DOI 10.31891/2307-5732-2022-311-4-237-240
УДК 004.8

СКОПІВСЬКИЙ С.Я.
Національний університет «Львівська політехніка»
https://orcid.org/0000-0000-0000-0000
e-mail: skopik.stepan@gmail.com

АНАЛІЗ МЕТОДІВ ПРОГНОЗУВАННЯ ІНФЕКЦІЙНИХ ЗАХВОРЮВАНЬ

В роботі наведено аналіз різноманітних підходів до прогнозування захворювання на COVID-19. Досліджено сучасні роботи вчених в області поширення інфекційних хвороб, застосування машинного навчання до статистичних даних про розповсюдження COVID, його взаємодії з іншими захворюваннями.

Ключові слова: COVID-19, прогнозування, машинне навчання.

Stepan SKOPIVSKY
Lviv Polytechnic National University

ANALYSIS OF INFECTIOUS DISEASES FORECASTING METHODS

At the end of 2019, COVID showed the world its unpreparedness and inability to resist the modification of the influenza virus. The onset of a global pandemic, the spread of the disease, and the large number of deaths have led not only to the search for control of the virus, but also to the possibility of predicting its spread. While some scientists have developed a vaccine, others have studied the prospects of the virus, filling the planet and predicting the number of deaths under certain conditions. Using statistical data, the researchers developed maps of the spread of the virus, possible future targets, and even estimated possible deaths from various strains of COVID-19.

The main task of data forecasting is to create some models from the provided data set in order to provide useful and correct forecasting of future or unknown values of another data set.

For many years, standard statistical methods and mainly the intuition of the doctor, his knowledge and experience have been used to predict the risk of the disease, the occurrence of complications in the patient, the spread of the disease among other people. This approach to disease assessment often leads to unwanted biases, errors and large losses. In the modern field of medicine, such an assessment also has a negative impact on the quality of services provided to patients. With the availability of electronic health data, more reliable and advanced computational approaches have emerged, such as machine learning and big data analysis. The emergence of new methods in data mining has led to the study and application of forecasting methods in the field of disease. In the literature, most relevant studies have used one or more machine learning algorithms to predict a particular disease. For this reason, comparing the effectiveness of different controlled machine learning algorithms for disease prediction is the main focus of this study.

Keywords: COVID-19, forecasting, machine learning.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Аналіз даних та прогнозування є одним із найбільш поширених напрямів машинного навчання у сучасному світі. Аналізуючи дані можна дізнаватись дуже багато корисної інформації, яку в подальшому можливо використати для різних цілей, зокрема і спрогнозувати майбутню поведінку досліджуваного об'єкта. Експоненційне збільшення часу в даних зробило важким отримання корисної інформації з даних. Традиційні методи показали велику продуктивність; однак їх прогнозна потужність обмежена, оскільки традиційний аналіз стосується лише первинного аналізу, тоді як аналітика даних стосується вторинного аналізу. Видобуток даних – це «діставання» або видобуток даних з багатьох вимірів або перспектив за допомогою інструментів аналізу даних, щоб знайти попередню невідому закономірність і зв'язок в даних, які можуть бути використані як дійсна інформація; крім того, він використовує цю видобуту інформацію для побудови прогнозної моделі. Інтенсивно і широко використовується багатьма організаціями, особливо в галузі охорони здоров'я.

Великі дані та машинне навчання мають великий потенціал для постачальників медичних послуг систематично використовувати дані та аналітику, щоб виявити модель, яка раніше була невідома. Медичні дані – це одні із найбільш корисних, але в деякій мірі одній з найскладніших даних для аналізу. Аналітика даних з її обіцянкою ефективно виявляти цінну закономірність шляхом аналізу великої кількості неструктурованих, неоднорідних, нестандартних та неповних даних про охорону здоров'я. Машинне навчання не тільки прогнозує, але й допомагає у прийнятті рішень, і все частіше його помічають як прорив у постійному прогресі, метою якого є поліпшення якості догляду за пацієнтами та зниження вартості медичного обслуговування. Адже аналізуючи персональні та медичні дані людини є змога проаналізувати стан її здоров'я та спрогнозувати стан здоров'я в майбутньому.

