БІОЛОГІЧНІ ТА МЕДИЧНІ ПРИЛАДИ І СИСТЕМИ

УДК 004.032.26

О. К. КОЛЕСНИЦЬКИЙ, Ю. О. ЖУРАВСЬКА

Вінницький національний технічний університет, Вінниця

ДОСЛІДЖЕННЯ ВПЛИВУ ТИПУ МЕТРИКИ НА ТОЧНІСТЬ КЛАСТЕРИЗАЦІЇ НЕЙРОННОЮ МЕРЕЖЕЮ КОХОНЕНА У ЗАДАЧІ МЕДИЧНОГО ДІАГНОСТУВАННЯ ЗА АНАЛІЗОМ КРОВІ

Анотація: У даній статті був проведений огляд відомих метрик та експериментально досліджено точність роботи системи медичного діагностування за аналізом крові на основі нейронної мережі Кохонена при використанні різних метрик. У даній задачі метрика використовується для визначення відстані між вектором вхідного набору показників загального аналізу крові та вектором значень центру кластеру, який відповідає певному діагнозу пацієнта. Встановлено, що тип метрики впливає на точність кластеризації. Для запропонованої системи медичного діагностування на основі експериментальних досліджень було встановлено, що найвища точність діагностики забезпечується при використанні зваженої Евклідової відстані.

Ключові слова: нейронна мережа Кохонена, метрика, міра відстані, медична діагностика, кластеризація.

Аннотация: В данной статье был проведён обзор известных метрик и экспериментально исследовано точность работы системы медицинского диагностирования по анализу крови на основе нейронной сети Кохонена при использовании разных метрик. В данной задаче метрика используется для определения расстояния между вектором входящего набора показателей общего анализа крови и вектором значений центра кластера, который соответствует определённому диагнозу пациента. Установлено, что тип метрики влияет на точность кластеризации. Для предложенной системы медицинского диагностирования на основе экспериментальных исследований было установлено, что наивысшая точность диагностики обеспечивается при использовании взвешенного Евклидового расстояния.

Ключевые слова: нейронная сеть Кохонена, метрика, мера расстояния, кластеризация, медицинская диагностика.

Abstract: In current article a review of known metrics was held and the accuracy of the system of medical diagnostics by a blood test based on Kohonen neural network using different metrics was experimentally investigated. In this task the metric is used to determine the distance between the input vector set of general blood test indicators and a vector of cluster center values, which corresponds to a certain diagnosis of the patient. It was established, that the type of metric affects the accuracy of clustering. For the proposed system of medical diagnosis on the basis of experimental studies it was established, that the highest accuracy of diagnostics is provided using the weighted Euclidean distance.

Keywords: Kohonen neural network, metrics, distance measure, clustering, medical diagnostics.

Вступ

Сьогодення характеризується швидким ступенем автоматизації у майже всіх сферах діяльності людини, включаючи медичну галузь. В умовах недосконалості сучасного процесу діагностування за участі сімейного лікаря (терапевта), а саме – довготривалості, багатоетапності та можливої неточності, що пов'язана з відсутністю чітких алгоритмів діагностування, був запропонований метод діагностування з використанням показників загального аналізу крові [1,2]. Даний метод заснований на визначенні 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 допомагає виявити тип захворювання пацієнта та надати рекомендації щодо лікування [1]. Метод був реалізований з використанням кластеризації даних нейронною мережею Кохонена [1,2].

Реалізований метод діагностування вирішує проблеми, притаманні існуючим автоматизованим методам. Більшість систем орієнтовані на визначення конкретного захворювання, що через специфіку медичної галузі, має високу ймовірність виявитись хибними; мають невиправдану складність алгоритмів та їх програмної реалізації; можуть використовуватись лише медичними фахівцями; не можуть бути використані пацієнтами дистанційно без відвідування лікарні. При цьому метод діагностування за загальним аналізом крові з використанням нейронної мережі Кохонена визначає 42 хвороби; заснований на чіткому поетапному алгоритмі кластеризації даних; може бути використаний як медичними працівниками, так і пацієнтами; може використовуватись дистанційно без відвідування лікарні [1,2].

