УДК 004.42:616.1

А. О. САВЧУК*, студентка, Н. Н. ШАПОВАЛОВА, І. О. ДОЦЕНКО, ст. викладачі
Криворізький національний університет

ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ В МЕДИЧНИХ КОНСУЛЬТАТИВНО-ДІАГНОСТИЧНИХ СИСТЕМАХ

Мета. Розробити і теоретично обґрунтувати ефективність застосування методів машинного навчання у ранній діагностиці серцево-судинних захворювань на догоспітальному етапі. Об'єктом дослідження є програмний засіб виявлення серцево-судинних захворювань; предметом дослідження – розробка програмного модуля консультативно-діагностичної системи з використанням апарату машинного навчання.

Методи. Для вирішення поставлених завдань використовувалися наступні методи: загальнонаукові методи теоретичного дослідження: аналіз комплексу ознак серцево-судинних захворювань за їх основними класами, синтез отриманих даних і перетворення їх у векторну модель, формалізація методів отримання інформативних ознак, моделювання процесу класифікації отриманих даних за певними типами ознак, узагальнення; методи емпіричного дослідження: вивчення досвіду в області поставленого завдання, тестування отриманої моделі; методи об'єктно-орієнтованого проектування та програмування.

Наукова новизна. Теоретично обґрунтовано необхідність процедури відбору сукупності клінічних ознак і лабораторних досліджень, як таких, які є найбільш інформативними для раннього діагностування серцево-судинних захворювань. Процедуру виокремлення таких ознак реалізовано на основі стратегії навчання з вкладеною технікою відбору ознак lasso.

Практична значимість виконаної роботи полягає в значному прискоренні процесу ранньої діагностики серцево-судинних захворювань на догоспітальному етапі за рахунок відбору інформативних клінічних ознак на стадії пре-процессінгу даних. Досягнуто задовільного рівня якості моделі визначення приналежності пацієнта до групи ризику серцево-судинних захворювань за рахунок використання методу машинного навчання – градієнтного бустінга. В якості метрик якості обрано долю вірних відповідей, точність і повноту, баланс оптимальних значень метрик знайдено на основі використання гармонічного середнього між точністю і повнотою, – *F*-міри.

Результати. Розроблено програмний модуль консультативно-діагностичної системи, який дозволяє оперативно проводити ранню діагностику серцево-судинних захворювань на основі методу машинного навчання і є доцільним для використання у медико-діагностичних центрах.

Ключові слова: серцево-судинні захворювання, рання діагностика, машинне навчання, градієнтний бустінг, регуляризація, точність, повнота.

doi: 10.31721/2306-5451-2019-1-49-113-117

Проблема та її зв'язок з науковими і практичними задачами. На сьогоднішній день серцево-судинні захворювання (ССЗ) є найбільш поширеною причиною загибелі серед неінфекційних захворювань у всьому світі. У країнах з низьким і середнім рівнем доходу на серцево-судинні захворювання припадає 75% випадків смерті. До серцево-судинних захворювань відносяться: артеріальна гіпертонія, ішемічна хвороба серця, судинні захворювання головного мозку, а також серцева недостатність і порушення ритму серця. Ішемічна хвороба серця (ІХС) є найпоширенішим захворюванням, так у 2017 році Україна посіла перше місце в світі за смертністю від ІХС [1].

Джерелом таких високих показників є уникнення ранньої медичної діагностики через її дорожнечу та тривалість. Це унеможливлює проведення первинної профілактики – запобігання розвитку ССЗ, в результаті чого пацієнти, які вже страждають цими хворобами, змушені витрачати більше коштів на лікування, що призводить до зубожіння сімей і супроводжується зниженням рівня економіки країни.

Рішенням даної проблеми може стати застосування системи ранньої діагностики ССЗ на догоспітальному етапі.

