УДК 54.066

ОПРЕДЕЛЕНИЕ ВИРУСНОГО АНТИГЕННОГО ПРОФИЛЯ МЕТОДАМИ МАШИННОГО И ГЛУБОКОГО ОБУЧЕНИЯ НА ОСНОВЕ ТАБЛИЧНЫХ ДАННЫХ.

Сосновский И.А. (Санкт-Петербургский государственный университет)
isosnovskiii@mail.ru
Научный руководитель – к.х.н., профессор, директор НОЦ инфохимии Скорб Е.В.
(«Национальный исследовательский институт ИТМО»)

В настоящее время, для определения антигенного профиля патогена используют иммунологические методы, основанные на специфической реакции антиген-антитело (ИФА-анализ и его разновидности). Полученный от системы сигнал обрабатывается и анализируется сотрудниками лаборатории - живыми людьми. Для автоматизации рутинного процесса обработки данных, упрощения и снижения стоимости процедуры ИФА-анализа предполагается использование алгоритмов глубокого обучения приспособленных к восприятию табличных данных.

При конвертации сигнала от системы ИФА в данные, наиболее удобным и доступным представлением сигнала является формат табличных данных (xlsx, csv, txt и тд). Накопленные знания о методах машинного обучения свидетельствуют, что в данный момент наиболее впечатляющих успехов среди них достигли алгоритмы глубокого обучения (нейросети). Данный тип алгоритмов прекрасно справляется с задачами компьютерного зрения и обработки естественного языка, но, однако, до сих пор отсутствуют доказательства, что нейросети справляются лучше с предсказаниями на табличных данных.
Традиционно, для табличных данных используются алгоритмы на основе деревьев (случайные леса), популярность такого подхода обусловлена достаточной репрезентативностью деревьев, они хорошо интерпретируются и быстро обучаются. Однако, иерархическое обучение деревьев вызывает проблемы при оптимизации, обычно такие подходы используют жадный поиск параметров и локальную оптимизацию, также деревья не способны обучаться непрерывно и требуют глобальной статистики для выбора узловых точек.
В это же время - глубокие нейронные сети, основанные на алгоритме обратного распространения ошибки, демонстрируют большой потенциал в непрерывном обучении и оптимизации параметров, но при этом часто недостаточно репрезентативны, плохо поддаются интерпретации и долго обучаются.

Для автоматизации обработки результатов проведения иммунологического тестирования предлагается использовать новый тип канонической архитектуры нейросетей (TabNet, NODE). Данные типы нейросетевой архитектуры специально разрабатывались для обработки табличных данных.
NODE состоит из слоев - как и стандартная нейросеть. На каждом слое находится несколько дифференцируемых деревьев, деревья решений обучаются с использованием метода обратного распространения ошибки, выход из дерева является взвешенной суммой листьев.
В TabNet, напротив, каждый слой архитектуры является шагом дерева решений, содержащим в себе блок из нескольких полносвязных слоев нейросети, который занимается определением важности полученной информации на входе.
Таким образом, два вышеперечисленных подхода пытаются совмещать преимущества глубоких нейросетей и случайных деревьев, при этом нивелируя недостатки обоих методов.

Представляется возможным использование новых архитектур при автоматизации обработки табличных данных полученных от систем ИФА, как при анализе оптического сигнала, так и электрохимического.
Особенно стоит отметить возможность обработки вольтамперометрических параметров, полученных от иммуносенсоров на основе печатных электродов, так как данный тип сигнала имеет высокое разрешение и в то же время слишком сложную структуру для обработки вручную. Использование электрохимических иммуносенсоров оправдано рядом преимуществ: относительная дешевизна и простота производства электродов и измеряющего прибора, высокая мобильность измеряющего прибора, быстрый анализ не требующий наличия высокой степени компетенций у человека проводящего его.

Сосновский И.А. (автор)	
Скорб Е.В. (научный руководитель)	