Использование методов машинного обучения для оценки тяжести течения COVID-19

А.В. Кузнецова
ИБХФ им.Н.М.Эмануэля РАН, ФБУН «Центральный НИИ Эпидемиологии» Роспотребнадзора
Москва, Россия
azforus@yandex.ru

О.В. Сенько
ФИЦ «Информатика и управление» РАН, ФБУН «Центральный НИИ Эпидемиологии» Роспотребнадзора
Москва, Россия
senkoov@mail.ru

Е.М. Воронин
ФБУН «Центральный НИИ Эпидемиологии» Роспотребнадзора
Москва, Россия
emvoronin@yandex.ru

И.А. Демина
ГБУЗ ГКБ им. С. П. Боткина ДЗМ
Москва, Россия
doctor.demira@gmail.com

А.А. Плоскирева
ФБУН «Центральный НИИ Эпидемиологии» Роспотребнадзора
Москва, Россия
antoninna@mail.ru

О.А. Кравцова
ФБУН «Центральный НИИ Эпидемиологии» Роспотребнадзора
Москва, Россия
dbrl.olga@gmail.com

***Аннотация* — В работе представлены результаты исследования возможности оценивания тяжести течения COVID-19 по набору клинических показателей с использованием методов машинного обучения и интеллектуального анализа данных.**

Ключевые слова -- COVID-19, SARS-CoV-2, пневмония, компьютерная томография, машинное обучение, интеллектуальный анализ данных

1. Введение

Одним из наиболее тяжелых осложнений при заболевании COVID-19, вызываемого вирусом SARS-CoV-2, является пневмония различной степени тяжести, которая может привести к неблагоприятным исходам. В рамках существующих Временных клинических рекомендаций «Профилактика, диагностика и лечение новой коронавирусной инфекции COVID-19» (версия 16 от 18.08.2022) наличие и степень тяжести пневмонии подтверждается таким объективным диагностическим исследованием как компьютерная томография (КТ). Особого внимания требуют пациенты со среднетяжелыми формами пневмонии, при которых степень поражения легких составляет от 25 до 100% (КТ2 – КТ4).

Наша работа посвящена выявлению факторов риска неблагоприятного течения COVID-19 и прогнозированию развития у госпитализированных пациентов пневмоний различной степени тяжести на основании анализа значений ряда результатов клинической лабораторной диагностики (КЛД) и клинических данных. Для ответа на этот вопрос традиционно используются параметрические и непараметрические статистические критерии, позволяющие оценить значимость различий в сравниваемых группах по каждому из показателей. Недостатком такого подхода часто является недостаточная оценка взаимозависимости результатов КЛД и клинических данных со степенью тяжести пневмонии, подтвержденной методом КТ. В настоящей работе для изучения взаимосвязи таких переменных используется метод оптимальных достоверных разбиений (ОДР). Научная и практическая значимость нашей работы заключается в оценке с какой точностью степень тяжести пневмонии, подтвержденная объективными данными КТ, может быть предсказана по результатам КЛД и клинических данных с использованием методов машинного обучения. Иными словами, степень тяжести пневмонии предсказывается по результатам КЛД и клинических данных с использованием алгоритма, настроенного (обученного) по данным КЛД и клинических показателей пациентов.

2. Данные

Значения 105 клинико-лабораторных показателей анализировались в двух группах пациентов COVID-19: группе из 113 пациентов без пневмонии или с легкой формой пневмонии (КТ0 - КТ1), а также в группе из 31 пациента со среднетяжелой формой пневмонии (КТ2 – КТ4).

3. Методы анализа данных

При анализе данных использовался стандартный непараметрический критерий Манна-Уиттни. Использовалась также метод оптимальных достоверных разбиений, представляющий собой технологию анализа данных, основанную на построении оптимальных разбиений пространства объясняющих переменных. При этом разбиение считается оптимальным, если оно позволяет наилучшим образом разделить сравниваемые группы. Использовалась версия метода ОДР, в которой оптимальные разбиения ищутся в одном из двух множеств разбиений:

- семейство разбиений интервалов значений отдельных переменных на два интервала с помощью одной граничной точкой;

- семейство разбиений двумерной области совместных значений пар переменных на четыре подобласти с помощью границ, параллельных координатным осям.

Оптимальные разбиения ищутся по обучающей выборке $S=\{(y_1, \mathbf{x}_{1)}, \ldots, (y_{m,,} x_m)\}$, где y_i - значение бинарной прогнозируемой переменной $Y \in \{0,1\}$, $\mathbf{x}_j$ - вектор значений потенциальных объясняющих переменных $X_1, \ldots, X_n$. Внутри указанных семейств производится поиск разбиения, для которого достигает максимум функционала

$$Q = \sum_{i=1}^{k} (\hat{Y} - \hat{y}_i)^2 m_i \,, \qquad (1)$$

где $\hat{Y}$- среднее значение Y во всей выборке, $\hat{y}_i$ - среднее значение Y внутри подобласти (квадранта) разбиения с номером i, m_i - число объектов выборки S, для которых значения соответствующих переменных попадают в квадрант с номером i.