Аналіз досліджень і публікацій. Медичні консультативно-діагностичні системи – це такі системи, які використовуються для визначення стану пацієнта в даний період часу, а також аналізу розвитку патологій в майбутньому і видачі рекомендацій щодо подальшої тактики лікування. Вхідною інформацією для таких систем є дані, які були отримані в результаті діалогу з користувачем. Оскільки пацієнт може неправильно інтерпретувати симптоми, а лікар – не мати достатню кваліфікацію для вирішення виниклої проблеми, вхідна інформація доповнюється даними, отриманими за допомогою приладо-комп'ютерних систем, і являє собою більш об'єктивну оцінку того, що відбувається. На виході МКДС пропонує лікарю рішення проаналізова-

ної проблеми, обґрунтування прийнятого системою висновку, а також можливі альтернативи. Таким чином, МКДС підтверджують гіпотези лікаря і сприяють підвищенню його кваліфікації.

На сьогоднішній день, для діагностики пацієнта, як правило, використовуються експертні системи. Це такий підвид МКДС, який передбачає імітацію лікарської логіки для прийняття рішення в певній предметній області на основі розробленої бази знань, що містить досвід висококваліфікованих фахівців [2]. Експертні системи класифікують як системи штучного інтелекту, що передбачають використання евристичних алгоритмів для знаходження вирішення поставленого завдання. Для коректної роботи даних систем необхідно постійне втручання фахівців – інженерів по знаннях, з метою перевірки адекватності роботи системи та внесення змін до бази знань. Таким чином, обсяги, які міститься в системі інформації постійно зростають, а значить зростає і обсяг правил обробки цієї інформації, тим самим збільшуючи час для знаходження рішення. До того ж існує колосальна безліч діагностичних правил, неочевидних на перший погляд, що робить дуже важкою можливість врахувати всі чинники для адекватної роботи системи.

Іншим підвидом МКДС є системи штучного інтелекту, що передбачають використання методів машинного навчання. Дані системи вимагають менше місця для зберігання інформації і працюють в рази швидше експертних систем, що є їх безперечною перевагою, не дивлячись на додатковий час, необхідний для їх навчання на початку процесу розробки. Крім цього, системи, побудовані за допомогою методів машинного навчання, можуть використовуватися для пошуку неочевидних на перший погляд залежностей. В діагностиці медичних патологій, такими факторами можуть послужити захворювання, які часто зустрічаються, з нехарактерними для них симптомами, симптоми, які важко виявляються, а також пошук рідкісних захворювань і класифікації різних захворювань з однаковими, на перший погляд, симптомами.

В даний час виділяють такі способи машинного навчання: класичне навчання (без вчителя, з учителем), навчання з підкріпленням і глибоке навчання (нейронні мережі). Нижче розглянуті приклади реалізації деяких з цих способів.

Санджив Дж. Шах та ін. [3] запропонували своє рішення для прогнозування виживання людей, які страждають серцевою недостатністю. Вони створили модель навчання без вчителя за 46 ознаками, яка вирішувала завдання кластеризації пацієнтів на три окремі феногрупи. У якості навчальної вибірки було досліджено дані 397 пацієнтів від 53 до 77 років. Використавши модель навчання з вчителем, вони спрогнозували чотири ймовірних наслідки для кожної з груп: госпіталізація по причині серцевої недостатності, серцево-судинна госпіталізація, смерть та сукупний результат серцево-судинної госпіталізації чи смерті. До переваг методу можна віднести прискорення процесу обчислення для великої кількості об'єктів за рахунок вирішення задачі кластеризації. В якості недоліка можна виділити постійну необхідність коригувати шаблон кластера, тобто дослідження повинно бути підтверджено іншими когортами.

Модель, запропонована Эдвардом Чоєм [4] – це алгоритм прогнозування діагностики серцевої недостатності з використанням методу глибокого навчання – вентильних рекурентних вузлів. Навчальні дані були отримані від пацієнтів первинної медичної допомоги Sutter Palo Alto Medical Foundation (приблизно 58 652 000 медичних кодів, призначених пацієнтам). Перевагою методу є більша ефективність в прогнозуванні порівняно з іншими класами нейронних мереж. До недоліків відносяться потреба у великій навчальній вибірці та складність реалізації.