Для удосконалення запропонованого методу медичного діагностування [1,2], а саме – підвищення точності визначення діагнозу, були внесені певні модифікації до математичної моделі нейронної мережі Кохонена. У даній статті розглянуто різні метрики (міри відстані), що у процесі навчання та безпосередньої роботи мережі застосовуються для визначення відстані між вхідним вектором та всіма векторами мережі.

Отже, мета статті – дослідження точності кластеризації нейронної мережі Кохонена при використанні різних метрик у задачі медичного діагностування за аналізом крові та, згідно з результатами експериментів, обрання метрики, яка забезпечує найвищу точність визначення діагнозу у порівнянні з іншими. Сформулюємо задачі, що потрібно виконати для досягнення мети дослідження:

1) Розглянути найчастіше використовувані метрики;

2) Провести експериментальні дослідження, що полягають у підрахуванні кількості правильно визначених діагнозів пацієнтів при використанні певної метрики;
3) Обрати метрику, що забезпечить найбільшу точність діагностування при її використанні, тобто таку метрику, при використанні якої було правильно визначено найбільшу кількість діагнозів.

Процес діагностування на основі нейронної мережі Кохонена

Запропонована система медичного діагностування за аналізом крові заснована на роботі нейронної мережі Кохонена. Дана мережа складається з *m* нейронів, що утворюють прямокутну решітку на площині – шар [3]. Всі вони мають однакову кількість входів *n* та отримують на свої входи один і той самий вектор вхідних сигналів $x = (x_1 \ldots x_n)$. На виході *j*-го лінійного елемента отримуємо сигнал, що розраховується за формулою:

$$y_j = w_{j0} + \sum_{i=1}^{n} w_{ij} x_i , \quad (1)$$

де y_j – вихідний сигнал *j*-го нейрону, w_{ij} – ваговий коефіцієнт *i*-го входу ($i = \overline{1,n}$, n = 13) *j*-го нейрону ($j = \overline{1,m}$, m=43), w_{j0} — пороговий коефіцієнт.

Робота нейронної мережі складається з двох етапів: навчання та безпосередньо діагностування.

Навчання традиційної нейронної мережі Кохонена відбувається наступним чином. На початку роботи відбувається ініціалізація мережі. Початковим ваговим коефіцієнтам присвоюються значення показників загального аналізу крові таким чином, щоб одному діагнозу відповідав один центр кластеру.

Після цього деякий вхідний вектор з набору навчальних векторів вибирається і встановлюється на вході нейронної мережі. На цьому етапі різниця між вхідним вектором та всіма векторами обчислюються за формулою 2, що являє собою формулу Евклідової метрики, яка у подальшому буде розглядатись.

$$D_{lj} = \left| \overline{X}_l - \overline{C}_{ij} \right| = \sqrt{(x_{1l} - c_{1j})^2 + \ldots + (x_{nl} - c_{nj})^2} , \quad (2)$$

де D_{lj} – відстань між l-м вектором вхідних значень та j-м нейроном мережі.

Після проходження шару лінійних елементів, сигнали посилаються на обробку за правилом «переможець забирає все»: серед вихідних сигналів y_i 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.

Після цього нейронна мережа обирає нейрон-переможець з переліку визначених центрів кластерів, тобто такий, щоб його ваговий вектор був схожий на вхідний за формулою:

$$D_{lj_{\max}} = \min_j D_{lj} , \quad (3)$$

де $D_{lj_{\max}}$ – відстань між нейроном-переможцем та l-м вектором вхідних значень.

Після цього проводиться корекція вагових векторів нейрона-переможця та сусідніх нейронів. Функція сусідства визначає міру сусідства нейронів та зміну векторів ваг. Вона повинна поступово уточнювати їх значення, тому функція сусідства задається у вигляді функції швидкості навчання $0 < \alpha(t) < 1$, що монотонно спадає з кожною послідовною ітерацією та визначає наближення значення вагових векторів нейронів до вектору вхідного набору, при цьому із збільшенням кроку зменшується уточнення. У якості функції сусідства використовується функція «Мексиканський капелюх» [3]:

$$h\left(D_{lj_{\max}}, t\right) = \exp\left(-\frac{{D_{lj_{\max}}}^2}{\sigma^2(t)}\right)\left(1 - \frac{2}{\sigma^2(t)} {D_{lj_{\max}}}^2\right), \quad (4)$$

