УДК 004.94 https://doi.org/10.51301/vest.su.2021.v143.i1.11

А. Кабдуллин
Институт кибернетики и информационных технологий, Satbayev University,
Алматы, Казахстан

ПОСТРОЕНИЕ НЕЙРОННОЙ СЕТИ С ПОМОЩЬЮ КОРРЕЛЯЦИОННОГО АНАЛИЗА ПРИЗНАКОВ ДЛЯ ПРЕДСКАЗАНИЯ РАННЕГО РИСКА РАЗВИТИЯ ИШЕМИЧЕСКОЙ БОЛЕЗНИ СЕРДЦА

Аннотация. В результате обзора статей, посвященных прогнозированию данных недугов были выявлены недостатки при диагностики ранней стадии. Работники здравоохранения диагностируют ишемическую болезнь сердца полагаясь на значения электрокардиограммы, анализа крови и прочих, но нельзя отметить и человеческий фактор и как показывает практика существует огромный риск не правильного диагноза пациентов на ранней стадии. По данным Всемирной организации здравоохранения, «Сердечно-сосудистые заболевания (ССЗ) являются основной причиной смерти во всем мире – ежегодно от ССЗ умирает больше людей, чем от какой-либо другой болезни. Существует огромное количество методов принятия решений ранней диагностики ишемической болезни сердца (ИБС), включая технологии машинного обучения. Данная статья посвящена изучению работ нейронных сетей с помощью корреляционного анализа признаков для предсказания риска развития ишемической болезни сердца.

Ключевые слова: нейронная сеть, система прогнозирования, машинное обучение, оптимизация нейронных сетей, медицина, ишемическая болезнь сердца, Python, Keras.

Введение. В 2030 г. от ССЗ, в основном от болезней сердца и инсульта, умрет около 23,6 млн человек. По прогнозам, эти болезни останутся основными отдельными причинами смерти» [1]. Чаще всего медицинские специалисты ставят диагнозы на основе результатов электрокардиографии, ангиографии и анализа крови. На раннем этапе болезнь диагностируется с трудом [2,3], однако для эффективного лечения очень важна ранняя диагностика. Диагнозы ставятся на основе личного опыта и квалификации медицинских работников, что приводит к увеличению рисков допущения ошибок, задержки необходимого лечения, следовательно, и времени лечения, таким образом, наблюдается существенное увеличение затрат на лечение пациента. Чтобы избавиться от этих недостатков, было проведено множество исследований в области клинических систем поддержки принятия решений с использованием таких технологий, как Data mining и машинное обучение.

Существует много исследований для прогнозирования ишемической болезни сердца с использованием нейронных систем или машинного обучения. За последние несколько лет было проведено огромное количество исследований в области раннего прогнозирования ишемической болезни сердца. В 2017 году нейронная сеть Арабасади [2], основанная на генетическом алгоритме, также Нарайн [3] использовала квантовую нейронную сеть. Но, как показывает практика, большинство предсказанных систем, использующих нейронные системы, не дают значительных результатов.

Методы. Программные средства, созданные на основе нейронных сетей, дали значимые итоги во время клинических испытаний. Несмотря на это, специалисты в области медицины остаются недовольными свойством «черного ящика» [4], которое нередко встречается у технологий данной категории. В результате обучение прогностических моделей осуществляется без понимания корреляции входных и выходных признаков нейросети. Для создания медицинских прогнозов, нейронная сеть должна выражать логику собственного функционирования, а это является серьезным недостатком**.**

Работа посвящается разработке прогностической системы, которая позволит сделать прогнозы о риске развития ИБС, основываясь на показателях нейросети и взаимосвязанного изучения признаков.

В процессе обучения, нейросеть определяет сложную корреляцию между входной и выходной информацией, а также обобщить полученные данные. Эффективный процесс обучения нейросетей способствует созданию информативных результатов для получения важных прогностических данных.

Систему, создающую прогнозы о риске возникновения ишемического заболевания сердца, решили создать, основываясь на многослойной нейросети прямой распространенности, так как эта архитектура отличается простотой и удобством для создания медицинских прогнозов.

Входные данные могут быть количественными и категориальными; под этим термином имеют в виду историю болезни. Система будет устроена так, чтобы данные о категориях передавались отдельно и обрабатывались специальным скрытым слоем. Это необходимо для наиболее удобной конфигурации данных для обработки.

Главным правилом анализа категориальной информации в нейросетях заключается в возможности охарактеризовать каждую из категорий собственным сигналом. Для этого следует применить векторные представления, которые помогут сопоставить вектора из действительных чисел к категориям. Наиболее широко применяется метод прямого кодирования. В нем каждый из существующих столбцов представляет одно вероятное значение признака.

В ситуации с множеством классов классификации количество нейронов в слое выхода равняется количеству необходимых классов. Их функция активации подразумевает Softmax или нормированную экспоненциальную функцию.