Л. Ясницький та ін. представили діагностичну систему у вигляді моделі нейронних мереж, яка за 51 ознаками діагностує 9 різних захворювань серцево-судинної системи, таких як: ішемічна хвороба серця, хронічна серцева недостатність, стенокардія стабільна, миготлива аритмія, аритмія і блокади серця, гіпертонія, стенокардія нестабільна, інфаркт міокарда, екстрасистолія. Для навчальної вибірки було оброблено 116 анкет, серед яких були як анкети здорових людей, так і анкети тих, чиї захворювання підтверджені ЕКГ діагностикою та лабораторними тестами. Перевагою методу можна вважати адекватні прогнози щодо захворювання ішемічної хвороби серця. Недоліком є велика похибка прогнозування інших восьми діагнозів [5].

Постановка завдання. На основі певних клінічних симптомів і лабораторних досліджень формується простір ознак, який описує стан пацієнта. Необхідно визначити приналежність об’єкту до певного рівню захворюваності.

В ході дослідження використовувались відкриті дані, з фонду Клівлендської клініки, відомі як Клівлендська база хвороб серця. Набір містить 14 ознак для 303 осіб, які страждають на хвороби серця. Пацієнти розділені на п’ять груп за рівнем захворюваності. Кожна група має рівень

від 0 (хвороба не виявлена) до 4. Серед ознак вік, стать, артеріальний тиск, біль у грудях, рівень цукру у крові, результати ЕКГ, кількість крупних сосудів тощо.

Досліджувана задача відноситься до класу задач навчання за прецедентами (supervised learning). Кожен прецедент являє собою пару «об'єкт – відповідь» [6]. Потрібно знайти функціональну залежність відповідей від описів об'єктів і побудувати алгоритм, який бере на вході опис об'єкта і видає на виході відповідь [7, 8].

Викладення матеріалу та результати. Задачу пропонується вирішити популярним ансамблієвим методом штучного інтелекту – градієнтним бустінгом над вирішальними деревами. На сьогоднішній день метод градієнтного бустінгу є одним з кращих способів спрямованої побудови композиції [9].

Ансамбль, або композиція, будується з простих базових алгоритмів. В якості базового алгоритмам використовується звичайна модель машинного навчання, така як дерево рішень (рис. 1).

Бустінг – це спосіб побудови композицій з дерев рішень, в рамках якого базові алгоритми будуються послідовно, один за одним і кожен наступний алгоритм будується таким чином, щоб виправляти помилки вже побудованої композиції (1).

$$a_N(x) = \sum_{n=1}^{N} b_n(x), \tag{1}$$

де $b_n(x)$ – базові алгоритми (дерева рішень) на просторі ознак x.

Для вирішення поставленого завдання будемо використовувати програмну бібліотеку з відкритим кодом XGBoost для Python. Перед тим, як почати процедуру навчання і валідації моделі, необхідно виокремити ті ознаки і результати клінічних досліджень, які є інформативними для постановки діагнозу. Для цього використаємо стратегію навчання моделей з вкладеною технікою відбору ознак – Lasso [10]. Ідея техніки полягає у накладанні штрафу за надвеликі значення параметрів моделі (зазвичай такі значення з'являються при мало інформативних ознаках), і зведенні їх до нуля. В якості штрафної функції використовується L1- норма вектора вагових коефіцієнтів (2)

Рис. 1 Схема додавання базових алгоритмів до композиції

$$\|\omega\|_1 = \sum_{j=1}^{d} |\omega_j|, \tag{2}$$

де ω – вектор вагових коефіцієнтів, d – кількість ознак.

