УДК 519.6

Анализ методов машинного обучения для решения задач медицинского профиля

Д.П. Налимов
АлтГУ, Барнаул

Методы машинного обучения применяются в разнообразных областях знаний: психология, генетика, экономика, геология, геофизика и др. В частности, важным и актуальным является их приложение к медицинским задачам. Спектр задач очень разнороден: диагностика заболеваний и состояния пациента, создание индивидуальной терапии, проверка эффективности препарата и т.д. [1].

В данном исследовании использовалась база данных выписок пациентов детского возраста с различными заболеваниями мочеполовой системы Алтайской краевой клинической детской больницы в период с 2010 по 2017 гг. формата *.docx.

Были решены проблемы некорректного представления данных в выписках пациентов и была написана программа, извлекающая необходимую информацию в полуавтоматическом режиме. Заболевания мочеполовой системы имеют свои особенности, поэтому в соответствии с рекомендациями врача были выбраны 33 признака для извлечения и использования в методах машинного обучения, а также 3 диагноза: пиелонефрит, гломерулонефрит и тубулоинтерстициальный нефрит. Размер базы данных составил 2881 пациент.

Целью работы являлся анализ методов машинного обучения для повышения точности и сокращения времени диагностики заболеваний мочеполовой системы у детей.

Анализ проводился на языке программирования python с использованием библиотек Scikit-Learn, NumPy, SciPy, Matplotlib и Pandas [2].

Предварительный анализ данных показал, что распределение заболеваний пациентов имеет дисбаланс, присутствует большое количество пропущенных значений и показателей с коэффициентом вариации близким к нулю. Первую проблему решает сбалансированная метрика качества Matthews correlation coefficient (MCC) [3]. Вторую проблему пытается решить метод заполнения пропусков на основе K ближайших соседей, а третью – метод рекурсивного отбора признаков на базе результата работы соответствующей модели.

После предварительного анализа были применены различные методы машинного обучения: дерево решений (DT), случайный лес (RF), градиентный бустинг (GB), логистическая регрессия (LR), метод K

ближайших соседей (KNN) и многослойный перцептрон (MLP). Результаты их работы были сравнены между собой, а с целью попытки повышения качества были применены: метод рекурсивного отбора признаков на базе результатов работы каждой модели и заполнение пропусков с помощью метода K ближайших соседей.

В таблице 1 указано качество классификации на валидационном наборе данных:

(1) – без отбора признаков, без заполнения пропусков;

(2) – с отбором признаков, без заполнения пропусков;

(3) – без отбора признаков, с заполнением пропусков;

(4) – с отбором признаков, с заполнением пропусков.

Таблица 1 – Качество классификации

	DT	RF	GB	LR	KNN	MLP
(1)	0.433	0.548	**0.591**	0.472	0.408	0.503
(2)	0.385	0.532	**0.584**	0.533	0.429	0.530
(3)	0.330	0.542	**0.545**	0.438	0.369	0.448
(4)	0.314	0.540	**0.547**	0.421	0.427	0.458

В итоге градиентный бустинг дал лучшее качество по сравнению с остальными методами. Метрика качества MCC, рассчитанная на результатах его работы, составила: 0.591 без применения методов заполнения пропусков и отбора признаков, 0.584 без заполнения пропусков и с отбором признаков, 0.545 с заполненными пропусками и без отбора признаков, 0.547 с применением методов заполнения пропусков и отбора признаков.

В ходе исследования значительно повысить качество методов с помощью отбора информативных признаков не удалось, а рассмотренный метод заполнения пропусков в худшую сторону повлиял на результат работы классификации, что говорит о его неприменимости к используемой базе данных пациентов.

Библиографический список

1. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. – М.: ДМК Пресс, 2015. – 400 с.

2. Documentation of scikit-learn 0.19.1 [Электронный ресурс]. –URL: http://scikit-learn.org/stable/documentation.html.

3. Boughorbel S. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric / S. Boughorbel, F. Jarray, M. El-Anbari // PLoS ONE 12(6). – 2017. – 17 p.

УДК 519.16

К вопросу о реализации динамических лабиринтов

Н.В. Павлова, В.В. Смирнов, Т.М. Тушкина
БТИ (филиал) АлтГТУ, г. Бийск

В настоящее время возрастает роль математических методов исследования многих направлений прикладных инженерных наук, в частности, материаловедения [1]. Становится актуальной разработка моделей структур материалов со сложными и ценными физическими и технологическими свойствами. В данной работе в общем виде формулируется задача о поведении некоторой твёрдой частицы в активной многофазной среде. Последняя представляет собой реализацию динамического лабиринта [2]. Одна из фаз среды лабиринта может образовывать непроницаемую для частицы стенку. Также непроницаемыми являются боковые стороны лабиринта.

Частица имеет возможность перемещаться в свободном пространстве случайным образом со скоростью n шагов в единицу времени. При этом лабиринт изменяется на каждом k-ом шаге. В исходных данных присутствует информация о концентрации непроницаемой фазы. Эта концентрация является искомым параметром, определяющим возможность выхода частицы из лабиринта за конечное число шагов [3].

Лабиринт генерируется на двумерной квадратной решетке, каждая ячейка которой может быть либо занята с вероятностью $0 \le p \le 1$, либо свободна с вероятностью $(1 - p)$. Множество занятых ячеек, связанных с ближайшими соседними занятыми ячейками кратчайшим расстоянием, образуют кластеры. Кластер в данном случае будет иметь характер случайного фрактального образования.

Разработана алгоритмическая процедура, которая создаёт матрицу $n \times n$ прямоугольников, каждый из которых частью стенки (закрашивается) только в случае, если сгенерированное для него случайное число p превышает некоторое заданное значение [4].

На рисунке 1 показан результат обращения к разработанной процедуре, реализованный на языке компьютерной математики Maple при $n = 25$ и $s = 0{,}515$.