Данная функция обобщает логистические функции, сжимающие K-мерный вектор z в K-мерный вектор (z) из вещественных чисел, находящихся в диапазоне 0-1, а их сумма равняется 1.

Каждый пример для обучения способен демонстрировать значение, которое моделирует актуальное распределение вероятностей. Для сравнения двух имеющихся вероятностей следует иметь определенную меру. В роли такой меры можно использовать бинарную кросс-энтропию. Она подразумевает перемену квадратичной функции на кросс-энтропию. Это позволяет справиться с насыщением нейронов и уменьшить продолжительность обучения нейросети [5]. Данное явление можно определить следующей формулой:

x – пример для обучения;

t – желаемые итоги;

o – полученные итоги.

После ознакомления с формулой, можно сделать вывод, что вложение в функцию стоимости является низким, но только в случае, если фактический вывод приближен к желаемому итогу. Во время использования данной функции кривая становится более крутой, нежели исходная плоскость на аналогичной кривой квадритичной функции стоимости. Данная крутизна помогает ускорить процесс обучения, поскольку это является характерной чертой квадритичной функции стоимости.

Этот метод имеет один существенный недостаток: большие потери физической памяти, приводящие к увеличению длительности обучения нейросети и даже ухудшению конечных данных. Еще один недостаток прямого способа кодирования – это потеря данных в случае, если важно точное расположение существующих категорий.

Настоящая работа подразумевает наличие специализированных слоев, работающих по методу поисковых таблиц. Каждый из категориальных признаков имеет тензор указанных размеров, заполняемый векторами из случайных чисел.

После обучения отображающего слоя тензор сохраняет вектора, которые ближе всего характеризуют категориальную информацию для решения определенной задачи.

Следовательно, чтобы образец функционировал и анализировал информацию о категориях физическая память может использоваться в границах, фиксированных потребителем. Векторные представления информации будут представлять собой оптимальный способ решить задачу.

Использование функции «выпрямитель» показало лучшие результаты в процессе активации функции нейронов скрытого слоя. Функция f(x)=max(0,x) позволяет реализовать простой пороговый переход в нуле. Нейроны с данной функции именуются ReLU. Главными преимуществами нейронов являются:

- ReLU создается методом порогового изменения матрицы активации в нуле, в то время как сигмоидальные и тангенциальные функции требуют осуществление операций, требовательных к емкости ресурса;
- Присутствие линейного характера и отсутствия насыщения функции, приводящее к ускорению сходимости стохастического градиентного спуска.

Такая нейронная сеть будет обучаться по данным, имеющим около двух классов результирующих признаков. Эти классы выглядят как варианты «болен» и «не болен». Вопреки этому, возможен анализ информации с большим количеством результирующих категорий.

В ситуации с бинарной классификацией, выходной слой нейросети включает в себя один нейрон с сигмоидальной функцией активности. Указанная функция обеспечивает обретение выходного значения в рамках от 0 до 1. Благодаря дифференцируемости можно применить технику обратной ошибки.

В некоторых случаях в роли функции стоимости нейронной сети используют категорийную кросс-энтропию, а в случае бинарного способа классификации применяют бинарную кросс-энтропию, являющуюся специализированной модификацией функций кросс-энтропии. Формула представлена ниже.

t – желаемый результат;

o – полученный результат.

После загрузки в систему набора данных для анализа, прежде всего следует заняться отсеиванием признаков, не влияющий на результирующий признак. Эта задача решается при помощи статистических способов, таких как U-критерий Манна-Уитни или критерий хи-квадрат Пирсона.

U-критерий Манна-Уитни представляет собой непараметрический критерий статистического вывода, который применяется для проверки разницы между представленными группами с помощью применения порядковых данных [7]. Небольшое значение критерия увеличивает вероятность достоверности различия между значениями.

В случае, если «xi-квадрат» обладает значением больше критического, актуален вывод о наличии взаимосвязи между определяемым фактором риска и исходом при соответствующей степени значимости.

Статистический анализ позволяет отбросить незначительные признаки, а дальнейшие процедуры необходимо совершать, основываясь на процессе обучения нейросети. Этот процесс возможен с понятием чувствительности Sen(X,xi), характеризующей вклад, который вносится признаком в результат функционирования модели. Этот параметр вычисляется в роли среднего значения перемены выхода сети с учетом добавления признаков исходной информации xi малого шума.

Для входа обученной нейросети используют набор данный, к которым добавляется шум, а затем происходит вычисления чувствительности. Позднее определения каждого признака происходит систематизация от меньшего к большему. Самый маленький вес имеют признаки с малой чувствительностью, а это означает, что их можно вычеркнуть. Нейросеть переобучается с упором на оставшиеся признаки, а затем проводится анализ производительности механизма и сравнение полученных данных со старой моделью. В случае, если производительность сохраняется, данный этап повторяют до получения минимального списка значимых признаков. Метод сбора значительных признаков приведен на следующем рисунке.

