БІОЛОГІЧНІ ТА МЕДИЧНІ ПРИЛАДИ І СИСТЕМИ

УДК 004.032.26

О. К. КОЛЕСНИЦЬКИЙ, Ю. О. ЖУРАВСЬКА

Вінницький національний технічний університет, Вінниця

ЗАСТОСУВАННЯ НЕЙРОННОЇ МЕРЕЖІ КОХОНЕНА ДЛЯ МЕДИЧНОГО ДІАГНОСТУВАННЯ ПАЦІЄНТІВ ЗА АНАЛІЗОМ КРОВІ

Анотація: В даній статті запропоновано нейронну мережу Кохонена, в якій центри кластерів визначаються заздалегідь згідно з нормальними та відхиленими від норми значеннями показників загального аналізу крові, що підвищує достовірність попереднього діагностування пацієнтів.

Ключові слова: нейронна мережа Кохонена, кластеризація, медична діагностика.

Аннотация: В данной статье была предложена нейронная сеть Кохонена, в которой центры кластеров определяются заранее в соответствии с нормальными и отклонёнными от нормы значениями показателей общего анализа крови, что повышает достоверность предварительной диагностики пациентов.

Ключевые слова: нейронная сеть Кохонена, кластеризация, медицинская диагностика.

Abstract: A Kohonen neural network, in which the cluster centers are determined in advance in accordance with normal and deviations from the normal values of general blood count parameters that increases the reliability of preliminary diagnostics of patients, is proposed in current article.

Keywords: Kohonen neural network, clustering, medical diagnostics.

Вступ

На даний час глобальний процес автоматизації зачепив практично всі сфери людської діяльності, включаючи медичну галузь. Існуючий процес діагностування пацієнтів лікарень є досить довгим та багатоетапним – хворий повинен пройти реєстрацію, консультацію у сімейного лікаря (терапевта), отримати направлення до лікаря, спеціалізація якого має ймовірність виявитись хибною відносно реальної хвороби пацієнта. Тому щоб підтвердити або спростувати діагноз потрібно здати низку аналізів, що може призвести через довготривалість даного процесу до загострення хвороби або поставлення хибного діагнозу. У даній ситуації раціонально використовувати можливості сучасних інформаційних технологій та широке розповсюдження мобільних інформаційних пристроїв у населення.

Існуючі автоматизовані методи діагностики також є недостатньо ефективними – більшість систем орієнтована на визначення конкретного захворювання, що через специфіку медичної галузі, має високу ймовірність виявитись хибним; мають невиправдану складність алгоритмів та їх програмної реалізації; можуть використовуватись лише медичними фахівцями. [1]

Таким чином, існує проблема недосконалості сучасних методів діагностування, що не забезпечують високу точність виявлення захворювання, швидкість (оперативність) та зручність використання із залученням персональних мобільних засобів (телефони, смартфони, планшети та ін.), не можуть бути використані пацієнтами дистанційно без відвідування лікарні.

Для усунення даної проблеми доцільно застосувати підхід, заснований на визначені показників загального аналізу крові, що проводиться при більшості захворювань і профілактичних обстеженнях, і включає в себе 13 показників, кожен з яких при відхиленні від норми передбачає схильність до певних хвороб. Діагностування за допомогою аналізу отриманих показників аналізу крові допоможе виявити тип захворювання пацієнта та надати рекомендації щодо лікування. [2] Для реалізації даного підходу раціонально використати кластеризацію даних, яка зможе забезпечити високу точність їх обробки, при цьому мати нескладну програмну структуру. [3]

Задачею дослідження є обрання методу кластеризації, який забезпечує найбільш якісну реалізацію сформульованого підходу до діагностування, що повинен гарантувати високу точність та швидкість виявлення типу захворювання. При цьому на основі значень показників загального аналізу крові потрібно визначити попередній діагноз пацієнту, тобто групу, до якої відноситься можливе захворювання.

Таким чином, **мета цієї статті** – аналіз відомих методів кластеризації даних та обрання найточнішого для забезпечення достовірного попереднього діагностування пацієнтів на основі аналізу крові з метою надання рекомендацій щодо лікування.

