DOI: https://doi.org/10.15276/ict.01.2024.33
УДК 004.93

Гетерогенний ансамблевий класифікатор у комп'ютерних системах медичної діагностики

Поворознюк Анатолій Іванович[1)]
Д-р техн. наук, професор каф. Комп'ютерної інженерії та програмування
ORCID: https://orcid.org/.0000-0003-2499-2350; ai.povoroznjuk@ Scopus Author ID: 55225664000
Поворознюк Оксана Анатоліївна[1)]
Канд. техн. наук, доцент каф. Комп'ютерної інженерії та програмування
ORCID: https://orcid.org/0000-0001-7524-5641; povoks76@gmail.com. Scopus Author ID: 55817007400
Філатова Ганна Євгенівна[1)]
Д-р техн. наук, професор каф. Комп'ютерної інженерії та програмування
ORCID: https://orcid.org/0000-0003-1982-2322; filatova@gmail.com. Scopus Author ID: 56448583600
[1)] Національний технічний університет «Харківський політехнічний інститут» вул. Кирпичова, 2. Харків, 61002, Україна

АНОТАЦІЯ

Робота присвячена вирішенню актуальної науково-технічної задачі розробки системи підтримки прийняття діагностичних рішень в медицині на основі розробленої моделі гетерогеного ансамблевого класифікатора, який в якості базових моделей реалізує два підходи до розробки діагностичних висновків: ймовірнісний, на основі аналізу навчальної вибірки та експертна інформація про структуру симптомокомплексів кожного захворювання. В якості ймовірнісної складової обгрунтовано вибір методу порівняння з еталоном, в якому діагностуємі стани пацієнтів представляються їхніми еталонами в просторі ознак. В якості еталона кожного класу вибирається геометричний центр угруповання класу в просторі діагностичних ознак. Експертні знання щодо структури симптомокомплексу формалізуються шляхом вираження симптомокомплексу захворювання у вигляді числових інтервалів лінгвістичних зміних "нижче норми", "норма", "вище норми". Розглянуто різні варіанти врахування думки експертів про структуру симптомокомплексів у ансамблевому класифікаторі. Тестова перевірка розробленого класифікатора були проведені на реальних медичних даних, яка підтвердила його ефективність.

Ключові слова: медичний діагноз; ансамблевий класифікатор; метод порівняння з еталоном; симптомокомплекс; експертна інформація

Актуальність. Розвиток високотехнологічного суспільства сприяє впровадженню досягнень науково-технічного прогресу в такій важливій, але недостатньо формалізованій сфері діяльності, як медична діагностика. На сьогодні існує сім рівнів медичних інформаційних систем (МІС) [1-2]: від електронних медичних карт до інтелектуальних систем підтримки прийняття рішень, які використовують синергетичні бази даних, методи штучного інтелекту [3] та можливості телемедицини [4].

Більшість із цих систем спрямовані на автоматизацію проведення обстеження пацієнтів у різних предметних галузях охорони здоров'я, які пов'язані з виявленням і реєстрацією діагностичних ознак (маркерів), обробкою біомедичних сигналів та зображень [5, 6].

Незважаючи на різноманіття сучасних систем МІС і складність проблем, які вони вирішують, однією з досі невирішених проблем оптимізації є синтез діагностичних вирішальних правил для забезпечення надійності та точності діагностики.

Загальний опис проблеми. Діагноз (грец. διάγνωσις — розпізнавати, ідентифікувати) – медичний висновок про функцію здоров'я, морфологічний стан людини, наявні захворювання чи травми, нормальне функціонування органів і систем, а також зміни в усьому організмі, або про причини, що призвели до летального результату. Тобто діагноз, це висновок щодо стану пацієнта, виражений термінами, передбаченими прийнятою класифікацією та номенклатурою захворювання.

Традиційна первинна діагностика ґрунтується на систематичному огляді хворого, аналізі анамнезу хворого, скарг та об'єктивних симптомів захворювання, які виявлені під час фізичного обстеження – огляду, перкусії тощо, а також на результатах інструментального

обстеження та лабораторних аналізів. На інтуїтивному рівні практикуючий лікар використовує ці міркування, щоб поставити діагноз, який відповідає принципам доказової медицини. Виявлені ознаки (симптоми) захворювання лікар групує в синдром (декілька симптомів спільного патогенезу), на підставі чого робить остаточний висновок про можливе захворювання. Слід відмітити, що при постановці діагнозу важливу роль відіграє майстерність і досвід лікаря.