де $h(D_{lj_{\max}},t)$ – топологічна функція сусідства, що залежить від часу навчання *t* та відстані від вхідного вектора до нейрона переможця, $D_{lj_{\max}}$ – відстань від вхідного вектора до нейрона переможця, яка знаходиться за формулою 3, *t* – час навчання, *σ* – функція, що визначає радіус сусідства. На початку функціонування програми вона включає весь простір сенсорного поля (сітки), але з часом, його значення зменшується.

Після обчислення топологічної функції ваги всіх нейронів переобчислюються за формулою:

$$W_{ij}(t+1) = W_{ij}(t) + \alpha(t)h(D_{lj_{\max}},t)(X_l(t) - W_{ij}(t)) \tag{5}$$

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 [3,4].

Для оптимальних розрахунків запропонованої мережі, що використовується в роботі програмного модуля, була також введена процедура попередньої обробки вхідних даних [5], при якій значення ознак, що утворюють вхідний вектор, приводяться до заданого діапазону [0…1]. Використовується нормалізація вхідних даних, виконувана за формулою:

$$x_0' = \frac{(x_0 - x_{\min})}{x_{\max} - x_{\min}} \tag{6}$$

де x_0' – значення після проведеної нормалізації; x_0 - значення, що підлягає нормалізації; [$x_{\min}$, $x_{\max}$]- інтервал вхідних значень *x*.

Діагностування проводиться шляхом подання на вхід мережі набору даних, які представляють собою значення показників загального аналізу крові, у вигляді вектора $x = (x_1 \ldots x_n)$ з тестової вибірки, що порівнюється із векторами даних визначених кластерів. Мережа має певний набір кластерів, їх кількість – 43. Отже, вирішується задача пошуку мінімальної відстані між вхідним вектором X (x_1, ... x_{13}) та одним з визначених центрів кластерів $C_1(c_1 \ldots c_{13})$ $\ldots C_{43}$ $(c_1 \ldots c_{13})$. Де 13 – кількість показників аналізу крові, 43 – кількість кластерів та, відповідно, можливих діагнозів.

Після визначення діагнозу у випадку наявності хвороби відбувається перевірка на випадок існування ще однієї хвороби пацієнта. Для цього відбувається повторне діагностування, але з огляду вилучається центр кластера, що відповідає визначеній хворобі. Якщо пацієнт здоровий або після повторного діагностування виявляється здоровим, діагностування припиняється, та пацієнту надається діагноз «здоровий» або вже визначені діагнози, відповідно.

Експериментальні дослідження з використанням різної метрики

Розглянемо основні використовувані метрики:

1) Евклідова відстань (7) – найбільш загальний тип відстані. Розраховується як геометрична відстань між точками у багатовимірному просторі. У двох- або трьохвимірному випадку – це довжина відрізку прямої, що з'єднує дані точки. Для використання даної метрики потрібно нормувати кожну ознаку, якщо вони вимірюються в різних діапазонах. Застосування Евклідової відстані виправдано якщо властивості (ознаки) об'єкта однорідні за фізичним сенсом та однаково важливі для задачі.

$$D_{lj} = \sqrt{\sum_{i=1}^{n}(x_i - c_i)^2} \tag{7}$$

де D - відстань між *l*-м вхідним вектором X та *j*-м центром кластера C; x_i – значення *i*-го показника X; c_i – значення *i*-го показника C [6-10].

2) Квадрат евклідової відстані (8) – використовується, щоб надати великі ваги більш віддаленим один від одного об'єктам [6-8].

$$D_{lj} = \sum_{i=1}^{n} (x_i - c_i)^2 \tag{8}$$

3) Зважена Евклідова відстань (9) – використовується при завданні довільних ваг для ознак. Застосовується в тих випадках, коли кожній і-ій властивості можливо призначити деяку вагу wi, пропорційно ступеню важливості ознаки у задачі. Визначення ваг, як правило, пов’язано з додатковими дослідженнями, наприклад, організацією опитування експертів та обробкою їхньої думки.

$$D_{lj} = \sqrt{\sum_{i=1}^{n} w_i (x_i - c_i)^2} \tag{9}$$

де w_i – ваговий коефіцієнт i-ї ознаки, зазвичай приймають $0 < w_i < 1$ [6, 9].

