

ПРОГНОЗ РИЗИКУ СЕРЦЕВОГО НАПАДУ З ВИКОРИСТАННЯМ WEKA

Тези представляють рішення проблеми класифікації для покращення продуктивності використання штучних нейронних мереж та розробку функцій задля вдосконалення відомого набору даних для прогнозування серцевих нападів. Дослідження спрямоване на підвищення прогностичної потужності та зменшення обсягу набору даних за допомогою методів глибокого навчання для потреб клінічної діагностики.

Підвищити точність прогнозування рівня ризику серцевого нападу можливе шляхом зменшення розміру набору необроблених даних. Це призводить до більш надійних, менш трудомістких і менш дорогих клінічних діагнозів.

Оптимальне зменшення розмірності та шумів у наборі даних про хвороби серця та застосування алгоритмів класифікації з використанням штучних нейронних мереж є перспективними для підвищення його прогностичної здатності до клінічно прийнятного стандарту.

Ключові слова: машинне навчання, набір даних про хворобу серця, ризик серцевого нападу, очищення від шуму, ефективність діагностики.

Серцевий напад, також відомий як інфаркт міокарда, виникає, коли артерія, яка постачає насичену киснем кров до серця, блокується. Така недостатність кровотоку може спричинити серйозне пошкодження серцевого м'яза та загрожувати життю. За даними Всесвітньої організації охорони здоров'я (ВООЗ), приблизно одна третина всіх смертей у світі викликана серцево-судинними захворюваннями, і приблизно 85% цих смертей є результатом серцевих нападів. Як наслідок, прогноз ризику серцевих нападів продовжує бути в центрі уваги клініцистів та експертів з машинного навчання [1].

Однак важливо зазначити, що точність прогнозів із цього набору даних лише незначно зросла з 1988 [2] до 2021 [1], досягнувши приблизно 90% [1]. В принципі, це вже було б досить точно, хоча, можливо, не для медичної діагностики. Тим не менш, ця відсутність прогресу свідчить про те, що може знадобитися перегляд і вдосконалення початкового набору даних за допомогою сучасних методів розробки функцій.

Набір даних, згаданий вище, датується 1988 роком і об'єднує чотири окремі набори даних. Він містить 13 атрибутів і одну додаткову цільову змінну, охоплює 303 екземпляри та не має відсутніх значень, як стверджується в [2]. Джерело також вказує, що попередня версія набору даних мала 76 атрибутів, але опублікована версія є більш короткою.

Таблиця 1

Опис набору даних [2]

ID	Назва атрибута	Тип	Опис	Одиниця	Діапазон
1	age	ціле число	Вік пацієнтів	років	29 - 77
2	sex	номінальний	Стать		0, 1 = жінка, чоловік
3	cp	номінальний	Тип болю в грудях		0 = Типова стенокардія, 1 = Атипова стенокардія, 2 = Неангінальний біль, 3 = Безсимптомний
4	trtbps	справжній	Артеріальний тиск у спокої	мм рт.ст	94 -200
5	chol	справжній	Холестерин за допомогою датчика ІМТ	мг/дл	126 - 564
6	fbss	номінальний	рівень цукру в крові натще > 120 мг/дл		0, 1 = false , true
7	restecg	номінальний	Результати ЕКГ у спокої		0, 1,3 = Нормальний, ST,?
8	thalachh	ціле число	Досягнуто максимального пульсу	уд/хв	71 - 202
9	oldpeak	справжній	Депресія ST, спричинена фізичним навантаженням відносно відпочинку	?	0 – 6,2
10	slp	номінальний	нахил піку навантаження на		0,1,2 = вгору, рівно, вниз

ID	Назва атрибута	Тип	Опис	Одиниця	Діапазон
			сегмент ST		
11	caa	ціле число	Кількість великих судин, забарвлених при рентгеноскопії		0 - 4
12	thall	номінальний	Результат стрес-тесту на талій		0 - 3
13	exng	номінальний	Стенокардія фізичного навантаження		0,1 = Ні, Так
14	output	номінальний	Цільова змінна		0,1 = менше шансів серцевого нападу, більше шансів серцевого нападу

З 303 випадків у 138 визначено низький ризик серцевого нападу, тоді як у 165 визначено високий ризик серцевого нападу. Набір даних містить вісім номінальних і п'ять числових атрибутів цілого чи дійсного (з плаваючою точкою) типів. П'ять числових атрибутів охоплюють різні масштаби, діапазони та природу.

Варто також нагадати застереження [3] щодо деяких проблем із набором даних:

- Розподіл даних у різних вікових групах, причому більшість тих, хто демонструє захворювання серця, зустрічаються у людей віком від 40 до 75 років;
- Чоловічі та жіночі записи не однаково представлені в даних, а кількість чоловічих записів значно вища, ніж жіночих записів;
- Кількість екземплярів здається занадто малою для ефективного використання алгоритмів глибокого навчання (DL).

Використання програмного забезпечення WEKA

Використовувалося програмне забезпечення WEKA на основі Java (v.3.9.6). Він включає три алгоритми класифікації, що використовують нейронні мережі (DL): DL4jMlp, багатошаровий перцептрон (MLP) і голосований перцептрон (VP).

Перш за все, початковий набір даних був стандартизований, що означає, що всі атрибути, крім цільової змінної, були відцентровані за середніми значеннями та нормалізовані на їх стандартні відхилення. Деякі класифікатори включають стандартизацію або нормалізацію як варіант, але краще зробити це своєчасно. Зауважимо, що таке перетворення було єдиним для вихідного набору даних.