В МІС різних рівнів формулювання комп'ютерної діагностики формально представлено класичною задачею класифікації, де модель об'єкта діагностики є «чорним ящиком», в якій шукається зв'язок між формалізованими станами об'єкта діагностики Y і вектор вхідних ознак X, тобто $Y = f(X)$ [1]. Цей зв'язок визначається на етапі навчання моделі класифікатора шляхом аналізу навчальної вибірки, що складається з пацієнтів із підтвердженим діагнозом. Однако, не завжди вдається отримати репрезентативну вибірку в існуючих медичних базах даних, особливо при діагностиці рідкісних захворювань. Крім того, практикуючі лікарі не завжди довіряють результатам комп'ютерної діагностики, та з обережністю інтерпретують результати діагностики.

Тому перспективним напрямом досліджень є розробка таких діагностичних методів, які поєднують вказані підходи традиційної та формалізованої комп'ютерної діагностики.

Метою дослідження є синтез гетерогенного ансамблевого класифікатора, який враховує два підходи до формування діагностичного висновку: ймовірнісний метод на основі аналізу навчальних вибірок та експертну інформацію щодо структури симптомокомплексів кожного захворювання.

Розробка моделі ансамблевого гетерогенного класифікатора. Поєднання окремих класифікаторів реалізован в ансамблевих методах машинного навчання [7, 8] для того, щоб зменшити вплив випадкових помилок або недоліків окремих класифікаторів для досягнення більш точних і надійних результатів, ніж використання кожної базової моделі класифікації окремо. Основними агрегованими методами машинного навчання є бегінг (Bootstrap Aggregating), бустинг (Boosting) і стекінг (Stacking) [9].

Ідея бегінг-ансамблю полягає в тому, щоб створити кілька випадкових підвибірок даних за методом вибірки з повтором, навчити різні моделі на цих підвибірках, а потім об'єднати їхні прогнози.

Принцип работи бустинг-ансамблів полягає в навчанні послідовних моделей, причому кожна нова модель зосереджується на попередніх помилках, що підсилює важливість правильного прогнозування екстремальних випадків.

Стекінг-ансамбль – це комбінація передбачень із різних моделей із використанням метамоделей, навчених на виходах базових моделей. Основною перевагою цього методу є можливість розглядати різні базові класифікатори для ефективного поєднання результатів, але він вимагає великих обчислювальних витрат і складних додаткових налагоджень параметрів.

Використання одного і того самого методу машинного навчання для створення базових моделей в рамках ансамблю є прикладом гомогенного (однорідного) ансамблю. У цьому сценарії однотипні класифікатори навчаються незалежно на різних навчальних даних, а їхні результати об'єднуються в остаточний результат.

Гетерогенні (неоднорідні) ансамблі дозволяють поєднувати різні базові методи машинного навчання для досягнення бажаної різноманітності та складності. Цей підхід визнає, що різні алгоритми можуть успішно фіксувати різні аспекти даних, а різні моделі можна комбінувати для більш детального аналізу.

Таким чином, для реалізації вищезазначених методів у комп'ютеризованій системі підтримки прийняття діагностичних рішень, рекомендується розробити гетерогенний ансамблевий класифікатор для прийняття діагностичних рішень.

В якості ймовірнісної компоненти гетерогенного класифікатора в роботі використовується метод порівняння з еталоном. Розглянемо особливості використання методу порівняння з еталоном в якості базової моделі. Цей метод використовується для аналізу кількісних

характеристик у випадку, коли класи Ω_m $(m=\overline{1,M})$ формують компактні множини об'єктів сферичної форми у функціональному просторі ознак.

Тоді кожен з класів Ω_m $(m=\overline{1,M})$ може бути представлений моделлю його еталона ω^{mr}, в якості якого обирається геометричний центр класу. На етапі навчання координати кожної моделі класу (еталона) розраховуються за формулою

$$x_i^{mr}=\frac{1}{n_m}\sum_{j=1}^{n_m}x_i^j\ ,\ (i=\overline{1,p}),(m=\overline{1,M})\ (1) \qquad (1)$$

де x_i^{mr} – *i*-та координата еталона класу *m*;

n_m – кількість об'єктів класу *m* у навчальній вибірці;

x_i^j – *i*-та координата *j*-го об'єкта класу *m у* навчальній вибірці*;*

p - розмір координатного простору (кількість діагностичних ознак);

M – кількість класів, на які виконується класифікація невідомих об'єктів (кількість захворювань, що діагностуються у цій галузі медицини).