Оценка статистической значимости эмпирических закономерностей в методе ОДР производится с помощью перестановочного теста, заключающегося в многократном сравнении значении функционала Q (1) со значениями этого функционала на случайных выборках, полученных из исходной выборки путём многократных перестановок значений переменной Y относительно фиксированных позиций векторов переменных X_1, …, X_n. При оценке значимости двумерных закономерностей используется принцип бритвы Оккама в следующей форме. Более сложная модель должна быть использована только в том случае, если она опровергает нулевую гипотезу об исчерпывающем описании зависимости с помощью простой модели. Метод ОДР входит в пакет Data Master Azforus DMA (https://azforus.com/), который использовался для анализа данных. Пакет включает ряд известных методов автоматической классификации с учителем, в число которых входят логистическая регрессия, градиентный бустинг, метод опорных векторов, метод статистически взвешенных синдромов. Большое число анализируемых показателей делает необходимой коррекцию на множественное тестирование. Нами использовался в данном случае известный метод Бонферрони-Холма.

4. РЕЗУЛЬТАТЫ

Использование критерия Манна-Уиттни и одномерных моделей метода ОДР выявило существенные различия между сравниваемыми группами, описанными в разделе 2. В таблице I приведены семь показателей, различия по которым остаются значимыми даже после коррекции по Бонферрони-Холму. Данные показатели могут быть получены простым наблюдением за пациентами или с использованием общедоступных в настоящее время оксиметров. В их число вошли числовые показатели длительность пиретической лихорадки в днях (ПЛД), уровень сатурации крови О2 (Сат.О2), частота дыхания (ЧД). Также значимыми оказались бинарные показатели: наличие астенизации (Астен.), снижение веса (СВ), нарушение сна (НС), выпадение волос (ВВ). Границы, полученные по методу ОДР, указаны в столбце «Граница» для числовых показателей. Результаты коррекции значимости по U-критерию приведены в столбце МТ-БХ.

Результаты для двумерной модели представлены на Рис. 1. Квадратиками обозначены случаи, соответствующие отсутствию пневмонии или лёгкому течению пневмонии, звёздочками обозначены случаи, соответствующие тяжёлому течению пневмонии. Из рисунка видно, что при продолжительности пиретической лихорадки выше 3 дней и уровне сатурации ниже 98,5% абсолютно преобладают случаи с тяжёлым течением пневмонии (квадрант III).

Таблица I. СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ ПОКАЗАТЕЛЕЙ

Показатель	Граница	U-критерий	МТ-БХ
ПЛД	3,5	0,00002	0,0014
СВ	-	0,000001	0,00003
НС	-	0,000003	0,00023
ВВ	-	0,00001	0,0007
Сат, О2	98,5	0,00002	0,0014
Астен,	-	0,000031	0,0022
ЧД	16,5	0,00017	0,012

Из 9 случаев, удовлетворяющих указанным условиям, 8 соответствует тяжёлому течению пневмонии. Наоборот, в квадранте I, при уровне сатурации выше 98,5 и длительности пиретической лихорадки ниже 4 дней, соответствуют 71 случай без пневмонии или лёгким течением пневмонии и только 4 случая с тяжёлым течением пневмонии.

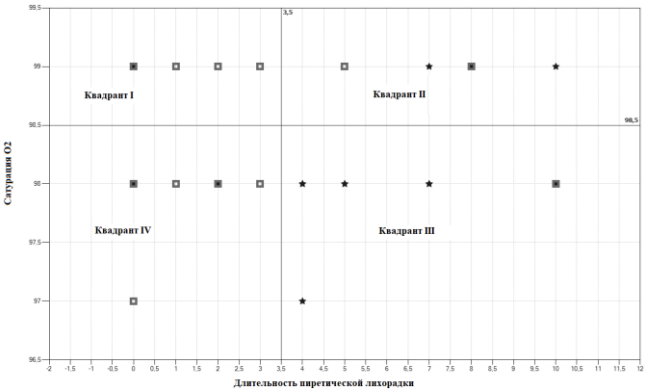

Рис. 1. Двумерное разбиение по показателям сатурации крови О2 и длительности пиретической лихорадки

Возможность диагностики тяжёлого течения пневмонии изучалась с использованием методов распознавания, вошедших в пакет Data Master. Наилучшие результаты с ROC AUC выше 0,8 были получены с использованием линейного дискриминанта Фишера, логистической регрессии и метода СВС. Коллективному решению по этим трём методам соответствует ROC AUC=0,845, чувствительность 0,82, специфичность 0,81, F-мера 0,88.

Литература

[1] Senko, O.V. Chapter 8 - Search of regularities in data: optimality, validity and interpretability / O.V. Senko, A.V. Kuznetsova, I.A. Matveev, I.S. Litvinchev – Advances of Artificial Intelligence in a Green Energy Environment,Academic Press, 2022. – 385 p.