Аналіз відомих методів кластеризації

Основними методами кластеризації є група ієрархічних алгоритмів, алгоритм k-means та його похідні, дерево мінімального покриття, метод найближчого сусіда, алгоритм нечіткої кластеризації, нейронні мережі та генетичні алгоритми.

Ієрархічні алгоритми спираються на вихідну «гіпотезу компактності»: у просторі об'єктів всі близькі об'єкти повинні ставитися до одного кластеру, а всі різні об'єкти відповідно повинні знаходитися в різних кластерах, що може призвести до великої похибки при кластеризації даних. Основним недоліком

методу k-means є те, що потрібно заздалегідь задавати *k* - кількість кластерів та еталонів, що не завжди можливо зробити раціонально. Метод є дуже чутливим до початкових наближень значень центрів. Дерево мінімального покриття може виділяти кластери довільної форми, але воно має громіздку структуру та ненаочність подання результатів кластеризації. Алгоритм найближчого сусіда простий у реалізації, швидко виконується, але, як і інші «жадібні» алгоритми, може видавати неоптимальні рішення. Нечіткі алгоритми недоцільно використовувати, якщо заздалегідь невідомо число кластерів, або необхідно однозначно віднести кожен об'єкт до одного кластеру. Для реалізації кластеризації даних мережею Кохонена використовується універсальний кластеризатор – нейронна мережа, передбачається навчання мережі без вчителя та самоорганізація мережі, існує суттєва простота реалізації та гарантоване отримання відповіді після проходження даних по шару нейронів. Недоліками є робота тільки з числовими даними, мінімізація розмірів мережі та необхідність задання кількості кластерів. Генетичні алгоритми не гарантують виявлення глобального рішення за прийнятний час, не гарантують, що знайдене рішення буде оптимальним та мають переваги перед іншими алгоритмами лише при дуже великих розмірах завдань і відсутності впорядкованості у вихідних даних, коли альтернативою є метод повного перебору варіантів. [4]

Оскільки для кластеризації результатів показників загального аналізу крові потрібні лише числові значення та велика точність в їх визначенні, для реалізації даної задачі була обрана мережа Кохонена, найефективніша нейронна мережа кластеризації числових даних. [5]

Оскільки стандартна реалізація мережі Кохонена не передбачає визначеної кількості кластерів, та їх початкові центри приймають випадкові малі значення, була проведена модифікація даної мережі, що полягає у попередньому визначенні кількості кластерів та їх центрів, які відповідають нормальним та відхиленим від норми значенням показників загального аналізу крові.

Модернізація нейронної мережі Кохонена

Задача, що розглядається, запропонована до вирішення модифікованою мережею Кохонена. Дана мережа використовує неконтрольоване навчання та навчальна множина складається лише із значень вхідних змінних.

Мережа розпізнає кластери в навчальних даних і розподіляє дані до відповідних кластерів. Якщо в наступному мережа зустрічається з набором даних, несхожим на жодний із відомих зразків, вона відносить його до нового кластеру. Якщо в даних містяться мітки класів, то мережа спроможна вирішувати задачі класифікації.

Шар Кохонена складається з деякої кількості n паралельно діючих лінійних елементів. Всі вони мають однакову кількість входів m та отримають на свої входи один і той самий вектор вхідних сигналів $x=(x_1 \ldots x_m)$. На виході *j*-го лінійного елемента отримуємо сигнал, що розраховується за формулою (1).

$$y_j = w_{j0} \sum_{i=1}^{m} w_{ji} x_i \qquad (1)$$

де w_{ji} — ваговий коефіцієнт *i*-го входу j-го нейрону, w_{j0} — пороговий коефіцієнт.

Після проходження шару лінійних елементів сигнали посилаються на обробку за правилом «переможець забирає все»: серед вихідних сигналів y_j шукається максимальний; його номер $j_{\max} = \arg\max_j \{y_j\}$. Остаточно, на виході сигнал з номером $j_{\max}$ дорівнює одиниці, всі інші — нулю. Якщо максимум одночасно досягається для декількох $j_{\max}$, то приймають всі відповідні сигнали рівними одиниці.