На етапі класифікації (діагностики) обчислюється відстань від точки об'єкта, який підлягає класифікації $R(\omega,\ \omega^{mr})$ в просторі ознак до кожного еталонного об'єкта ω^{mr} класу *m* за формулою

$$R(\omega,\omega^{mr})=\sqrt{\sum_{i=1}^{p}(x_i^{\omega}-x_i^{\omega^{mr}})^2}\ ,\ (m=\overline{1,M})\ , \qquad (2)$$

де $x_i^{\omega^{mr}}$ – *i*-та координата еталонного об'єкта класу *m* ;

x_i^{ω} – *i*-та координати об'єкта, який підлягає класифікації;

p – розмір координатного простору (кількість діагностичних ознак);

M – кількість класів, на які виконується класифікація.

Об'єкт, що підлягає класифікації відноситься до класу Ω_t, відстань до еталону якого $R(\omega,\ \omega^{tr})$ буде мінімальною:

$$R(\omega,\omega^{tr})=\min_{m=1,M}R(\omega,\omega^{mr})\ , \qquad (3)$$

де $R(\omega,\ \omega^{mr})$ відстань між об'єктом ω та еталоном ω^{mr} класу Ω_t розраховується за виразом (2).

Врахування експертної інформації шодо структури симптомокомплексу. Інформацію про симптоми захворювання в неформалізованому виді можна отримати з різних медичних довідників [10, 11], і її можна розглядати як експертну оцінку захворювання, розроблену кількома поколіннями лікарів. Виділяють такі типи симптомів, що складають симптомокомплекс: *патогномонічні* (однозначно вказують на наявність певного захворювання – маркери захворювань), *специфічні* (проявляються при певному захворюванні, але чітко не вказують на його прояв), *неспецифічні* (можуть проявлятися при конкретному захворюванні, але не вказують однозначно на його наявність).

Структура симптомокомплексу представляє думку експертів щодо конкретного діагнозу, а вплив кожного симптому розраховується на основі певної лінгвістичної змінної, що описує відповідний симптом (наприклад, «висока температура» або «високий артеріальний тиск»).

Крім того, кожному симптому x_j присвоюється певний експертний рейтинг, який вказує його вагу в симптомокомплексі. Ці рейтинги приймають значення e_0, e_1, e_2, e_3 , де e_0 –вага патогномонічних симптомів захворювання; e_1 - вага специфічних симптомів; e_2 - вага неспецифічних симптомів; e_3 - вага показників, яке не належать до синдрому цього захворювання.

При цьому справедлива нерівність

$$e_0 \geq e_1 \geq e_2 \geq e_3, \ \sum_{i=0}^{3} e_i = 1. \quad (4)$$

Розглянемо порядок врахування думок фахівців щодо структури симптомокомплексу при створенні ансамблевого методу класифікації. При постановці діагнозу лікарі часто користуються поняттям норми того чи іншого симптома і ділять динамічний діапазон зміни симптомів на три піддіапазони: «нижче норми», «норма» і «вище норми». Подібним чином визначаються бінарні ознаки, для яких лінгвістичними змінними є два терми «проявляється ознака» або «не проявляється ознака».

Отже, кожен патологічний синдром (еталон кожного класу) в даному випадку визначається діагностичними симптомами x_j, кожен з яких приймає одне зі значень наведених вище лінгвістичних змінних. Опис типових наборів симптомів разом із відповідними варіантами лінгвістичних змінних і є формалізованою експертною оцінкою еталонів для кожного класу захворювання.

Якщо відомі значення динамічного діапазону зміни всіх діагностичних ознак, а також порогові значення «норми», то центри відповідних діапазонів: «нижче норми», «норма» і «вище норми» і є кількісним представленням еталона ω^{ml} класу Ω_l, який базується на висновках експертів щодо структури симптомокомплексу, а не на навчальній вибірці.

В остаточному підсумку, різні методи (методи на основі навчальної вибірки та методи на основі експертної оцінки симптомокомплексу) призначені для вирішення однієї і тієї задачі класифікації. Спираючись на різні вихідні позиції (статистику та експертну думку), вони висвітлюють загальну проблему з різних сторін. Природно вважати перспективним їх використання в якості базових моделей гетерогенного ансамблевого класифікатора, тому пропонуються наступні сценарії їх спільного використання:

1) Класична діаграма стекінг-ансамблевого гетерогенного класифікатора, в якому результати класифікації різних моделей об’єднуються в кінцевий результат за допомогою метамоделі.