Оскільки штрафна функція є сумою модулів вагових коефіцієнтів, а модульна функція не є гладкою, тому при оптимізації функціоналу помилки градієнтним методом, робиться припущення про те, що в точці нуль похідна функції дорівнюватиме нулю (хоча формально похідна в нулі не існує) [11, 12]. Похідна модуля є константною величиною, тому градієнтний спуск прямує до нуля з постійною швидкістю, а потрапивши до нього, там і залишається. Таким чином, неінформативні ознаки отримують нульові значення вагових коефіцієнтів, тим самим спрощуючи модель і водночас запобігають ефект перенавчання. Цей ефект спостерігається, коли на виборці, яка побувала у навчанні, модель показує прийнятні значення якості, а на новому наборі даних – неприйнятні, або зовсім неадекватні.

Після проведеного відбору, побудуємо ансамблієву модель XGBoost:

```
from sklearn.model_selection import GridSearchCV
clf = xgb.XGBClassifier()
parameters = {
     "eta"              : [0.05, 0.10, 0.15, 0.20, 0.25, 0.30 ],
     "max_depth"        : [ 3, 4, 5, 6, 8, 10, 12, 15],
     "min_child_weight" : [ 1, 3, 5, 7 ],
     "alpha"            : [ 0.0, 0.1, 0.2 , 0.3, 0.4 ],
     "colsample_bytree" : [ 0.3, 0.4, 0.5 , 0.7 ] }
grid = GridSearchCV(clf,
                    parameters, n_jobs=4,
```

```
                    scoring="neg_log_loss",
                    cv=3)
grid.fit(X_train, Y_train)
```

Оскільки налаштування методу залежить від багатьох гіперпараметрів, таких як кількість базових алгоритмів у композиції, глибини дерев, критерію зупину побудови дерева, обсягу вибірки, яка припадає на один базовий алгоритм, коефіцієнтів регуляризації [14], було прийняте рішення підібрати оптимальні значення гіперпараметрів за допомогою методу GridSearchCV, який реалізує метод «підгонки» і «оцінки» (рис. 2).

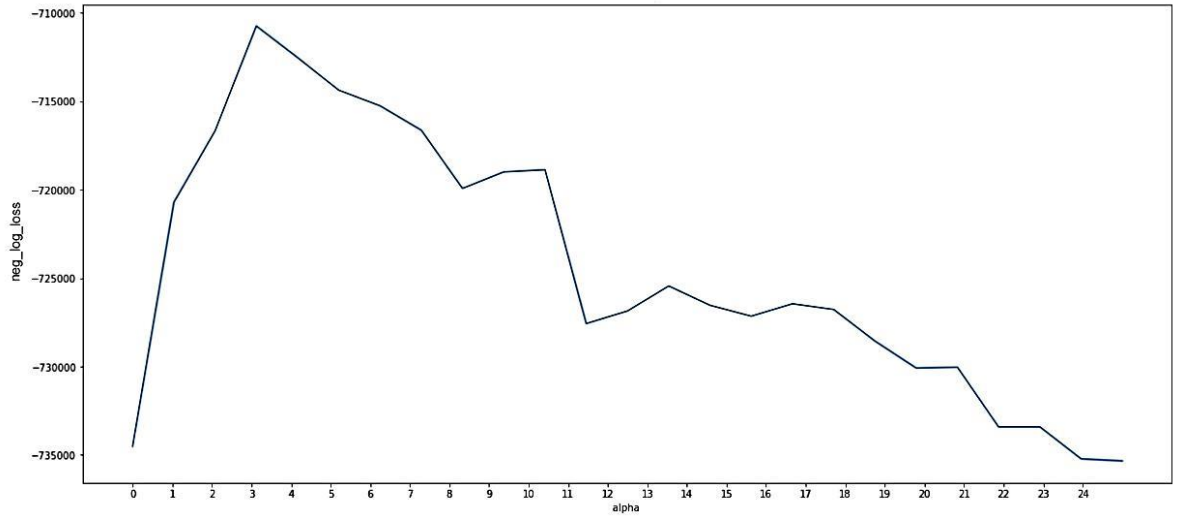

Рис. 2 Графік залежності якості алгоритму від коефіцієнта регуляризації

Побудовано композицію з 5000 дерев глибиною 4. Ці параметри були підібрані емпірично, виходячи з міркувань розумного співвідношення витрат часу на навчання моделі і прийнятної величини якості класифікатора.