Мережа Кохонена має два прошарки: вхідний і вихідний, що називають самоорганізованою картою. Елементи карти розташовуються в деякому просторі - як правило двовимірному, який передбачено в реалізації запропонованого підходу (рис. 1).[6]

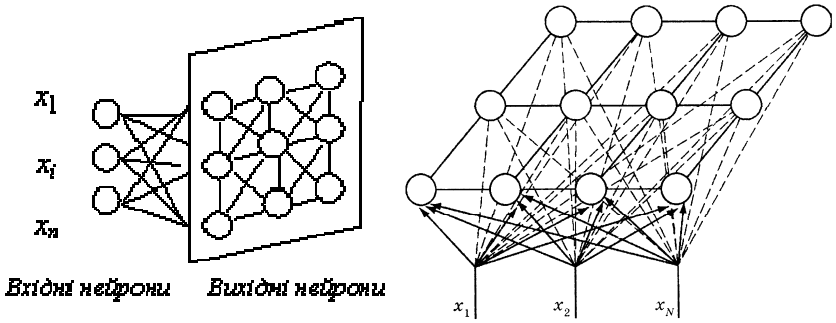

Рисунок 1 – Двомірна структура мережі Кохонена

Навчання розроблюваної мережі відбувається наступним чином. На початку роботи визначається кількість кластерів та їх центри. Після цього деякий вхідний вектор з набору навчальних вибирається і встановлюється на вході нейронної мережі. На цьому етапі відмінності між вхідним вектором та всіма векторами обчислюються за формулою (2).

$$D_{ij} = \left| X^l - W_{ij} \right| = \sqrt{(x_1 - w_{ij1})^2 + ... + (x_n - w_{ijn})^2} \,, \tag{2}$$

де i та j – показники нейронів у вихідному шарі. Після цього нейронна мережа обирає нейрон-переможець з переліку визначених центрів кластерів, тобто такий, щоб його ваговий вектор був схожий на вхідний за формулою (3).

$$D(k_1, k_2) = \min_{i,j} D_{i,j} \,, \tag{3}$$

де k_1 та k_2 – показники нейрона-переможця. Після цього проводиться корекція вагових векторів переможця та сусідніх нейронів. Ближні нейрони до переможця визначаються топологічною функцією сусідства, яка розраховується за формулою 4.

$$h(p,t) = \exp\left(-\frac{p^2}{\sigma^2(t)} \right)\left(1 - \frac{2}{\sigma^2(t)} p^2 \right), \tag{4}$$

де p – відстань до нейрона переможця, яка знаходиться за формулою 5.

$$p = \sqrt{(k_1 - i)^2 + (k_2 - j)^2} \tag{5}$$

де σ – функція, що визначає радіус сусідства. На початку функціонування програмного модулю вона включає весь простір сенсорного поля (сітки), але з часом, значення її зменшується.

У якості функції сусідства була обрана функція «Мексиканський капелюх», що згідно з наведеними результатами експериментів у таблиці 1, забезпечує більшу точність розподілу даних на кластери.

Після обчислення топологічної функції ваги всіх нейронів переобчислюються за формулою (6).

$$W_{ij}(t+1) = W_{ij}(t) + \alpha(t) h(p,t)(X^l(t) - W_{ij}(t)) \tag{6}$$

де $\alpha(t)$ – функція швидкості навчання, яка також змінюється з часом. Якщо нейрон є переможцем, або сусіднім до нього, його вектор ваг оновлюється або залишається незмінним в іншому випадку. На кожному кроці нейронна мережа визначає нейрон, чий ваговий вектор найбільш схожий до вхідного, та коригує його ваги та ваги сусідів, щоб наблизити їх до вхідного вектора (рис. 2).