Ранжування та вибір атрибутів: скорочений набір даних

Marple 2023, зокрема його пакети програм «LinearAlgebra» та «Statistics», використовувався в цьому розділі у співпраці з WEKA. Ми вибрали вісім алгоритмів WEKA для ранжування атрибутів, і результати можна знайти в таблиці 2 разом із результатами ранжування для всіх 13 атрибутів (за винятком цільового атрибута) початкового набору даних. У таблиці 2 (третій стовпець) і тексті нижче автори використовують ідентифікатори атрибутів з таблиці 1.

Таблиця 2

Ранжування атрибутів (ознак)

немає	Оцінювач WEKA	Ранги атрибутів (найкращі спочатку)
1	Фільтр рейтингу кореляції	13,9,10,8,3,12,11,2,1,7,4,5,6
2	Оцінювач функцій Gain Ratio	13,9,8,3,12,10,11,2,1,7,6,4,5
3	Інформаційний рейтинговий фільтр	13,3,12,10,9,8,11,1,2,7,6,5,4
4	Оцінювач функцій OneRule	13,3,12,9,11,10,8,1,2,7,4,5,6
5	Фільтр рейтингу RELIEF	12,3,13,2,11,9,7,10,8,1,6,4,5
6	PairwiseCorrelationAttributeEval	3,3,9,12,8,10,11,2,1,7,6,5,4
7	Симетричний фільтр ранжирування невизначеності	13,3,9,12,8,10,11,2,1,7,6,4,5
8	Оцінювач функцій J48	13,3,12,9,8,11,10,1,2,7,6,5,4

З таблиці 2 видно, що найчастіше першу позицію (ранг 1) займає ознака з ID 13, тобто «exng» (стенокардія фізичного навантаження). При цьому атрибути з ідентифікаторами 4,5,6 ("trtbps", "chol", "fbs") завжди мають три нижніх ранги.

Результатом процедури ранжирування є матриця рангів із 13 стовпцями (що представляють кількість характеристик) і вісьмома рядками (що відповідають кількості оцінювачів). Медіанний ранг для кожного атрибута (стовпця) можна обчислити разом із його середнім відхиленням. Автор віддає перевагу медіані над середнім, враховуючи цілочисельний характер рангів у згаданій вище матриці.

References

- Gupta, S.K., Shrivastava, A., Upadhyay, S.P., and Chaurasia, P.K. (2021). A Machine Learning Approach for Heart Attack Prediction. International Journal of Engineering and Advanced Technology.10,(6). <https://doi.org/10.35940/ijeat.f3043.0810621>

2. Janosi A., William S., Pfisterer M., and Detrano R., (1988). Heart Disease. UCI Machine Learning Repository. <https://doi.org/10.24432/C52P4X>
3. Ruqiya, Muhammad Tayyab Yaqoob, Abid Muhammad Khan, Aftab Ahmed Shaikh1, Noman Khan. Review on Cleveland Heart Disease Dataset using Machine Learning. Quaid-e-Awam Univ Res J Eng Sci Technol. 2023;21:87-98. <http://dx.doi.org/10.52584/QRJ.2101.11>

DOI 10.34132.mspc2025.01.08.42

**Olga Yaremchuk
Mariia But**

HEART ATTACK RISK PREDICTION WITH WEKA

The thesis presents a solution to the classification problem to improve the performance of artificial neural networks and the development of functions for improving a well-known dataset for predicting heart attacks. The study aims to increase the predictive power and reduce the size of the dataset using deep learning methods for clinical diagnostic purposes.

It is possible to improve the accuracy of heart attack risk prediction by reducing the size of the raw data set. This leads to more reliable, less time-consuming and less expensive clinical diagnoses.

Optimal dimensionality and noise reduction in the heart disease dataset and the application of classification algorithms using artificial neural networks are promising to improve its predictive ability to a clinically acceptable standard.

Keywords: machine learning, heart failure dataset, survival chances, noise cleaning, diagnostics performance

Відомості про автора / Information about the Author

Ольга Яремчук, аспірантка кафедри комп'ютерної інженерії, ст. викладач кафедри медичної біології та фізики, мікробіології, гістології, фізіології та патофізіології Навчально-наукового медичного інституту, Чорноморський національний університет ім. Петра Могили, Миколаїв, Україна. E-mail: olga.yaremchuk.77@gmail.com, ORCID. 0000-0002-0891-4216

Olga Yaremchuk, PhD student, Department of Computer Engineering, Senior Lecturer of the Department of Medical Biology and Physics, Microbiology, Histology, Physiology, and Pathophysiology, Educational and Scientific Medical Institute, Petro Mohyla Black Sea National University Mykolayiv, Ukraine, E-mail: olga.yaremchuk.77@gmail.com, ORCID. 0000-0002-0891-4216

Марія Бут, студентка 5 курсу Навчально-наукового медичного інституту Чорноморського національного університету імені Петра Могили, м. Миколаїв, Україна. E-mail: mbut2003@gmail.com

Mariia But, 5th year student, Educational and Scientific Medical Institute of the Petro Mohyla Black Sea National University, Mykolaiv, Ukraine. E-mail: mbut2003@gmail.com

© О. Яремчук, М. Бут
29.05.2025.

Стаття отримана редакцією