После проведения таких операций возникает нейросетевой механизм, который уже можно использовать. Модель прошла обучение на медицинских источниках информации и

лишается проблем «черного ящика». Это происходит благодаря структуре, способной учитывать все вариации взаимосвязи признаков.

Средство для разработок должны обладать высокой скоростью и результативностью. Для этого выбирают такие технологии:

- язык программирования Python;
- библиотека нейросети Keras [8];
- библиотека pandas для выбора и анализа данных [9].

Указанные средства отличаются удобством применения, высокими показателями функционирования, большими возможностями и производительностью.

Результаты. Анализируемая выборка представляет собой каталог информации об истории болезней пациентов клиники Кливленда [10]. База данных включает в себя 303 записи и 76 атрибутов, однако специалисты рекомендуют работать с 14 из атрибутов. В следующей таблице описывают характерные черты исследуемой выборки.

Таблица 1 - **Описание выборки из исходящих данных**

Designation	**Feature Name**	**Characteristic Type**
age	Age	Continuous
sex	Floor	Categorical
cp	Type of chest pain	Categorical
trestbps	Residual blood pressure	Continuous
chol	Cholesterol	Continuous
fbs	Fasting blood sugar	Categorical
restecg	ECG result at rest	Categorical
thalach	Maximum heart rate during thallium stress test	Continuous
exang	Induced angina pectoris	Categorical
oldpeak	ST segment depression caused by resting exercise	Continuous
slope	ST segment peak slope	Categorical
ca	The number of large vessels with fluoroscopy	Discrete
thal	Thallium stress test result	Categorical
num	Heart disease	Categorical

В роли обучающей выборки по большей части выступали записи, а другая часть представляла собой проверочную выборку для определения производительности.

Использование критериев Манна-Уитни и xi-квадрат демонстрирует незначительность признаков fbs и restecg для итога работы. В связи с этим целесообразно убрать значения указанных атрибутов из выборок для проверки и обучения.

Таким образов, первая модель нейросети состоит из 11 нейронов на входе, 4 – со скрытым слоем и с одним выходным. Модель, обученная на нормализации обучающей выборки, включает в себя 212 записей.

В итоге отбора признаков нейросети, наиболее значительными признаками называют sex, thal и cp, а менее важным признаком называют age. Сеть переобучается данными, не включенными в указанный признак. Это привело к упадку производительности модели 91% до 86%.

Характеристики взаимосвязи признаков определяются в соответствии с взаимодействием на изменение чувствительности. Оказалось, что признаки взаимосвязи способны влиять на изменения чувствительности друг друга при помощи увеличения характеристик единичного признака. Между собой коррелируют такие признаки, как chol, cp, exang, sex, slope, thal, thalach и trestbps.

Конечная модель обладает производительностью, которую можно сравнить с классификатором SVM, полносвязной нейросетью и Байесовским классификатором.

Классификация методом SVM преследует цель разработки алгоритмически действенных способов выстроения оптимальной гиперплоскости разделения в пространстве высоко размеренных признаков.

Байесовский классификатор основывается на использовании теоремы Байеса, предположения которой отличаются строгостью о независимости входящих данных.

Кроме точности, в роли оценивающих критерий использовали положительные и отрицательные значения предсказания. Положительное предсказывающее значение символизирует возможность заболевания во время положительных прогнозов, в то время как негативное предсказывающее значение определяет возможность отсутствие болезни в период отрицательного прогноза.

Таблица 2 - **Сопоставление методов предсказания**

PPV	NPV	Accuracy	
Proposed model	0.91	0.91	91.22%
SVM	0.85	0.86	85.67%
Naive Bayes Classifier	0.83	0.87	85.67%
Fully connected neural network	0.89	0.89	90.11%

Следующая таблица показывает, что SVM – это классификатор, который вместе с классификатором Байеса продемонстрировал менее эффективные результаты, чем нейросеть. Указанная модель выступает лучше нейросети, поскольку она способна ликвидировать неважные признаки во время обучения.

Чувствительный метод отбора признаков повествует о возможности развития ИБС. Следовательно, итоговая модель обладает высокими показателями производительности и отличается эффективностью в создании прогнозов на тему вероятности возникновения ИБС.

Заключение. Существует множество исследований о создании прогнозов на тему вероятности возникновения ИБС, и чаще всего в процессе исследований применяются нейросети и линейные регрессоры. В результате, можно сделать вывод о том, что для создания высокоэффективных нейросетей следует рассматривать особенности трактовки признаков модели и отбросить незначительные показатели.