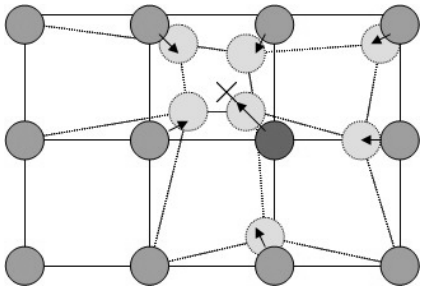

Рисунок 2 – Оновлення нейрону-переможця та його сусідів та підштовхування в сторону вхідного вектору, що на рисунку позначений «×». Суцільні та пунктирні лінії визначають ситуацію до та після оновлення, відповідно

Кожен вхідний вектор з навчальної вибірки представляється нейронній мережі, і навчання триває або деяке фіксоване число циклів, або доки різниця між вхідним і ваговим векторами досягне заданого значення ε. Різниця між сусідніми нейронів зменшується з плином часу, і, отже, вони організуються в групи (кластери), які відповідають одному з класів з навчальної множини. [7]

Для оптимальних розрахунків мережі була також введена процедура попередньої обробки вхідних даних, при якій значення ознак, що утворюють вхідний вектор, приводяться до деякого заданому діапазону. Нормалізація необхідна, оскільки вихідні значення ознак змінюються в достатньо великому діапазоні, тому існує ймовірність некоректної роботи нейронної мережі з такими даними. Так, в одному вхідному векторі можуть міститися значення, що відрізняються один від одного на кілька порядків: стандартний рівень гемоглобіну – 120-180г/л, колірного показнику – 0,85-1,15%. Такий дисбаланс між значеннями показників може викликати нестійкість роботи моделі, погіршити результати навчання і уповільнити його процес. Після нормалізації всі значення вхідних ознак будуть приведені до деякого вузького діапазону, а саме - [0 ... 1] що дозволяє мережі працювати з даними більш коректно.

Для реалізації запропонованого підходу до діагностування була використана нормалізація вхідних даних за формулою (7).

$$y = \frac{(x - x_{\min})(d_2 - d_1)}{x_{\max} - x_{\min}} + d_1 \tag{7}$$

де: x- значення, що підлягає нормалізації, $[d_1, d_2]$- інтервал значень x, $[x_{\min}, x_{\max}]$- інтервал, до якого буде зведене значення x. [8]

Програмна реалізація діагностування на основі нейронної мережі Кохонена

Описаний підхід до діагностування був реалізований за допомогою мови програмування C# (рис.3). Вхідні дані представлені у вигляді стовпців даних, де перший стовпець відповідає імені пацієнту, другий – його статі, стовпці з третього по одинадцятий відповідають показникам загального аналізу крові: гемоглобін, еритроцити, колірний показник, ретикулоцити, тромбоцити, ШОЕ, лейкоцити, паличкоядерні, сегментоядерні, еозинофіли, базофіли, лімфоцити, моноцити.

