети, градиентный бустинг, обратные задачи

1. Введение

В представленной работе приведено решение обратной задачи (ОЗ) по определению концентраций компонентов объекта с помощью лазерной флуоресцентной спектроскопии (ЛФС) и спектроскопии оптического поглощения (СОП) и моделей машинного обучения, построенных на основе собранных данных. Конечной целью работы является точный контроль выведения с уриной нанокомплексов, состоящих из углеродных точек (УТ) и противоракового препарата – доксорубицина (Д), для чего требуется определить их концентрации. Несмотря на то, что УТ обладают стабильной люминесценцией и отличными сорбционными свойствами, сама по себе биологическая среда (урина) обладает достаточно интенсивной и весьма вариабельной широкополосной люминесценцией.

2. Экспериментальная часть

Физический эксперимент проводился аналогично описанному в [1]. Исследуемый объект - 624 суспензии УТ и Д в урине, разбавленной дистиллированной водой в 10 раз, диапазон концентраций для УТ 0-1.2 мг/л (шаг: 0.05 мг/л), для Д 0-1 мг/л (шаг 0.042 мг/л).

2.1. Физические измерения

Возбуждение сигнала флуоресценции осуществлялось с помощью диодного лазера с длиной волны 405 нм и 532 нм, запись спектра с разрешением 2 нм производилась в диапазонах

420-800 нм и 565-800 нм, соответственно. Спектры поглощения записывались в диапазоне 190-800 нм с разрешением в 1 нм.

2.2. Вычислительные эксперименты

Для снижения зависимости результатов от статистики тестового набора данных было произведено 5 случайных разбиений массива примеров оригинальных измерений на тренировочный, валидационный и тестовый наборы в соотношении 7:2:1. На Рисунке 1 представлена средняя абсолютная ошибка, рассчитанная на тестовых наборах. Для всех моделей машинного обучения проведен поиск оптимальных внутренних параметров, критерий оптимальности рассчитывался на валидационном наборе.

3. Выводы

Результат подтвердил выводы авторов, полученные в рамках работы над ОЗ спектроскопии ионов металлов [2]. При разделении массива данных на локальные поднаборы и построении «кусочного» решения на основе комитета регрессоров (случайный лес, градиентный бустинг) лучшие результаты могут быть получены и при отсутствии «гладкой» статистики в данных (т.е. когда имеет место значительный разброс значений измерений), что продемонстрировано в данной работе на примере решения ОЗ для спектров флуоресценции УТ. Однако в целом при не слишком большом количестве обучающих примеров преимущество имеют методы, связанные с построением единой аппроксимационной модели, для которых более существенным оказывается отсутствие сильной вариабельности данных (спектры поглощения).

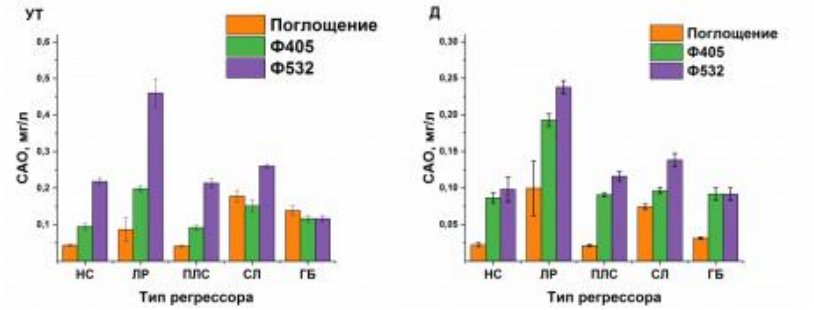

Рисунок 1: Сравнение значений средней абсолютной ошибки, полученной в результате решения обратной задачи спектроскопии по определению концентрации углеродных точек (слева) и доксорубицина (справа) различными методами машинного обучения (НС — нейронные сети, ЛР — линейная регрессия, ПЛС — метод проекций на латентные структуры, СЛ — случайный лес, ГБ — градиентный бустинг) на массиве экспериментальных данных спектров поглощения и флуоресценции (зеленый — возбуждение на длине волны 405 нм, синий — возбуждение на длине волны 532 нм)

4. Благодарности

Исследование выполнено за счёт гранта Российского Научного Фонда (проект №19-11-00333).

5. Литература

[1] Sarmanova, O.E. A method for optical imaging and monitoring of the excretion of fluorescent nanocomposites from the body using artificial neural networks // Nanomedicine: Nanotechnology, Biology, and Medicine. – 2018. – Vol. 14(4). – P. 1371-1380.

[2] Efitorov, A.O. Solution of Multi-parameter Inverse Problem by Adaptive Methods: Efficiency of Dividing the Problem Space // Lecture Notes in Computer Science. – 2017. – Vol. 10614. – P. 751-752.