Прогностична аналітика має силу перетворити галузь охорони здоров'я. Її переваги охоплюють якість та ефективність догляду за пацієнтами, а також ефективність та результативність медичного персоналу та організацій. Це також має позитивні наслідки для прибутковості лікарень та інших організацій охорони здоров'я.

Особливо доцільною є прогностична аналітика у сфері прогнозування того чи іншого захворювання людини. Досить часто здача аналізів для прогнозування стану пацієнта займає досить багато часу, зусиль та коштів, проте не завжди вистачає необхідних ресурсів і як результат це може призводити до летальних випадків. Саме тому машинне навчання є досить хорошим помічником у цьому напрямку і може допомогти зберегти велику кількість людських життів.

Аналіз досліджень та публікацій

Досліджуючи дану проблемну область було проаналізовано багато статей, аналогів. У роботі [1] було розроблено алгоритм машинного навчання для прогнозування майбутньої інтубації серед пацієнтів з діагнозом або підозрою на COVID-19. Це ретроспективне когортне дослідження пацієнтів, яким поставлено діагноз або, які перебувають під дослідженням на перенесення COVID-19. Алгоритм машинного навчання був навчений прогнозувати наявність інтубації в майбутньому на основі попередніх показників, лабораторних та демографічних показників. Ця стаття є хорошою для аналізу, оскільки розроблений у ній алгоритм може бути використаний для виявлення пацієнтів з високим ризиком можливості захворіти.

У роботі [2] описують параметри коагуляційної функції хворих на COVID-19 та виявлення факторів ризику розвитку важкої хвороби. У ній було проведено багатовимірний аналіз регресії Кокса для виявлення потенційних біомаркерів для прогнозування прогресування захворювання. За результатами цього аналізу, була побудована номограма та оцінена точність прогнозування за допомогою калібрувальної кривої, кривої прийняття рішення, кривої клінічного впливу та аналізу Kaplan–Meier.

У роботі [3] пропонується відрегульована модель випадкового лісу, підсилена алгоритмом AdaBoost. Модель використовує географічні, туристичні, медичні та демографічні характеристики пацієнта для прогнозування важкості перенесення та можливого результату, одужання або смерті. Модель має точність 94% та оцінку F1 0,86, на використаному наборі даних аналіз виявляє позитивну кореляцію між статтю пацієнтів та смертністю, а також вказує на те, що більшість пацієнтів мають вік від 20 до 70 років.

Робота [4] показує процес очищення даних, який являє собою очищення відсутніх значень, використовуючи інтерполяцію сплайнів та ентропію-кореляцію. Очищення даних потім піддається процесу вилучення ознак за допомогою Principle Component аналізу. Для вибору оптимальних функцій представлений алгоритм Dragon Fly, а результуючий вектор функцій подається до мережі Deep Belief.

Залежність між раковим та коронавірусним захворюваннями досліджується у роботі [5]. Це теоретична робота, в якій досліджується вплив певного захворювання в людини на можливість захворіти COVID-19.

У роботі [6] використано декілька нейронних моделей, а саме ResNet18, ResNet50, SqueezeNet та DenseNet-121, для ідентифікації COVID-19 на аналізованих рентгенівських знімках грудної клітки. Аналізуючи цю статтю, можна детальніше розібратись із даними моделями та можливо використати їх у подальших дослідженнях.

У праці [7] поставлено за мету виявлення взаємозв’язку між певними клінічними показниками (такими як лейкоцити, тромбоцити, вік, стать, наявність суміжних захворювань і т.д.) і тяжкістю коронавірусної хвороби COVID-19 та дослідження їхньої ролі у прогнозуванні тяжкості захворювання COVID-19. Пошук взаємозв’язку був здійснений за допомогою багатовимірної логістичної регресії. Тут здійснено аналіз взаємозалежності багатьох показників стану пацієнта та захворюваністю на COVID-19.

Описано проведення ретроспективного дослідження у 70 безсимптомно інфікованих пацієнтів, підтверджених тестами на нуклеїнові кислоти в провінції Хунань, Китай, з 28 січня 2020 року по 18 лютого 2020 року в роботі [8]. Для оцінки потенційних предикторів появи симптомів була наведена модель регресії Кокса. Як результат було виявлено те, що пацієнти, які курять чи мають легеневу хворобу, були досить схильними до безсимптомного перенесення хвороби, і необхідно бути пильним до цих пацієнтів.