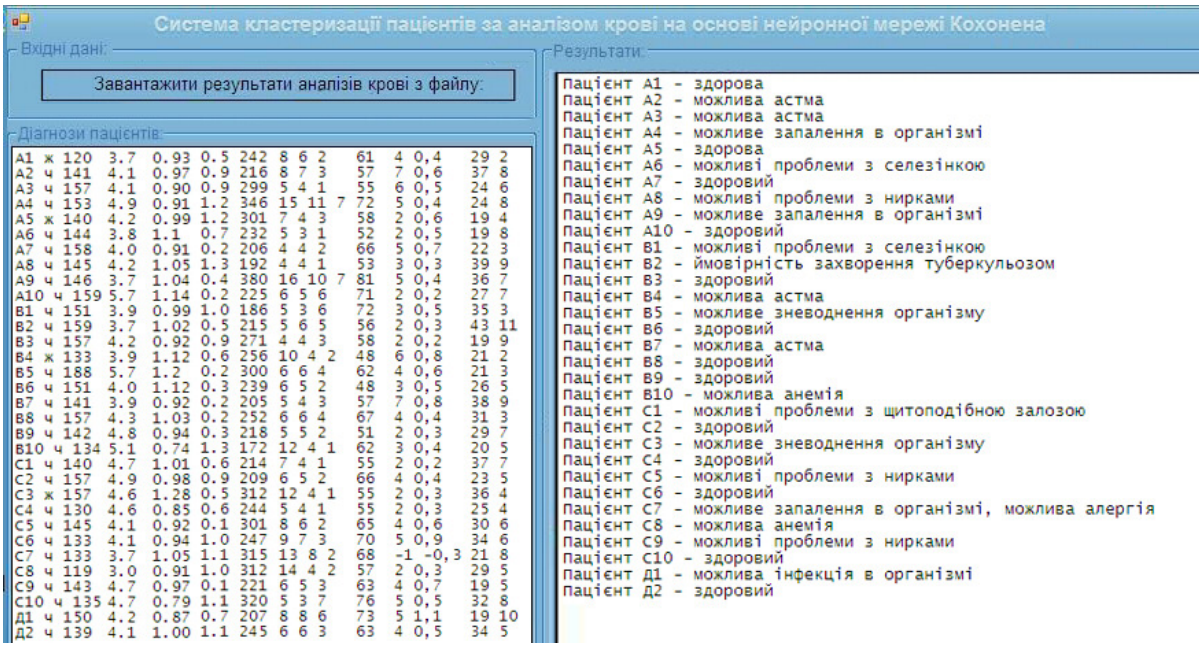

Рисунок 3 – Приклад роботи програми

Результати тестування програми попереднього діагностування пацієнтів

При тестуванні програми перевірялась достовірність її роботи при виборі різних функцій сусідства[6, 9]:

1) Функція Гауса (8);

$$h(p,t) = \exp\left(\frac{p^2}{2\sigma^2(t)}\right) \tag{8}$$

2) Функція «Мексиканський капелюх» (9);

$$h(p,t) = \exp(-\frac{p^2}{\sigma^2(t)})(1-\frac{2}{\sigma^2(t)}p^2) \quad (9)$$

3) Функція «Французький капелюх» (10);

$$h(p) = \begin{cases} 1, & |p| \le \alpha \\ -\frac{1}{3}, & \alpha < |p| \le 3\alpha \\ 0, & |p| < 3\ \alpha \end{cases} \quad (10)$$

Для вибору найефективнішої з представлених функцій сусідства, проведемо порівняння їх точності на основі вибірки у 1000 вхідних векторів, де 800 – навчальна вибірка, 200 – тестова; $\varepsilon=10^{-4}$ (табл.1), для проведення експериментів була створена база даних, основана на безкоштовних загальнодоступних базах даних показників аналізів крові та відповідних хвороб. [10-11].

Таблиця 1 – Функції сусідства та досягнена достовірність діагностики

Функція сусідства	Кількість правильно розпізнаних вхідних векторів (з 200)	Кількість неправильно розпізнаних вхідних векторів (з 200)	Достовірність діагностики
Функція Гауса	166	34	83%
Функція «Мексиканський капелюх»	186	14	93%
Функція «Французький капелюх»	180	20	90%

Таким чином, з результатів експерименту видно, що обрана для подальшої реалізації функція «Мексиканський капелюх» дає найбільшу точність кластеризації.

Проведемо порівняння кластеризації звичайною мережею Кохонена та запропонованою модифікованою мережею з наперед заданими кластерами та їх центрами на основі точності розбиття вибірки у 10, 100, 500 та 1000 вхідних векторів тестової вибірки та $\varepsilon=10^{-4}$ (табл.2).

Таблиця 2 – Досягнена достовірність діагностування звичайною мережею Кохонена та запропонованою модифікованою мережею

Обсяг вибірки	Класична нейронна мережа Кохонена			Модифікована нейронна мережа Кохонена		
	кількість правильно розпізнаних вхідних векторів	кількість неправильно розпізнаних вхідних векторів	Достовірність діагностики	кількість правильно розпізнаних вхідних векторів	кількість неправильно розпізнаних вхідних векторів	Достовірність діагностики
10	7	3	70%	8	2	80%
100	73	27	73%	82	18	82%
500	355	145	71%	394	106	78,8%
1000	716	284	71,6%	803	197	80,3%

У результаті експериментальних досліджень було встановлено, що запропонований метод має більш високі показники точності роботи у порівнянні із класичним, що дає можливість значно підвищити якість діагностування пацієнтів.