У цьому випадку для кожного класу Ω_m $(m = \overline{1, M})$ створюються дві основні моделі класифікації шляхом створення окремих еталонів (ω^{mr} та ω^{m}). Базові моделі відносять кожний новий об’єкт ω, що підлягає класифікації, до класу Ω_t згідно виразами (2) та (3). Якщо результати класифікації кожної базової моделі відрізняються одна від іншої, результати цих моделей агрегуються і остаточний діагноз D_k формується за виразом

$$\min(R(\omega, \omega^{mr}), R(\omega, \omega^{ml})) \rightarrow D_k . \quad (5)$$

2) Агрегація моделей при визначені координат еталонів класів. Остаточний варіант запропонованого в даній роботі ансамблевого класифікатора полягає в агрегації результатів розрахунку координат еталонів класів різними моделями в методі порівняння з еталоном. У цьому випадку координати кожного еталону класу Ω_m $(m = \overline{1, M})$ розраховуються на етапі навчання за формулою

$$x_i^m = k_1 x_i^{mr} + k_2 x_i^{ml}, \quad (6)$$

де $k_i > 0$, $\sum k_i = 1$, $i = \overline{1, 2}$.

Вагові коефіцієнти k_i відповідають ступеню довіри до кожної складової, а саме довіра до репрезентативності навчальної вибірки k_1 або до формалізованої експертної оцінки симптомокомплексу певного класу захворювання k_2. Надалі нові об'єкти класифікуватимуться за стандартними алгоритмами (2) і (3) методу порівняння з еталоном. Однак діагноз, визначений запропонованим ансамблевим класифікатором, є підтримкою прийняття діагностичного рішення, яке остаточно приймається лікарем.

Результати тестування. Для тестової перевірки розробленого класифікатора створено навчальну вибірку, яка включає 200 пацієнток.

Діагностичними ознаками є 9 показників клінічних досліджень крові та сечі, за якими

визначено діагностичні стани захворювань молочної залози: D1 – кіста молочної залози; D2 – мастопатія; D3 – ліпома; D4 – фіброаденома; D5 – практично здорова.

Результати діагностики наведені в Таблиці, прийняті наступні умовні позначення: $P(x)$ – результати використання методу порівняння з еталоном; $\mu(x)$ – результат використання детермінованої логіки шляхом інтерпретації симптомокомплексу в термах нечіткої логіки; Ω – результат використання розробленого класифікатора; N – кількість правильно ідентифікованих об'єктів; % – відсоток від загальної кількості правильно ідентифікованих об'єктів.

У виразі (6) кожен компонент має однакову вагу, тобто $k_1 = k_2 = 0{,}5$.

За результатами Таблиці частка правильно класифікованих об’єктів за результатами клінічного аналізу при діагностиці захворювання молочної залози методом порівняння з еталоним становить 88 %, за детермінованою логікою – 87.5%, використання розробленого класифікатора – 94 %, що підтверджує функціональність та ефективність розробленого методу діагностики.

Висновки. В роботі обґрунтована необхідність синтезу класифікатора для задач медичної діагностики, який об’єднує два методи формування діагностичних висновків: ймовірнісний підхід на основі аналізу навчальних вибірок і метод, заснований на формалізації знань експерта про структуру симптомокомплексу. Розроблено математичну модель гетерогенного ансамблевого класифікатора, в якому базовими моделями є класифікатори з використанням методу порівняння з еталоном та на основі формалізації експертних знань про структуру симптомокомплексів. Запропоновано альтернативи агрегації результатів базових моделей. Результати тестування на основі реальних медичних даних підтвердили працездатність та продемонстрували ефективність діагностики. Подальші дослідження спрямовані на застосування розроблених правил прийняття рішень не лише до кількісних параметрів, але й до різних діагностичних даних, таких як біомедичні сигнали та зображення.

Таблиця. **Результати діагностики**

код	Кількість пацієнток	Результати діагностики					
		$P(x)$		$\mu(x)$		Ω	
		N	***%***	***N***	***%***	***N***	***%***
D1	40	35	87	34	85	38	95
D2	60	54	90	53	88	56	93
D3	20	17	85	17	85	19	95
D4	30	26	87	24	80	26	87
D0	50	45	90	47	94	49	98
Разом	200	177	88	175	87.5	188	94

СПИСОК ЛІТЕРАТУРИ

1. Аврунін О. Г., Бодянський Є. В., Калашник М. В., Семенець В. В., Філатов В. О. «Сучасні інтелектуальні технології функціональної медичної діагностики: монографія». *Харків : ХНУРЕ*. 2018.

2. Тимчик С. В., Злепко С. М., Костішин С.В. «Класифікація медичних інформаційних систем і технологій за інтегральним сукупним критерієм». *Системи обробки інформації*. 2016; 3 (140): 194–198.

