Обнаружение признаков болезни сердца с использованием методов машинного обучения

Д.Ф. Пирова[1], Б.Э. Забержинский[1], А.В. Машков[1]

[1]Самарский государственный технический университет, Молодогвардейская 244, Самара, Россия, 443100

Аннотация. В данной работе исследуется возможность использования методов машинного обучения для обнаружения сердечно-сосудистых заболеваний сердца. Для анализа было взято 187 записей ЭКГ, из которых 80 принадлежат здоровым пациентам, 90 соответствуют пациентам, которые больны инфарктом миокарда и 17 пациентов с кардиомиопатией. Сигнал каждой записи был предварительно обработан. Результатом предварительной обработки явился общий сегмент, состоящий из 600 отсчётов. Для обнаружения признаков болезни сердца использовались такие методы, как: «Случайный лес», классическая логистическая регрессия, метод опорных векторов и нейронная сеть, состоящая из трёх слоёв.

1. Введение

Сердечно-сосудистые заболевания являются основной причиной утраты здоровья в большинстве развитых стран, поэтому преждевременное обнаружение сердечно-сосудистых заболеваний является очень важной проблемой. Для того, чтобы обнаружить сердечные заболевания, врач анализирует результаты электрокардиографии.

Существует много различных методов для анализа ЭКГ-сигнала. Например, методы основанные на слепом разделении сигнала [1], многослойном методе опорных векторов [2] и другие.

Цель данного исследования – изучение применимости методов машинного обучения для обнаружения ЭКГ признаков, которые свидетельствуют о некоторых отклонениях сердца от нормы.

Для анализа электрокардиограммы была выбрана база данных PhysioNet [3], в которой представлено в общей сложности 594 записи ЭКГ. Для каждой записи был создан файл с личными данными пациента, включающими в себя такие данные, как возраст и диагноз, а также файл ЭКГ-записи в формате .dat. Для работы с имеющейся базой данных была использована библиотека Wfdb языка программирования Python., так как данные электрокардиограммы предоставляются в numpy-массивах, работа с которыми не предоставляет затруднений.

В представленной работе на первом этапе, исходя из полученных сигналов электрокардиограмм пациентов, был извлечен общий сердечный цикл, который позволил уменьшить общий объем данных и шаблонизировать все данные. Это было сделано с помощью библиотеки biosppy языка программирования Python. В-первую очередь, был использован метод Гамильтона, который позволил обнаружить R-пики. После нахождения данных зубцов,

сигнал был разделен в окрестности 600 отсчетов относительно каждого. Затем массивы, включающие в себя 600 отсчётов, были усреднены, в последвии чего был получен общий сердечный цикл для сигнала электрокардиограммы.

На втором этапе представлено решение задачи бинарной классификации с использованием различных методов машинного обучения.

2. Предварительная обработка сигнала

В один массив были объединены два набора данных больных пациентов с кардиомиопатией и инфарктом. Таким образом, было получено два массива, один из которых состоял из 80 записей здоровых пациентов, а другой из 107 записей больных пациентов. Используя библиотеку biosppy были найдены R-пики каждого сигнала и выделены окрестности из 600 точек около каждого R-пика. Полученные сердечные циклы, состоящие из 600 точек, были усреднены для получения общего сегмента для ЭКГ-сигнала. Данная обработка была проведена для всех ЭКГ-сигналов. На рисунке 1 показан набор сердечных циклов здоровых пациентов.