Існуюча інтерактивна система діагностування Diagnos.ru, що має найбільшу в світі базу захворювань (близько 240) заснована на методі нечіткої логіки. При цьому точність виявлення захворювання складає в середньому 68% [12], що нижче зазначеного отриманого результату.

Висновки

У роботі запропоновано модифіковану нейронну мережу Кохонена, в якій центри кластерів визначаються заздалегідь згідно з нормальними та відхиленими від норми значеннями показників загального аналізу крові з використанням функції сусідства «Мексиканський капелюх». Це рішення покладено в основу підходу до діагностування пацієнтів на основі загального аналізу крові.

Отже, була вирішена задача знаходження підходу та його реалізації до визначення типу хвороби згідно результатів загального аналізу крові, що характеризується простотою числової реалізації, більш високою точністю роботи у порівнянні з аналогами.

Список літератури

1. Алгоритми діагностування пацієнтів [Електронний ресурс]. Режим доступу: http://www.vidal.by/vracham/Informatsiya-dlya-spetsialistov /Nevrologiya-psihiatriya/Algoritmy-diagnostiki-i-vedeniya-patsientov/ - Назва з екрану

2. Загальний аналіз крові [Електронний ресурс]. Режим доступу: http://ukrhealth.net/zagalnij-analiz-krovi/

3. Обзор методов кластеризации [Електронний ресурс]. Режим доступу: http://www.dialog21.ru/Archive/2001/volume2/2_26.htm - Назва з екрану

4. Симахин В.А., Кашинцев А.В. Кластеризация медицинских данных с помощью нейросетей [Електронний ресурс]. Режим доступу: http://neurocomp.ru/klasterizaciya-medicinskix-dannyx-s-pomoshhyu-nejrosetej/ - Назва з екрану

5. О С. Амосов. Интеллектуальные информационные системы. Нейронные сети и нечеткие системы: Учеб. Пособие. - Комсомольск-на-Амуре: ГОУВПО «КнАГТУ», 2004. -104 с.

6. Руденко О.Г., Бодянский Е.В. Искусственные нейронные сети – Харьков, 2005. – 407с.

7. Дяченко В.А., Михаль О.Ф. Адаптивное параллельное обучение модифицированной самоорганизующейся карты Кохонена [Електронний ресурс]. Режим доступу: http://archive.nbuv.gov.ua/portal/natural/bionint/ 2012_1/Mihal1.pdf

8. Дмитро Парфенович – «Нейронні мережі – від теорії до практики»[Електронний ресурс]. Режим доступу - http://www.mql5.com/ru/articles/497 - Назва з екрану

9.Применение модернизированной вейвлет-функции «Французская шляпа» для аппроксимации продольного распределения магнитного поля в магнитных реверсивных фокусирующих системах. Сравнение её показателей. [Електронний ресурс]. Режим доступу: http://www.moluch.ru/archive/44/5367/ - Назва з екрану.

10. Laboratory Procedure Material. Complete Blood Count using HMX – NHANES 2007-200810. Mayo clinic. Diseases and conditions [Електронний ресурс]. Режим доступу: http://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/basics/tests-diagnosis/con-20020865 - Назва з екрану.

11. CLL Patient Databases [Електронний ресурс]. Режим доступу: https://patientdatabases.org/wp/about

12. Диагноз.ру. Инновационный медицинский центр [Електронний ресурс]. Режим доступу: http://www.diagnos.ru/ - Назва з екрану.

Відомості про авторів

Колесницький Олег Костянтинович - кандидат технічних наук, доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет.

Журавська Юлія Олександрівна – студентка групи 1КН-10 Вінницький національний технічний університет.