УДК 519.67

к. ф.-м. н., доцент Третиник В. В., студентка Рижкова Д. О.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

КОНЦЕПТУАЛЬНА МОДЕЛЬ СИСТЕМИ ДІАГНОСТИКИ ХВОРОБ СИСТЕМИ КРОВООБІГУ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Abstract

Violeta Tretynyk, assoc. prof., PhD; Daria Ryzhkova student

Mathematics and software for diagnosing diseases of the circulatory system based on machine learning methods

This paper presents the evaluation of mathematical and software tools for diagnosis of circulatory system diseases using real medical data from MIMIC-III. Several classical machine learning methods are applied, including logistic regression, k-nearest neighbors (KNN), support vector machines(SVM), decision trees, and ensemble learning method (random forests, boosting). Their performance is compared to identify the most accurate approach for predicting the risk of cardiovascular diseases.

Вступ

Згідно з дослідженням ВООЗ проведеному у 2019 році загалом 60.1% населення мали 1-2 фактори ризику виникнення хвороб системи кровообігу, а одна третина населення (32,8%) – 3-5 факторів ризику, а 7.1% – не мали жодного; 23.4% населення з групи віком 40-69 років мають від 30% виникнення таких хвороб у наступні 10 років [1]. Відповідно до даних серцево-судинні хвороби є основною причиною смерті в Україні (2019 р.) [2].

Більшість систем та досліджень, що стосуються діагностики хвороб системи кровообігу зазвичай фокусуються на поширених хворобах серця (найпоширенішою для досліджень в сфері науки продані є ішемічна хвороба серця) і зазвичай використовують візуальні та сигнальні дані, що суттєво відрізняється від даної роботи, де використовуються дані з лабораторних досліджень. Ці дані охоплюватимуть декілька підгруп хвороб із групи хвороб системи кровообігу за ICD-9 (International Statistical Classification of Diseases, Міжнародна статистична класифікація хвороб, МКХ-9). Також в даній роботі буде виконано порівняння поширених методів машинного навчання для задач класифікації.

Постановка задачі

Метою даної роботи є розробка математичного та програмного забезпечення для задачі діагностики хвороб системи кровообігу на основі методів машинного навчання. Для визначення найкращого методу машинного навчання для поставленої задачі потрібно провести порівняльний аналіз існуючих методів. При проведенні класифікації кожному об'єкту (пацієнту) будуть надаватися мітки груп, до яких входить певний перелік хвороб. Кінцевим результатом роботи буде математичне, програмне та методичне забезпечення концептуальної моделі системи, що буде надавати ймовірність ризику виникнення певних хвороб системи кровообігу (їхній перелік визначатиметься за ICD-9 codes), пошук пацієнтів з бази даних лікарні, що входять у "зону ризику" для виникнення серцево-судинних хвороб.

Опис набору тренувальних даних

У даній роботі використовується набір даних MIMIC-III (Medical Information Mart for Intensive Care) [3]. Він містить дані пацієнтів, що перебували у відділенні інтенсивної терапії. Серед зазначених даних: демографічні дані, клінічні вимірювання, втручання, виставлення рахунків, словник медичної історії, фармакотерапія, клінічні лабораторні дослідження, медичні дані [4]. MIMIC-III - це деперсоналізовані медичні дані з американського Медичного центру Бет Ізраїль Діаконес (Beth Israel Deaconess Medical Center) у Бостоні, штат Массачусетс [4]. MIMIC-III складається з 26 таблиць і містить дані 53423 пацієнтів.