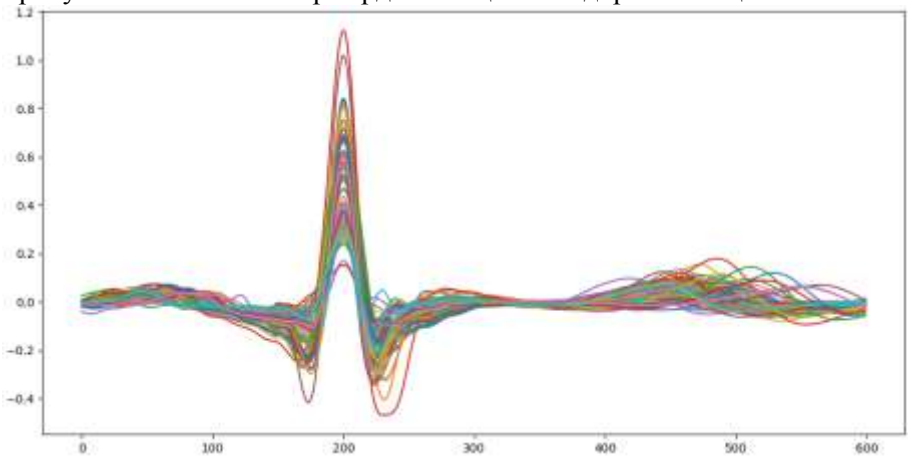

Рисунок 1. Набор сердечных циклов.

3. Построение и обучение моделей машинного обучения

Для обнаружения признаков болезни сердца были построены модели, основанные на таких методах, как: «Случайный лес», классическая логистическая регрессия, метод опорных векторов и нейронная сеть, состоящая из трёх слоёв (25, 5 и 1 нейрон). Опишем подробнее принцип работы каждого из методов:

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится зависимая переменная y, принимающая значения 0 и 1 и множество независимых переменных $x_1,...,x_n$ на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной [4]. Например, рассмотрим случай двух классов: $Y=\{-1,+1\}$. В логистической регрессии строится линейный алгоритм классификации $a: X \to Y$ вида

$$a(x,w)=\operatorname{sign}\left(\sum_{j=1}^{n} w_i f_j(x) - w_0\right)=\operatorname{sign}\langle x,w\rangle, \quad (1)$$

где w_j – вес j – го признака, w_0 – порог принятия решения, $w=(w_0,...,w_n)$ – вектор веса, $\langle x,w\rangle$ – скалярное произведение признакового описания объекта на вектор весов.

Основная идея метода опорных векторов заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Алгоритм работает в в следующем предположении: чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора. В данной работе в качестве ядра метода опорных векторов была взята линейная гиперплоскость.

Случайный лес состоит из множества деревьев решений. Они применяются в статистике, анализе данных и машинном обучении. На рисунке 2 наглядно продемонстрирована модель данного алгоритма. Каждое отдельное дерево — достаточно простая модель, которая имеет ветви, узлы и листья. В узлах записаны атрибуты, от значений которых зависит целевая функция. Далее по ветвям в листья попадают значения целевой функции. В процессе классификации нового случая нужно спуститься по дереву через ветви до листа, пройдя через все значения атрибутов по логическому принципу "ЕСЛИ-ТО". В зависимости от этих условий, целевой переменной будет присвоено то или иное значение или класс (целевая переменная попадет в конкретный лист). Цель построения дерева решений — создание модели, которая предсказывает значение целевой переменной в зависимости от нескольких переменных на входе [5].

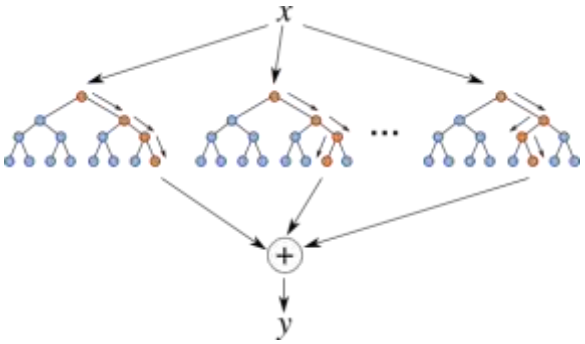

Рисунок 2. Алгоритм «Случайный лес».

Простейшая модель искусственной нейронной сети представлена на рисунке 3.

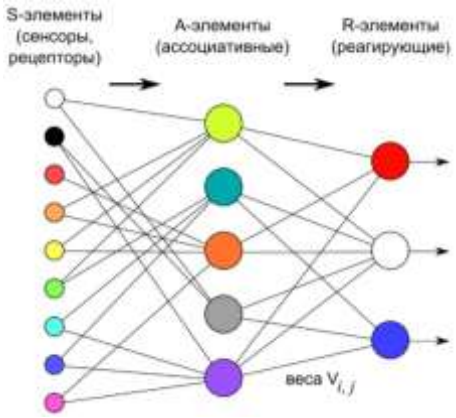

Рисунок 3. Простая модель искусственной нейронной сети.

