3. Вощинин А.П., Сотиров Г.Р. Оптимизация в условиях неопределенности. – Изд-во МЭИ (СССР); Техника (НРБ), 1989.

4. Оскорбин Н.М., Максимов А.В., Жилин С.И. Построение и анализ эмпирических зависимостей методом центра неопределенности // Известия Алтайского государственного университета. – 1998. – №1. – С. 37–40.

5. Шарый С.П. Конечномерный интервальный анализ. – Новосибирск, 2017.

6. Шарый С.П. Интервальный анализ или методы Монте-Карло? // Вычислительные технологии. – 2007. – Т. 12, №1. – С. 103–115.

7. Dyer M.E., Frieze A.M. On the complexity of computing the volume of a polyhedron // SIAM Journal of Computing 17 (5), 1988, pp. 967–974.

8. Cousins B., Vempala S. A practical volume algorithm // Mathematical Programming Computation 8 (2), 2016, pp 133–160.

УДК 004.89

Применение машинного обучения к задачам анализа историй болезней детей с заболеваниями почек

Д.П. Налимов
АлтГУ, Барнаул

На сегодняшний день такой раздел искусственного интеллекта, как машинное обучение имеет приложения в разнообразных областях знаний. Комбинируя в себе методы математической статистики, оптимизации, теории вероятностей, алгоритмов и графов, алгебры, математического анализа и других наук, машинное обучение позволяет решать множество задач: от кредитного скоринга и построения рекомендаций до генерации изображений и музыкальных композиций. Очень важным и перспективным является применение методов машинного обучения в медицине (в частности, в доказательной медицине). Сюда относится диагностика заболеваний, прогнозирование состояния пациента, создание индивидуальной терапии, проверка эффективности лекарственных препаратов и многое другое.

При применении методов машинного обучения для диагностики заболеваний возникает множество проблем. Например, исследуемые данные неструктурированны: выписки пациентов часто отличатся от шаблона, содержат массу опечаток, неточностей, а порой представлены в рукописном виде. Признаков зачастую огромное количество: различные показатели анализов, данные осмотра пациента, результаты терапий и прочее, поэтому процесс отбора наиболее значимых призна-

ков трудоемок и предполагает сотрудничество аналитика с экспертом (врачом).

В рамках данной работы были рассмотрены порядка 4000 выписок пациентов (с 2008 по 2017 года) с различными заболеваниями почек в формате *.doc, либо *.docx. Основной целью исследования являлось прогнозирование заболевания пациента на основе его показателей, т.е. решение задачи многоклассовой классификации.

Вначале необходимо было обезличить выписки и удалить из них ФИО (заменив его на уникальный идентификатор) и адрес. Уникальный идентификатор зависит непосредственно от ФИО, даты рождения и конкретной выписки (у одного пациента возможно наличие нескольких выписок). Предполагается, что вероятность встретить двух различных пациентов с одинаковыми показателями, которые перечислены ранее, пренебрежительно мала.

Далее необходимо было решить, какие признаки будут участвовать в формировании базы данных и разработать методы их извлечения (парсинг), причем каждое заболевание должно быть зашифровано согласно МКБ-10.

Заболевания почек имеют свои характерные особенности, поэтому совместно с экспертом-врачом было принято решение использовать показатели общего анализа крови (11 признаков), биохимического анализа крови (15 признаков) и общего анализа мочи (7 признаков). На этом этапе возникали серьезные трудности с извлечением признаков, т.к., помимо всевозможных опечаток, многие выписки (особенно разных годов) существенно отличались друг от друга. В ряде случаях значения отсутствовали.

Из 4000 выписок было обработано около 500, на основе которых и проводился анализ. В качестве показателей первоначально были выбраны общий и биохимический анализы крови.

Первым был опробован метод K-NN (K-ближайших соседей). Его довольно часто используют в медицинских задачах, т.к. результат легко интерпретируется: пациенту присваивается такой класс заболевания, который наиболее распространен среди его K ближайших соседей (другие пациенты) по некоторой метрике. К тому же база данных пациентов будет лежать локально на дисковом пространстве, поэтому выбор данного метода вполне обоснован. Однако K-NN выдал хорошую точность (0.86) только на объектах 2 класса (хронический пиелонефрит). Это говорит о том, что он верно распознал 86% объектов данного класса. Однако, точность составила 0.54, т.е. из всех объектов, которые были классифицированы как 2 класс, классификатор верно распознал только 54% объектов 2 класса. Наибольшая точность у 7

класса (гломерулонефрит) – 1.0, т.к. классификатор распознал всего лишь один объект из одного предсказанного, однако полнота составила 0.21, что является плохим результатом.

Согласно f1-мере (0.43) и верности (0.45), результаты классификации неадекватны.

Также в рамках данного исследования был применен метод случайного леса. Этот метод легко интерпретируем и позволяет моделировать стиль мышления врача: на основе сравнения значений показателей предполагать наличие какого-либо диагноза. Данный метод выдал более адекватный результат, за исключением того, что 7 класс не был обнаружен вообще и в качестве ответа не выдавался. f-мера равна 0.46, а верность – 0.53. Однако это не является достаточно хорошим результатом и не дает право говорить о применимости данного метода к задаче.

Для более тщательной проверки результатов методов необходимо наличие всей выборки и всех показателей, а также проверка на выбросы в данных, которые могут повлиять на точность.

В дальнейшем планируется обработать остальные выписки и параметры и добавить их в составленную базу. После этого необходимо будет выявить выбросы в данных, заполнить пропущенные значения (различными методами), применить более сложные методы машинного обучения и нейронные сети, сравнить результаты их работы и сделать визуализацию данных. Также существует необходимость в отборе более значимых признаков во избежание переобучения классификаторов и для улучшения их точности.

Библиографический список

1. Peter Flah Machine Learning. – United Kingdom. Cambridge, 2012.

2. Себастьян Рашка Python и машинное обучение. – Москва, 2017.

3. Abraham Jacob Frandsen Machine learning for disease prediction. – USA, 2016.

4. Tom Schaul, Justin Bayer, Daan Wierstra, Sun Yi, Martin Felder, Frank Sehnke, Thomas Rückstieß, Jürgen Schmidhuber. PyBrain. To appear in: Journal of Machine Learning Research, 2010.