Для побудови моделей будуть використовуватися таблиці: PATIENTS, ADMISSIONS, CHARTEVENTS, LABEVENTS, DIAGNOSES_ICD, D_ITEMS, D_LABITEMS. Вони надають необхідну інформацію про діагнози, демографічні дані та результати лабораторних досліджень пацієнтів. Для побудови моделей було обрано записи 388321 госпіталізацій (одного пацієнта могли госпіталізувати декілька разів). Початкове кодування класів хвороб наступне (у дужках зазначені ICD коди хвороб): «Перикардіальні захворювання» (420, 423), «Ендокардіальні захворювання» (421, 424), «Міокардіальні ураження» (422, 425), «Серцеві аритмії» (427), «Серцева недостатність» (428), «Інші серцеві захворювання» (426, 429, 390–392), «Хронічна ревматична хвороба серця» (393–398), «Гіпертонічна хвороба» (401–405), «Ішемічна хвороба серця» (410–414), «Хвороби легеневого кровообігу» (415–417), «Цереброваскулярні захворювання» (430–438), «Захворювання артерій, артеріол та капілярів» (440–449), «Хвороби вен і лімфатичних судин» (451–459).

Перед тренуванням моделей виконується очищення та попередня

обробка даних з наступними кроками: Resampling, SMOTE, MICE, IQR та нормування на відрізок [0; 1].

Модель системи

Демонстраційний варіант системи складається з 7 модулів: модуль «Набір навчальних даних», модуль «Набір тестових даних», модуль «Попередня обробка даних», модуль «Навчання моделі», модуль «Класифікація», модуль «Інтерпретація», модуль «Інтерфейс користувача». Таке «розбиття» розроблювального програмного забезпечення дозволяє досягти гнучкості та вносити зміни в окремий компонент без впливу на інші, а також додавати нові для розширення функціоналу в майбутньому.

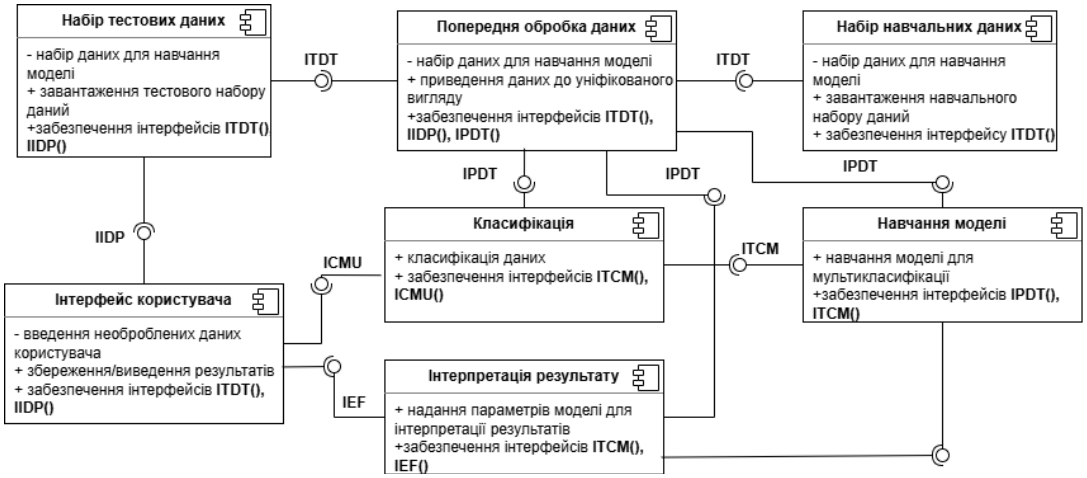

Рис. 1. Модель діагностичної системи. Діаграма компонентів у нотації UML

Серед алгоритмів машинного навчання будуть розглянуті: логістична регресія, k-найближчих сусідів (KNN), метод опорних векторів (SVM, Support Vector Machine), дерева рішень, ансамблеві методи (випадковий ліс, бустинг). Оскільки більшість методів сконструйовані для задач бінарної класифікації, то потрібно використовувати різні стратегії для розширення. Оскільки у цій роботі задача класифікації поставлена як багатоміткова варто використати стратегію «один проти всіх» (one-vs-rest strategy strategy, one-vs-all, OvA). Це дає додаткові можливості отримати не тільки загальну оцінку, а й побачити які класи розпізнаються гірше.