В данной работе была построенная искусственная нейронная сеть, состоящая из 3 слоёв. В качестве функций активации на каждом слое была выбрана функция ReLU. Так как перед нами типичная задача бинарной классификации, то была выбрана функция ошибки logloss.

$$logloss = -(y\log(p) + (1-y)\log(1-p)), \quad (2)$$

где y – двоичный индикатор (0 или 1) того, является ли метод класса правильной классификацией для наблюдения, p – прогнозируемая вероятность модели, что наблюдение относится к классу.

Каждый из описанных выше методов имеет свои достоинства и недостатки. В данной работе мы обучили все описанные выше модели, для того, чтобы определить какая модель позволит точнее определить болезнь сердца.

4. Экспериментальные исследования

Для экспериментального исследования был произведён запуск программы для набора данных, состоящего из 187 записей для каждого из методов машинного обучения. Перед обучением

были предварительно обработаны все сигналы. Обучение моделей проходило при использовании кросс-валидации. Для каждой модели были посчитаны характеристики: чувствительность, специфичность и точность. В таблице 1 показаны результаты обучения.

Таблица 1. Результаты нейронной сети в зависимости от взятых отведений.

Метод	Чувствительность	Специфичность	Точность
Случайный лес	0.879	0.775	0.874
Метод опорных векторов	0.879	0.675	0.791
Логистическая регрессия	0.888	0.7125	0.813
Нейронная сеть	0.832	0.7	0.775

5. Заключение

В данной работе было произведено сравнение различных алгоритмов машинного обучения для обнаружения сердечных заболеваний. Перед применением алгоритмов, все ЭКГ-сигналы были предварительно обработаны в соответствии с разделом 2. Исследование показало, что классическая логистическая регрессия лучше остальных алгоритмов показывает вероятность того, что больной субъект будет классифицирован именно как больной (чувствительность). Однако, лучшую специфичность и точность продемонстрировал алгоритм "Случайный лес". Наименее эффективные результаты были продемонстрированы нейронными сетями, что указывает на необходимость использования классических алгоритмов, так как опытным путем установлено, что не всегда нейронные сети дают лучшие результаты.

6. Литература

[1] Devika, M.G. Myocardial infarction fetection using hybrid BSS method / M.G. Devika, R.P. Aneesh // International Conference on Communication Systems and Networks. – 2016. – Vol. 1. – P. 167.

[2] Dhawan, A. Detection of acute myocardial infarction from serial ecg using multilayer support vector machine / A. Dhawan, B. Wenzel, S. George, I. Gussak, B. Bojovic, D. Panescu // 34th Annual International Conference of the IEEE EMBS. – 2012. – Vol. 1. – P. 2704.

[3] The PTB Diagnostic ECG Database // PhysioNet, 2016 – [Electronic resource]. – Access mode: https://physionet.org/physiobank/database/ptbdb/ (18.11.2019).

[4] Davydov, N.S. Myocardial infarction detection using wavelet analysis of ECG signal / N.S. Davydov, A.G. Khramov // CEUR Workshop Proceedings. – 2018. – Vol. 2212. – P. 31-37.

[5] Флах, П. Машинное обучение. Наука и искусство построения алгоритмов. Гарнитура, 2016. – С. 38-56.

Detecting Heart Disease Symptoms Using Machine Learning

D. Pirova[1], B. Zaberzhinskiy[1], A. Mashkov[1]

[1]Samara State Technical University, Molodogvardeyskaya str. 244, Samara, Russia, 443100

Abstract. This paper explores the possibility of using machine learning methods for detecting cardiovascular diseases. 187 electrocardiography (ECG) recordings were taken for analysis, of which 80 records are the results of healthy people, 90 ones correspond to patients with myocardial infarction and 17 ones – to patients with cardiomyopathy. The signal of each recording has been pre-processed. The pre-processing resulted in a common segment consisting of 600 samples. Such methods as the random forest algorithm, classical logistic regression, the support vector method and a neural network consisting of three layers were used for detecting heart disease symptoms.