3. Wu T., He S, Liu J., Sun S., Liu K., Han Q. L. and Tang Y. “A brief overview of ChatGPT: The history, status quo and potential future development”. *IEEE/CAA Journal of Automatica* Sinica, 2023; 10 (5): 1122–1136. DOI: https://doi.org/10.1109/JAS.2023.123618

4. Yang Y. T., Iqbal U., Horn-Yu Ching J. et al. “Trends in the growth of literature of telemedicine: A bibliometric analysis”. *Computer Methods and Programs in Biomedicine*. 2015;

122 (3): 471–479. DOI: https://doi.org/10.1016/j.cmpb.2015.09.008

5. Povoroznyuk A. I., Filatova A. E., Surtelb W., et al. “Design of decision support system when undertaking medical-diagnostic action”. *Optical Fibers and Their Applications*. Proc. of SPIE. 2015; 9816: 98161O1–98161O17. DOI: https://doi.org/10.1117/12.2229295.

6. Biloborodova T., Scislo L., Skarga-Bandurova I., Sachenko A., Molgad A., Povoroznjuk O., Yevsieiva Y. “Fetal ECG signal processing and identification of hypoxic pregnancy conditions in-utero Math Biosci Eng” . 2021; 18 (4): 4919–4942. DOI: https://doi.org/10.3934/mbe.2021250.

7. Kiflay A. Z., Tsokanos A., and Kirner R. “A network intrusion detection system using ensemble machine learning”. *International Carnahan Conference on Security Technology (ICCST)*. 2021. p. 1–6. DOI: https://doi.org/10.1109/ICCST49569.2021.9717397.

8. Sarkar A., Sharma H. S. and Singh M. M. “A supervised machine learningbased solution for efficient network intrusion detection using ensemble learning based on hyperparameter optimization”. *International Journal of Information Technology*. 2022; 15 (1): 423–434. DOI: https://doi.org/10.1007/s41870-022-01115-4

9. Hornostal O., Chelak V. and Gavrylenko S. “Research of Intelligent Data Analysis Methods for Identification of Computer System State”. *Proceedings of the 30th International Scientific Symposium Metrology and Metrology Assurance (MMA)*. 2020. p. 1–5. DOI: https://doi.org/10.1109/MMA49863.2020.9254252.

10. Мавров І. І. «Оптимальна діагностика та лікування захворювань шкіри та венеричних захворювань». Посібник для дерматологів і венерологів. Київ, Україна. *ТОВ «Доктор Медіа»*. 2007.

11. Григор'єва К. І. «Педіатрія: посібник для практикуючих лікарів». Київ, Україна. *Медпрес-інформ*. 2014.

DOI: https://doi.org/10.15276/ict.01.2024.33
UDC 004.93

Heterogeneous ensemble classifier in computer systems of medical diagnostics

Anatoly I. Povoroznyuk[1)]
Dr. Sc., Professor, Department Computer engineering and programming
ORCID: https://orcid.org/ 0000-0003-2499-2350 ; ai.povoroznjuk@ gmail.com. Scopus Author ID: 55225664000
Oksana A. Povoroznyuk[1)]
PhD, Associate Professor, Department Computer engineering and programming
ORCID: https://orcid.org/ 0000-0001-7524-5641; povoks76@gmail.com. Scopus Author ID: 55817007400
Anna Ye. Filatova[1)]
Dr. Sc., Professor, Department Computer engineering and programming
ORCID: https://orcid.org/0000-0003-1982-2322; filatova@gmail.com. Scopus Author ID: 56448583600
[1)] National Technical University “Kharkiv Polytechnic Institute”, 2, Kirpychova, Str . Kharkiv, 61002, Ukraine

ABSTRACT

The work is devoted to the solution of an actual scientific and technical problem tasks development of a support system for making diagnostic decisions in medicine based on the developed model of a heterogeneous ensemble classifier, which as basic models implements two approaches to development of diagnostic conclusions: probabilistic , based on the analysis of the training sample and expert information on the structure of symptom complexes of each disease. As a probabilistic component, the choice of the method of comparison with the standard, in which the diagnosed conditions of patients are represented by their standards in the space of signs, is justified. The geometric center of the class grouping in the space of diagnostic features is chosen as the benchmark for each class. Expert knowledge about the structure of the symptom complex is formalized by expressing the symptom complex of the disease in the form of numerical intervals of linguistic changes "below the norm", "norm", "above the norm". Various options for taking into account the opinion of experts about the structure of symptom complexes in the ensemble classifier are considered. Test verification of the developed classifier was conducted on real medical data, which confirmed its effectiveness.

Keywords: Medical diagnosis; ensemble classifier; method of comparison with the standard; symptom complex; expert information