У роботі [9] на основі ретроспективного аналізу було виявлено, що показник FIB підвищений у важких пацієнтів і був кращим, ніж кількість лімфоцитів та міоглобіну, для розрізнення загальних та важких пацієнтів. Дослідження також припустило, що гормональне лікування не має суттєвого впливу на COVID-19.

Було досліджено те, чи КТ точно виявляє важкість перенесення COVID-19 у роботі [10]. Зображення КТ витягували за допомогою LK2.1. Для виявлення суттєвих особливостей було використано двовибірковий t-тест або U-тест Манна-Уїтні. Метод мінімальної надмірності та максимальної релевантності (MRMR) був проведений, щоб знайти характеристики з максимальною кореляцією та мінімальною надмірністю. Потім ці особливості використовувались для побудови моделі текстури радіоміки для виявлення важких пацієнтів за допомогою багатовимірного методу логістичної регресії. Крім того, була побудована клінічна модель. Для оцінки ефективності двох моделей були проведені аналізи ROC. Співвідношення клінічних особливостей та особливостей текстури КТ аналізували за допомогою кореляційного аналізу Spearman.

Проаналізовані роботи досить цікаві та необхідні в даному баченні. Проте на даний час недостатньо досліджень, які б на основі певних показників могли передбачити з певною похибкою наявність COVID-19 у пацієнта.

Формулювання цілей статті

Метою роботи є дослідження підходів до машинного навчання прогнозування поширення інфекцій та визначення причин відхилень даних.

Виклад основного матеріалу

Традиційно для прогнозування ризику захворювання використовувались стандартні статистичні методи та інтуїція лікаря, знання та досвід. Така практика часто призводить до небажаних упереджень, помилок та великих витрат і негативно впливає на якість послуг, що надаються пацієнтам [11]. Зі

збільшенням доступності електронних даних про здоров'я, більш надійні та вдосконалені обчислювальні підходи, такі як машинне навчання, стали більш практичними для застосування та дослідження в області прогнозування захворювань. У літературі більшість відповідних досліджень використовували один або кілька алгоритмів машинного навчання для прогнозування певної хвороби. З цієї причини порівняння ефективності різних керованих алгоритмів машинного навчання для прогнозування захворювань є основним напрямком цього дослідження.

В навчанні з вчителем маємо деякі вхідні дані *x*, і прагнемо навчити певну функцію *y* відображати опрацьовані вихідні дані. Багатокласова класифікація – це завдання класифікації машинного навчання, яке складається з більш ніж двох класів або результатів. Існують сотні моделей для класифікації. Насправді часто можна взяти модель, яка працює для регресії, і перетворити її на модель класифікації. В основному так працює логістична регресія.

Оскільки моделі машинного навчання базуються на математичних рівняннях, можна інтуїтивно зрозуміти, що це може спричинити певні проблеми, якщо з'явиться змога зберігати категоріальні дані в рівняннях, оскільки потрібні лише числа в рівняннях. Тому категоріальні дані кодують у числові дані. Для перетворення таких даних використовують LabelEnoder. LabelEncoder – це об'єкт, який використовують і, який допомагає при передачі категоріальних даних у числові дані.

Після опрацювання даних класифікатор може вчитись прогнозувати результат на тренувальній вибірці даних.

Досягнувши достатньої точності на тренувальній вибірці, система може класифікувати раніше невідомі їй дані за допомогою натренованого класифікатора. Після тренування, класифікатор вже зможе вказати приналежність невідомого йому вектору ознак до одного із трьох очікуваних класів («хворий», «здоровий» та «можливо хворий» відповідно).

Decision tree – один із найпопулярніших алгоритмів машинного навчання. Це керований алгоритм машинного навчання, який використовується як для класифікації, так і для завдання регресії. Це модель, яка використовує набір правил для класифікації чогось.

Розгалуження у дереві базується на контрольних операторах або значеннях, а точки даних лежать по обидва боки вузла розбиття, залежно від значення конкретної ознаки. Структуру дерева рішень можна визначити за допомогою кореневого вузла, який є найважливішою функцією розділення.

