УДК 004.9

Хлопин И. В.

Предсказывание наличия болезни сердца у человека с помощью машинного обучения

Рекомендовано к публикации доцентом Гришкиным В. М.

1. Введение. В большинстве развитых стран частой причиной утраты здоровья являются сердечно-сосудистые заболевания, поэтому важно найти способ преждевременного прогнозирования этих болезней. Задача решается разными способами, один из самых эффективных из них – это метод машинного обучения, который может облегчить жизнь врачам, взяв на себя часть работы по обнаружению болезней сердца на основе имеющихся данных о пациенте. Например, можно сразу составить группу риска из пациентов, подозрительных на наличие сердечно-сосудистых заболеваний.

Довольно часто задачи в медицине решаются именно машинным обучением, например, при работе со снимками данный метод особенно эффективен [1].

2. Постановка задачи. Прогнозирование болезни сердца – это задача бинарной классификации. Для решения этой задачи имеется набор данных, который разделяет пациентов на два класса: имеется болезнь сердца – 1, и не имеется – 0. Подобным образом можно решать любые задачи диагностики заболеваний, либо определять наиболее целесообразный способ лечения.

Набором данных выступает “Heart Failure Prediction Dataset” [2]. Основная проблема – это маленький объем данных. В наборе примерно 900 объектов, а если исключить объекты с незаполненными признаками, останется 746. Маленький набор данных сказывается на точности решения задачи. Но чтобы улучшить положение, можно нормализовать численные параметры в наборе данных, а признаки с ограниченными вариантами значения разделить на отдельные признаки.

Хлопин Игорь Владимирович – студент, Санкт-Петербургский государственный университет; e-mail: khlopin2018@yandex.ru, тел.: +7(900)543-81-48

3. Структура данных. Каждый объект представлен парой $< x, y >$ (набор признаков, ответ). Вектор описания объекта x состоит из 11 признаков:

- Age: возраст пациента [лет];
- Sex: пол пациента [M: Мужчина, F: Женщина];
- ChestPainType: тип боли в груди [TA: Типичная стенокардия, ATA: Атипичная стенокардия, NAP: Неангинальная боль, ASY: Бессимптомный];
- Cholesterol: сывороточный холестерин [мм/дл];
- FastingBS: уровень сахара в крови натощак [1: если BS натощак $>$ 120 мг/дл, 0: в противном случае];
- RestingECG: результаты электрокардиограммы в состоянии покоя [Normal: Норма, ST: наличие аномалии зубца ST-T (инверсии зубца T и/или повышение или понижение ST $>$ 0,05 мВ), LVH: наличие вероятной или определенной гипертрофии левого желудочка];
- MaxHR: достигнутая максимальная частота сердечных сокращений [Числовое значение от 60 до 202];
- ExerciseAngina: стенокардия, вызванная физической нагрузкой [Y: Да, N: Нет];
- Oldpeak: oldpeak = ST [Числовое значение, измеренное в депрессии];
- STSlope: наклон сегмента ST пикового упражнения [Up, Flat, Down];
- RestingBP: артериальное давление в состоянии покоя [мм рт.ст.].

HeartDisease – это значение y (1 – есть болезнь сердца, 0 – норма). В данном наборе данных каждый объект описывают 11 признаков, которому соответствует значение (1 либо 0), показывающее наличие заболевания сердца. Следовательно, данную задачу можно отнести к задачам классификации.

3.1. Преобразование данных. Для удобства работы с данными: 1) признаки, принимающие произвольные значение, были отображены на отрезок [0, 1], где 0 – минимальное значение признака, 1 – максимальное значение признака; 2) признаки, принимающие значение из заранее известного небольшого множества вариантов,

были разделены на отдельные признаки (каждый вариант значения данного признака вынесен как отдельный признак).

4. Архитектура нейронной сети. Так как данные о пациенте представляют вектор признаков, размер которого достаточно мал, то целесообразно использовать архитектуру полносвязной нейронной сети [3]. Путём экспериментального подбора выбрана оптимальная архитектура полносвязной нейронной сети, состоящая из входного вектора, двух скрытых слоёв (по 5 нейронов каждый) и выходного вектора размерности два. Выбранная нейронная сеть реализована в среде Colaboratory [5], язык программирования – Python.

Функцией ошибки выбрана бинарная кросс-энтропия, которая хорошо подходит для задач классификации.

Метрика показывает, насколько качественно обучена сеть. Для решения задачи в качестве метрики взято отношение правильных предсказаний к общему количеству предсказаний (правильные/всего). Оптимизатором выбран Adam [4] с подобранным шагом обучения, равным 0,0001.

5. Подготовка данных перед обучением. Имеющиеся 746 объектов разделим на три группы (train – тренировочный набор данных, valid – валидационный набор данных, test – тестовый набор данных).

Путём корреляционного анализа выявлено, что существеннее всего коррелируются с наличием болезни сердца 7 признаков: Chest Pain Type ATA, Chest Pain Type ASY, MaxHR, Exercise Angine, Oldpeak, StSlopeUp, StSlopeFlat.

Обучив сеть на двух наборах данных (19 исходных признаков и 7 признаков с наибольшей корреляцией с болезнью сердца), можно сравнить их эффективность.

6. Обучение. Обучение сети происходит на тренировочном наборе данных train, оценка качества во время самого процесса обучения – на valid, а итоговый вывод о результате работы обученной сети делается на тестовом наборе данных test.

На каждой эпохе подаётся на вход набор данных (батч) размером с сам тренировочный набор данных (600 объектов), после вычисляется функция потерь на этом наборе, и с помощью оптимизатора изменяются весы нейронной сети.

