АНАЛИЗ ДАННЫХ В МЕДИЦИНСКИХ ЗАДАЧАХ

М. Д. Гершов

michaelgershov@yandex.by;

Научный руководитель — А. В. Кушнеров, старший преподаватель

Использование средств, предоставляемых искусственным интеллектом, в медицинской сфере позволяет сократить смертность пациентов за счёт более дешёвой, однако не менее качественной диагностики и обнаружения как инфекционных, так и хронических неинфекционных заболеваний, 80% случаев прогрессирования и заболевания которыми можно предотвратить, если выявлять их на ранней стадии и оказывать своевременное лечение. Целью исследования являлось проведение качественного анализа медицинских данных для построения моделей машинного обучения ради последующего применения полученных результатов для прогнозирования наличия сердечно-сосудистых заболеваний у пациентов реального медицинского учреждения. Обученная на предварительно подготовленных данных модель логистической регрессии показала удивительные 87% процентов recall-точности.

Ключевые слова: искусственный интеллект в медицине; машинное обучение; анализ данных; задача классификации; решающее дерево; случайный лес; логистическая регрессия.

ВВЕДЕНИЕ

Проблема диагностики и раннего выявления различных сердечно-сосудистых заболеваний имеет большее значение для выбора тактики лечения и прогноза развития. При постановке правильного и полного кардиологического диагноза часто требуется привлечение всех шести основных методов обследования: сбор анамнеза, визуальное обследование, электрокардиография, рентгенография грудной клетки, неинвазивные графические методы исследования – эхокардиография, радионуклидное сканирование, а в ряде случаев и специальные инвазивные методы исследования такие, как катетеризация полостей сердца и ангиокардиография.

В рамках исследования не ставилась задача применения алгоритмов анализа данных для каждого вышеописанного метода. Для решения проблемы была предпринята попытка создания модели машинного обучения, которая сможет только на основе анамнеза пациента определить вероятность наличия или отсутствия у него кардиологического заболевания. Это позволит врачу уже на начальном этапе определить, стоит ли назначать дополнительные дорогостоящие

обследования или же есть смысл обратиться к специалисту другого профиля.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

В процессе исследования был использован набор данных (из базы данных Кливленда), который на сегодняшний день является единственным, используемым специалистами по машинному обучению и анализу данных [1]. Выборка данных включает в себя 920 записей с данными реальных пациентов, каждая запись включает 16 атрибутов (табл.1).

Таблица 1

Описание признаков используемой в работе выборки данных

Признак	Описание
age	возраст в годах
sex	пол
dataset	место сбора данных
cp	тип боли в груди: без симптомов, отсутствие стенокардии, нетипичная стенокардия, типичная стенокардия
trestbps	кровяное давление в состоянии покоя в мм. рт. ст. при поступлении в больницу
chol	уровень сывороточного (общего) холестерина в мг/дл
fbcs	уровень сахара в крови натощак > 120 мг/дл
restecg	результаты электрокардиографии в состоянии покоя: в норме, неспецифические изменения комплекса ST-T, гипертрофия левого желудочка
thalch	достигнутая максимальная частота сердечных сокращений
exang	стенокардия, вызванная физической нагрузкой
oldpeak	депрессия ST, вызванная физической нагрузкой по сравнению с отдыхом
slope	наклон пикового сегмента ST: ровный (плоский), наклонный, нисходящий склон
ca	количество крупных сосудов, окрашенных при рентгеноскопии
thal	талассемия
num	целевой признак: отсутствие сердечных заболеваний, стадии сердечных заболеваний

На первом этапе исследования была проведена первичная обработка данных: удаление дубликатов и выбросов, а также корректирование нереалистичных значений. В связи с особенностями предметной области выбросами считались данные, находящиеся на расстоянии $1.7 \times IQR$, где IQR – межквартильный интервал.

Для заполнения пропусков использовался многомерный метод, основанный на группировке коррелирующих между собой признаков и выведении внутригруппового значения. Стандартизация признаков была необходима для того, чтобы числовые данные были правильно интерпретированы моделью машинного обучения. В первом случае для кодирования категориальных переменных использовался метод One-Hot Encoder (базируется на создании из одного категориального признака множества бинарных, которые показывают принадлежность к уникальному значению), а во втором – метод Target Encoder (использует целевую переменную для кодирования категориальных признаков, т. е. категориальные значения признаков заменяются сочетанием апостериорной вероятности целевой переменной с учетом конкретного категориального значения и априорной вероятности целевой переменной по всему обучающему набору).

На этапе обучения моделей мы использовали метод главных компонент PCA [2] (взяв соответственно 3, 5, и 7 вычисленных главных компонент), однако он не смог существенно улучшить результаты обучения. Подбор гиперпараметров моделей осуществлялся методом поиска по сетке (рассмотрение определенных комбинаций гиперпараметров и выбор той, которая обеспечивает наилучшие результаты модели), который реализует функция GridSearchCV; некоторые модели обучались с гиперпараметрами по умолчанию.

В ходе исследования было обучено в общей сложности 20 моделей: 7 моделей классификатора на основе дерева принятия решений, 7 – на основе случайного леса, 6 – на основе логистической регрессии.

Для предварительной обработки данных, построения и обучения моделей машинного обучения использовался язык программирования Python и ряд специализированных, написанных на этом языке библиотек.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Наиболее качественная построенная нами модель машинного обучения на основе логистической регрессии имеет следующие оценки точности (табл. 2).

Таблица 2

Результаты модели классификатора на основе логистической регрессии

	<i>precision</i>	<i>recall</i>	<i>f-1 score</i>	<i>support</i>
Здоровый	0.84	0.79	0.81	125
Больной	0.83	0.87	0.85	142

Столбец *precision* показывает долю объектов, названных классификатором положительными и при этом действительно

являющихся положительными, т. е. $\text{precision} = \text{True Positives} / (\text{True Positives} + \text{False Positives})$.

Столбец *recall* в свою очередь показывает, какой процент объектов положительного класса из всех объектов положительного класса нашел алгоритм, т. е. $\text{recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives})$.

Значения в столбце *F1-score* – это способ объединения *precision* и *recall* через вычисление среднего гармонического этих показателей, т. е. $\text{f1-score} = 2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$.

Последний столбец *support* показывает количество экземпляров каждого класса.

ЗАКЛЮЧЕНИЕ

Результаты исследования показали, что создаваемые уже сегодня решения на основе искусственного интеллекта способны, возможно, с некоторыми ограничениями, решить с высокой эффективностью стоящие перед специалистами-медиками задачи о раннем обнаружении сердечно-сосудистых заболеваний и, таким образом, существенно повлиять на здравоохранение во всём мире.

Кроме того, были обнаружены некоторые особенности при использовании различных моделей машинного обучения в работе с данными, взятыми для исследования из медицинской сферы.

Библиографические ссылки

1. UCI Heart Disease Data | Kaggle [Электронный ресурс]. – Режим доступа: <https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data?datasetId=888463> Дата доступа: 05.07.2023.
2. Померанцев А. Л. Метод главных компонент [Электронный ресурс]. – Режим доступа: <https://rsc.chemometrics.ru/ru/books/metod-glavnykh-komponent/> – Дата доступа: 05.07.2023.
3. Учебник по машинному обучению [Электронный ресурс]. – Режим доступа: <https://academy.yandex.ru/handbook/ml> – Дата доступа: 05.07.2023.
4. Диагностика ишемической болезни сердца интеллектуальной системой «АРМ-Кардиолог» / О. А. Ефремова [и др.] // Курский научно-практический вестник «Человек и его здоровье». 2014. №. 1.