Random Forrest – це ансамблевий метод навчання для класифікації, регресії та інших завдань, він діє шляхом побудови великої кількості дерев рішень під час етапу навчання, а також виведення класу, що є режимом класів (класифікація) або усередненого прогнозу (регресія) окремих дерев. Ліси випадкових рішень коригують звичку дерев рішень переоснащувати їх навчальний набір. Випадкові ліси переважно є кращими ніж дерева рішень, проте їх точність нижча, аніж підсилені градієнтами дерева. Однак характеристики даних можуть впливати на їх ефективність.

Простіше кажучи: випадковий ліс будує кілька дерев рішень та об'єднує їх, щоб отримати більш точний та стабільний прогноз.

Support vector machine (SVM) – це один з найпотужніших нестандартних керованих алгоритмів машинного навчання. На відміну від багатьох інших алгоритмів машинного навчання, таких як нейронні мережі, вам не потрібно робити багато налаштувань, щоб отримати хороші результати за допомогою SVM.

Ідея полягає в тому, щоб зіставити точки даних з простором, щоб отримати взаємне лінійне розділення між кожними двома класами. Це називається One-to-One підходом, який розщеплює в мультикласові проблеми в декількох бінарних завдання класифікації. Двійковий класифікатор для кожної пари класів.

Інший підхід, який можна використати, – це One-to-Rest. У цьому підході розбивка встановлюється на двійковий класифікатор для кожного класу.

Один SVM робить двійкову класифікацію і може розрізняти два класи так, щоб, згідно з двома підходами до розбиття, класифікувати точки даних із набору даних класів.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

За результатами проведеного аналізу можемо стверджувати, що на сьогодні не достатньо розкрите питання прогнозування штамів COVID, які б на основі певних показників могли передбачити з певною похибкою наявність COVID у пацієнта. Тому дана робота орієнтована на спростування проблеми прогнозування COVID, а також розкриття нових підходів щодо застосування традиційних методів для розв'язання проблеми прогнозування.

Література

1. Arvind V., Kim J. S., Cho B. H., Geng E., and Cho S. K. Development of a machine learning algorithm to predict intubation among hospitalized patients with COVID-19. Journal of Critical Care, vol. 62, pp. 25–30, 2021.

2. Bi X. et al. Prediction of severe illness due to COVID-19 based on an analysis of initial Fibrinogen to Albumin Ratio and Platelet count. Platelets, vol. 31, no. 5, pp. 674–679, 2020.

3. Iwendi C. et al. COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in Public Health, vol. 8, 2020.

4. Koppu S., Maddikunt P. K. R. a, and Srivastava G. Deep learning disease prediction model for use with intelligent robots. Computers and Electrical Engineering, vol. 87, 2020.

5. Li Q. et al. Cancer increases risk of in-hospital death from COVID-19 in persons <65 years and those not in complete remission. Leukemia, vol. 34, no. 9, pp. 2384–2391, 2020.

6. Minaee S., Kafieh R., Sonka M., Yazdani S., and Jamalipour Soufi G. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Medical Image Analysis, vol. 65, 2020.

7. Shang W. et al. The value of clinical parameters in predicting the severity of COVID-19. Journal of Medical Virology, vol. 92, no. 10, pp. 2188–2192, 2020.

8. Tao P.-Y. et al. Determination of risk factors for predicting the onset of symptoms in asymptomatic covid-19 infected patients. International Journal of Medical Sciences, vol. 17, no. 14, pp. 2187–2193, 2020.

9. Wang Y. et al. Clinical characteristics and laboratory indicator analysis of 67 COVID-19 pneumonia patients in Suzhou, China. BMC Infectious Diseases, vol. 20, no. 1, 2020.

10. Wei W., X.-W. Hu, Q. Cheng, Y.-M. Zhao, and Y.-Q. Ge. Identification of common and severe COVID-19: the value of CT texture analysis and correlation with clinical characteristics. European Radiology, vol. 30, no. 12, pp. 6788–6796, 2020.

11. Palaniappan S., Awang R. Intelligent heart disease prediction system using data mining techniques. In: Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on; 2008. p. 108–15. IEEE.

Рецензія/Peer review : 20.06.2022 р. Надрукована/Printed :02.08.2022 р.