Каждую тысячную эпоху фиксируются значения функции потерь

и метрики на тренировочном и валидационном наборах данных.

Обучение проводилось на наборе данных из всех имеющихся 19 признаков и на наборе данных, состоящем из признаков, наиболее существенно коррелирующих с наличием болезни сердца.

6.1. Результат обученния сети на 19 признаках. Для обучения сети на 19 признаках были выбраны следующие параметры обучения: количество эпох – 10 тысяч, шаг обучения оптимизатора Adam – 0,0001, размер батча – 600.

На каждой тысячной эпохе фиксировались значения функций потерь и метрик. На рис. 1. представлен график этих значений.

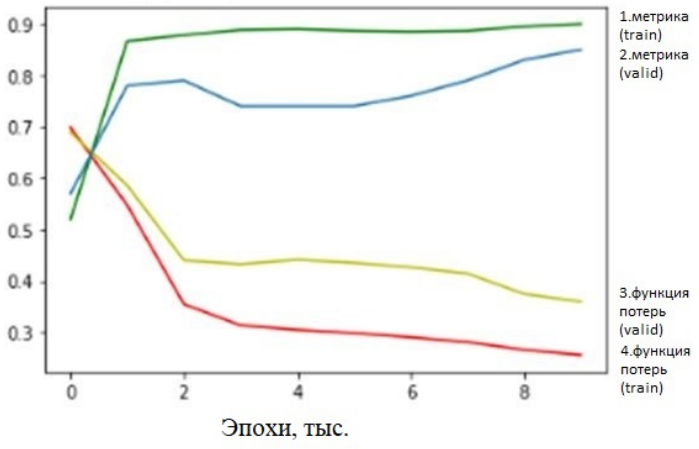

Рис. 1. График обучения на полном наборе признаков (10 тысяч эпох)

Метрика на тестовом датасете равняется 0,8. Значения обеих метрик на валидационном и тестовом датасетах примерно равняются 0,8.

Таким образом, нейронная сеть, обученная на полном наборе признаков, предсказывает наличие болезни сердца в 80% случаев.

6.2. Результат обученной сети на наборе из 7 признаков. Для сети с набором данных из 7 признаков выбраны следующиие параметры обучения: количествово эпох – 20 тысяч, шаг обучения оптимизатора Adam – 0,0001, размер батча – 600.

Точно так же на каждой тысячной эпохе фиксировались значения функций потерь и метрик. На рис. 2. представлен график этих значений.

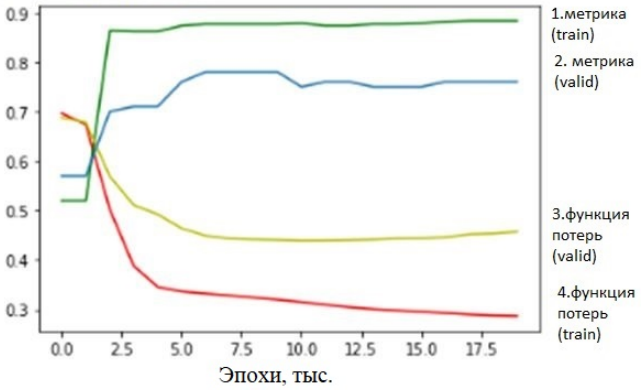

Рис. 2. График обучения на неполном наборе (20 тысяч эпох)

Вычисленная на тестовом датасете метрика равняется 0,71. Как и ожидалось, качество обучения уменьшилось с уменьшением количества признаков, на которых сеть обучалась.

Таким образом, нейронная сеть с 7 входными признаками (наиболее коррелирующими с наличием болезни сердца) верно предсказывает наличие болезни сердца в 71% случаев.

7. Вывод. Архитектура нейронной сети показала приемлемую работоспособность в задаче предсказывания наличия болезни сердца. При этом лучший результат показала первая нейронная сеть с 19 входными признаками.

В результате эксперимента можно сделать вывод, что полносвязная нейронная сеть с 19 входными признаками, с двумя скрытыми слоями (по 5 нейронов каждый), с бинарной кросс-энтропией в роли функции потерь и с оптимизатором Adam может предсказывать наличие болезни сердца в 80% случаев.

В случае, если не хватает данных о пациенте, но имеется набор данных из 7 признаков, наиболее существенно коррелирующих с наличием болезни сердца, то вторая нейронная сеть предскажет наличие болезни сердца в 71% случаев.

Предлагаемый подход по прогнозированию наличия болезни сердца может ускорить процесс диагностирования, повысить его качество и сэкономить время врачам.

Литература

1. Шиманчук Д., Кванчиани К., Блеканов И. Нейросетевые методы повышения качества медицинских изображений для обнаружения и сегментации опухолей [Электронный ресурс]: URL:http://hdl.handle.net/11701/32412 (дата обращения: 05.04.2022).

2. Kaggle. Обучающий набор данных [Электронный ресурс]: URL:https://www.kaggle.com/fedesoriano/heart-failure-prediction (дата обращения: 20.03.2022).

3. Popescu M., Balas V., Perescu-PopescuNikos L., Mastorakis N. Multilayer perceptron and neural networks [Электронный ресурс]: URL:https://www.researchgate.net/publication/228340819_Multilayer_perceptron_and_neural_networks (дата обращения: 05.04.2022).

4. Loshchilov I., Hutter F. Decoupled Weight Decay Regularization [Электронный ресурс]: URL:https://arxiv.org/abs/1711.05101 (дата обращения: 05.04.2022).

5. Github. Код проекта [Электронный ресурс]: URL:https://github.com/IgorvH-Developer/ColaboratoryFailes/blob/main/Heart_Failure_Prediction.ipynb (дата обращения: 20.03.2022).