Логістична регресія – метод бінарної класифікації. Зокрема його доречно використовувати в медицині, оскільки він не просто класифікує об’єкт, а й показує ймовірність належності його до певного класу, але цей метод погано працює за великої кількості ознак та даних, які мають складні зв’язки.

KNN – приклад лінивого машинного навчання. Його ідея полягає не у створенні певної функції на основі навчальних даних, а запам’ятовування їхнього набору і в припущенні, що схожі об’єкти знаходяться близько один

до одного за певною відстанню. Цей алгоритм має високу адаптивність до даних, але є дуже ресурсно дорогим.

SVM – метод навчання з учителем, що розглядає кожен вектор ознак у багатовимірному просторі, його ціль побудувати гіперплощини, що розділяють різні класи. SVM досить гнучкий у налаштуванні, але великі набори даних робить його неефективним.

Дерево рішень – метод машинного навчання з учителем. Саме дерево рішень представляє собою ациклічний граф, який використовується для прийняття рішень. Цей метод може класифікувати будь-який об'єкт, але він чутливий до дисбалансу класів.

Випадковий ліс – метод ансамблевого навчання. Його ідея полягає в поєднанні декількох дерев рішень з незалежним вибором ознак. Такі моделі мають високу точність, але низьку інтерпретацію.

Бустинг – метод ансамблевого навчання, який ітеративно створює декілька слабких моделей, і кожна наступна виправляє помилки попередньої. Хоча бустинг зазвичай працює краще за випадковий ліс, його ресурсні витрати можуть зробити його неефективним.

Навчання моделей здійснюється за допомогою конвеєру. Це допомагає застосувати необхідні перетворення для даних (заповнення пропусків, нормалізація), підбір параметрів та, власне навчання моделі.

Оцінка алгоритмів

Створена система має на меті працювати з багатьма мітками одночасно, через що звичайні метрики (наприклад, правильність (accuracy)) не відображатимуть результати адекватно. Тому варто застосовувати макро- та мікро-варіанти метрик. Наприклад, точність (precision) визначається формулою (1):

$$precision_{micro} = \frac{TP_1 + \cdots + TP_k}{TP_1 + \cdots + TP_k + FP_1 + \cdots + FP_k} \quad (1)$$

де $precision_i$, TP_i, FP_i, $i = 1, \ldots, k$ – точність, доля правильних спрацювань, доля хибних спрацювань для i-го класу відповідно.

На тестовому наборі даних серед розглянутих моделей найкращий результат показав алгоритм Random Forest за оптимальний час (коефіцієнт кореляції Метьюса дорівнює 0.479 при базових налаштуваннях моделі без попередньої обробки незбалансованості класів).

Висновки

У даній статті було запропоновано концептуальну модель системи діагностики хвороб системи кровообігу. Було наведено опис набору даних,

що використовується в даній роботі та наведено перелік алгоритмів машинного навчання, їхні переваги та недоліки. Після перевірки моделей на тестовому наборі даних найкращі результати дав Random Forest.

Подальша робота полягає в навчанні моделей на повних даних, ретельний підбір гіперпараметрів та створення повноцінного застосунку для демонстраційного варіанту системи.

Література

1. Всесвітня організація охорони здоров'я. Європейське регіональне бюро. (2020). STEPS поширеність факторів ризику неінфекційних захворювань. Україна 2019. Всесвітня організація охорони здоров'я. Європейське регіональне бюро. – Режим доступу до ресурсу: https://iris.who.int/handle/10665/336643
2. Серцево-судинні захворювання — головна причина смерті українців. Висновки з дослідження Глобального тягаря хвороб у 2019 році – Режим доступу до ресурсу: https://phc.org.ua/news/sercevo-sudinni-zakhvoryuvannya-golovna-prichina-smerti-ukrainciv-visnovki-z-doslidzhennya
3. Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/C2XW26
4. Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016). https://doi.org/10.1038/sdata.2016.35