Для отінення якості класифікатора будемо використовувати точність (precision) і повноту (recall). Ці метрики використовуються при оцінці здебільшого числа алгоритмів вилучення інформації. Точність показує наскільки можна довіряти класифікатору в разі спрацювання (алгоритм відносить об’єкт до певного класу, і цей об’єкт дійсно належить до цього класу). Повнота показує на якій частці справжніх об'єктів певного класу алгоритм спрацьовує. Зрозуміло що чим вище точність і повнота, тим краще. Але максимальна точність і повнота практично недосяжні одночасно і доводиться шукати якийсь баланс. Тому будемо додатково використовувати метрику, яка об'єднує у собі інформацію про точність та повноту алгоритму – *F*-міру. *F*-міра є гармонійне середнє між точністю і повнотою. Вона прямує до нуля, якщо точність або повнота прямують до нуля. Будемо розрахувати *F*-міру надавши різну вагу точності і повноті (3), тому що ми хочемо виявити як можна більше пацієнтів, схильних до ССХ, не менш, ніж 80%. Тоді ставлять завдання максимізації точності за умови того, що повнота має бути більшою за 0,8.

$$F = \left(\beta^2 + 1\right)\frac{Precision \times Recall}{\beta^2 Precision + Recall}, \qquad (3)$$

де β приймає значення в діапазоні 0 <β <1 якщо пріоритет віддається точності, а при β> 1 пріоритет віддається повноті.

Прийнята в роботу модель дає якість класифікації у розмірі 81%, при значеннях повноти і точності також 81%.

Висновки та напрямок подальших досліджень. Таким чином розроблений модуль діагностики серцево-судинних захворювань медичної консультативно-діагностичної системи з урахуванням найбільш інформативних лабораторних і клінічних ознак, відібраних за технологією lasso-регуляризації, є новим високоефективним засобом ранньої діагностики серцево-судинних захворювань, що дозволить підняти на більш новий рівень якість медичного обслуговування пацієнтів на догоспітальному етапі, які належать до групи ризику. Розроблений модуль надає лікарям якісний сучасний інструмент для знаходження оптимального рішення при постановці діагнозу. Основною особливістю програмного модуля є те, що довготривалий процес навчання

і валідації моделі, що робить прогнозні припущення, виконується на етапі запуску системи, а під час експлуатації система функціонує дуже швидко. Модуль має простий і інтуїтивно зрозумілий інтерфейс, прогнозований діагноз виводиться з вказанням рівню вірогідності.

В ході подальших досліджень отримані висновки планується перевірити з застосуванням інших стратегій і методів штучного інтелекту, зокрема, штучних нейронних мереж прямого розповсюдження.

Список літератури

1. **Д. С**. **Полякова.** «Этот хрупкий мир»: Украина – на 1-м месте по смертности от ишемической болезни сердца / **Полякова Д.С**. // Издательство «МОРИОН». – 2019.

2. **Б. А. Корбинский.** Консультативные интеллектуальные медицинские системы: классификации, принципы построения, эффективность / **Б.А. Корбинский**. // Врач и информационные технологии. – 2008. – С. 38–47.

3. **Sanjiv J. Shah**. Phenomapping for Novel Classification of Heart Failure With Preserved Ejection Fraction / **Sanjiv J. Shah, Daniel H. Katz, Senthil Selvaraj**. // Circulation. – 2015. – №131. – С. 269–279.

4. **Edward Choi.** Using recurrent neural network models for early detection of heart failure onset / **Edward Choi, Andy Schuetz, Walter F Stewart, Jimeng Sun**. // Journal of the American Medical Informatics Association. – 2017. – №24. – С. 361–370.

5. **Л. Н. Ясницкий.** Нейросетевая система экспресс-диагностики сердечно-сосудистых заболеваний / **Л. Н. Ясницкий, А. А. Думпер, А. Н. Полещук**. // Пермский медицинский журнал. – 2011. – №4. – С. 77–86.