Указанная модель нейросети с применением анализа корреляции признаков демонстрирует большую точность вероятности развития ИБС (92% точности) по сравнению с классификатором Байеса. Это означает, что нейронная сеть может быть полезной в процессе предварительной диагностики пациентов.

Способ создания медицинских прогнозов, разработанный в пределах данной работы, имеет большую значимость в области практического применения, поскольку внедрение данной системы поможет увеличить скорость и точность предварительной диагностики пациентов. Система обладает множеством преимуществ в сравнении с аналогами.

ЛИТЕРАТУРА

[1] WHO: Top-10 global causes of deaths, 2018 (https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death)

[2] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, A. A. Yarifard, “Computer Aided Decision Making for Heart Disease Detection Using Hybrid Neural network-Genetic Algorithm,” Computer Methods and Programs in Biomedicine, vol. 141, 2017, pp. 19 – 26.

[3] R. Naraini, S. Saxena, A.K. Goyal, “Cardiovascular Risk Prediction: A Comparative Study of Framingham and Quantum Neural Network Based Approach”, Patient Preference and Adherence, vol. 10, 2016, pp. 1259–1270.

[4] Sussillo, D., Barak, O. Opening the black box: lowdimensional dynamics in high-dimensional recurrent neural networks. Neural Computation, vol. 25, no. 3, 2013 - P. 626–649.

[5] MC.AI, “Softmax Function Beyond the Basics”, 2019, (https://mc.ai/softmax-function-beyond-the-basics/)

[6] Y. Bengio, I. Goodfellow, A. Courville, “Deep learning”, DMK Press, pp. 654-655, 2018.

[7] Kingma D.P., Ba J.L., “Adam: Method for Stochastic Optimization”, Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA. 7–9 May 2015.

[8] Official library guide of Keras (https://keras.io/)

[9] Official library guide of pandas (https://pandas.pydata.org/)

[10] Data set source for Heart Disease (https://archive.ics.uci.edu/ml/datasets/Heart+Disease)

[11] Q. Mao, F. Hu, Q. Hao, “Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey,” IEEE Communications Surveys & Tutorials, vol. 20 , pp. 2595 – 2621, June 2018.

[12] N.Y. Musaev, I.M. Belova, “System for predicting the possible infection of coronary heart disease using a neural network,” Innov: electronic scientific journal, 2018.

[13] Kim, J.K., Kang, S. Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis. Journal of Healthcare Engineering, vol. 2017, 2017 – 13p.

[14] R.S Michalski, J.G. Carbonell, T.M. Mitchell. Machine learning: An artificial intelligence approach. Springer Science & Business Media, 2013.

[15] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444

А. Кабдуллин

Кибернетика және ақпараттық технологиялар институты, Satbayev University,
Алматы, Қазақстан

ЖҮРЕК-ТҮРЕК АУРУЫНЫҢ ДАМУЫНЫҢ АЛҒАШҚЫ ТӘУЕКЕЛ БЕЛГІЛЕРІН БОЛЖАУ БОЙЫНША КОРРЕЛЯЦИЯЛЫҚ ТАЛДАУ КӨМЕГІМЕН НЕЙРОНДЫҚ ЖЕЛІНІ ҚҰРУ

Андатпа: Машиналық оқыту технологияларын қоса алғанда, жүректің ишемиялық ауруын (ЖСА) ерте диагностикалау үшін шешім қабылдау әдістері өте көп. Бұл мақала жүректің ишемиялық ауруының даму кауптын болжау үшін белгілердің корреляциялық талдауын қолдана отырып, жүйке желілерінің жұмысын зерттеуге арналған. Осы ауруларды болжауға арналған мақалаларды шолу нәтижесінде ерте сатыдағы диагностикадағы кемшіліктер анықталды.

Негізгі сөздер: биомедициналық бейнелеу, кардиология, машинамен оқыту.

A. Kabdullin

Institute of Cybernetics and Information Technology, Satbayev University, Almaty, Kazakhstan

NEURAL NETWORK USING FEATURE CORRELATION ANALYSIS TO PREDICT EARLY RISK OF CORONARY HEART DISEASE

Abstract. As a result of a review of articles devoted to predicting these ailments, shortcomings in the diagnosis of an early stage were identified. Health care workers diagnose coronary heart disease relying on the values of the electrocardiogram, blood test and others, but the human factor cannot be noted, and as practice shows, there is a huge risk of incorrect diagnosis of patients at an early stage. According to the World Health Organization, “Coronary disease” (CVD) is the leading cause of death worldwide - more people die from CVD every year than from any other disease.

There are a huge number of decision-making methods for early diagnosis of coronary heart disease (CHD), including machine learning technologies. This article is devoted to the study of the work of neural networks using correlation analysis of signs to predict the risk of developing coronary heart disease.

Keywords: Coronary Heart Disease, Neural Network, prediction system, machine learning, optimization of neural networks, medicine, Python, Keras.