6. Порівняльний аналіз методів оптимізації функціоналу якості моделей машинного навчання / **Н. Н. Шаповалова**, **О. Г. Рибальченко, Д. І. Куропятник**. // Вісник Криворізького національного університету : зб. наук. праць / національний університет Криворізький ; М-во освіти і науки України, ДВНЗ «КНУ». – Кривий Ріг, 2018. – Вип. 46. –С. 104 – 112.

7. **Луис Педро Коэльо**. Построение систем машинного обучения на языке Python / **Луис Педро Коэльо**, **Вилли Ричарт**. 2-е издание / пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2016. – 302 с.

8. **Петер Флах**. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. / **Петер Флах** . // Учебник / пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 408 с.

9. **Л. Н. Ясницкий.** Введення в штучний інтелект / **Л. Н. Ясницкий**. – видання 1-е. – Издательский центр «Академия», 2005. – 176 с.

10. **Л. В. Забуранна.** Оптимізаційні методи та моделі. Підручник. / **Л. В. Забуранна, Н. В. Попрозман, Н. А. Клименко, О. І. Попрозман, С. В. Забуранний**. – К.: __, 2014. – 372 с.

11. **В. В**. **Вітлінський** , Наконечний С. І., Терещенко Т. О. Математичне програмування / **В. В**. **Вітлінський** , **Наконечний С. І., Терещенко Т. О.//** Навч.-метод. посібник для самост. вивч. дисц. – К.: КНЕУ, 2001. – 248 с.

12. **Нейт Сильвер**. Сигнал и Шум. Почему одни прогнозы сбываются, а другие – нет / **Нейт Сильвер** // Азбука-Аттикус, КоЛибри, 2015. – 400 с.

13. **William M. Bolstad**. Introduction to Bayesian Statistics / **William M. Bolstad //** 2nd Edition // Wiley-Interscience; 2nd edition.

14. **Г. К**. **Вороновский**. Генетичні алгоритми, штучні нейроні мережі і проблеми віртуальної реальності./ **Г. К**. **Вороновский, Махотило К. В., Петрашев С. Н., Сергєєв С. А.** – Замовне. – Х.: ОСНОВА, 1997. – 112 с.

Рукопис подано до редакції 18.11.2019

УДК 622.7: 004.94

О.І. САВИЦЬКИЙ*, канд. техн. наук, доц., М.А. ТИМОШЕНКО, асистент
Криворізький національний університет

РОЗРОБКА АВТОМАТИЗОВАНОЇ СИСТЕМИ КЕРУВАННЯ СЕКЦІЄЮ ЗБАГАЧЕННЯ ЗАЛІЗНОЇ РУДИ З ВИКОРИСТАННЯМ ВІРТУАЛЬНИХ АНАЛІЗАТОРІВ ПРИ МОДЕЛЮВАННІ ЗАСОБАМИ ПРОГРАМОВАНОГО КОНТРОЛЕРА

Мета. Метою даної роботи є створення моделі ділянки секції збагачувальної фабрики з використанням віртуальних аналізаторів при моделюванні засобами програмованого контролера для моніторингу та контролю основних показників роботи технологічних механізмів та стану оброблюваного продукту на різних етапах операцій збагачення. Складність, інерційність, нестаціонарність та динамічність технологічних процесів, що відбуваються на збагачувальній фабриці, наявність складних зв'язків та рециклів між технологічними механізмами обумовлюють застосування нестандартного підходу для створення моделі секції, а саме розглядання моделі з точки зору застосування програмованих логічних контролерів і, відповідно, використання для створення моделі середовища програмування ПЛК. В умовах мультиагентного керування модель повинна розбиватись на окремі частини для кожного агента керування.

Методи дослідження. Підтверджено можливість застосування нетрадиційного програмного середовища для моделювання роботи секції збагачувальної фабрики. У той час коли для створення моделі об'єкту керування зазви-