Для застосування зваженої Евклідової відстані у задачі медичного діагностування за аналізом крові потрібно визначати ваги для кожного з 13 показників аналізу крові. При проведенні експериментальних досліджень були обрані такі значення вагових коефіцієнтів, щоб показники аналізу крові, які визначають більшу кількість хвороб (мають більшу важливість), отримали більше значення вагового коефіцієнту. Значення вагових коефіцієнтів: 1. гемоглобін – 0,5; 2. еритроцити – 0,7; 3. колірний показник – 0,267; 4. ретикулоцити – 0,53; 5. тромбоцити – 0,567; 6. ШОЕ – 0,9; 7. лейкоцити – 0,83; 8. паличкоядерні – 0,33; 9. сегментоядерні – 0,33; 10. еозинофіли – 0,33; 11. базофіли – 0,53; 12. лімфоцити – 0,4; 13. моноцити – 0,6.

4) Манхеттенська відстань (відстань міських кварталів) (10) – у порівнянні з Евклідовою відстанню вплив окремих великих різниць (вибросів) зменшується, оскільки вони не підносяться у квадрат [6-10].

$$D_{lj} = \sum_{i=1}^{n} \left(|x_i| - |c_i| \right). \tag{10}$$

5) Відстань Махаланобіса (11) – застосовується у випадку залежних компонент вектора спостережень та їх різної залежності у вирішенні задачі [9, 10].

$$D_{lj} = \left(x_i - c_i \right)^T \cdot K^{-1} \cdot \left(x_i - c_i \right), \tag{11}$$

де K^{-1} – коваріаційна матриця генеральної сукупності.

6) Косинус (Cosine) (12) – область значень цієї міри можуть знаходиться у межах -1 та +1. Об’єкти у багатовимірному просторі при цьому розглядаються як вектори [8,10].

$$D_{lj} = \frac{\sum_{i=1}^{n} (x_i, c_i)}{\sqrt{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} c_i^2)}}. \tag{12}$$

7) Відстань Мінковського (13) – визначається як корінь r-го ступеня із суми абсолютних різниць пар значень, взятих у p-му ступені. Звичайно при розрахунку цієї відстані допускається застосування лише квадратного кореня, у той же час як ступінь різниці значень можна обирати у межах від 1 до 4. Якщо даний ступінь взяти рівним 2, то отримаємо Евклідову відстань. Для задачі медичного діагностування за аналізом крові будуть використатись різні значення параметрів r та p [8, 10].

$$D_{lj} = \left(\sum_{i=1}^{n} |x_i - c_i|^p \right)^{1/r}, \tag{13}$$

де r – параметр, що відповідає за прогресивне зважження різниць за окремими координатами; p – параметр, що відповідає за поступове зважження різниць за окремими координатами.
8) Відстань Чебишева (14) – приймає значення найбільшого модуля різниці між значеннями відповідних властивостей (ознак) об'єктів. Ця відстань є корисною, коли потрібно визначити два об'єкти як «різні», якщо вони розрізняються по якій-небудь одній координаті [6-8].

$$D_{lj} = \max\left|x_i - c_i\right|. \tag{14}$$

Якщо обидва параметри – r та p — дорівнюють двом, то ця відстань співпадає з відстанню Евкліда.

При проведенні досліджень була взята навчальна вибірка у розмірі 800 векторів вхідних значень, що являють собою показники загального аналізу крові, тестова вибірка – 200 векторів. Правильно розпізнаними векторами вважаються ті, для яких вхідний вектор значень був присвоєний правильному кластеру, тобто набір показників пацієнта відповідає правильному діагнозу. Результати експериментальних досліджень занесені у табл. 1.

Таблиця 1 – Результати експериментів по визначенню точності діагностування в залежності від використовуваної міри відстані

№ п/п	Назва	Кількість правильно розпізнаних тестових векторів (з 200)	Кількість неправильно розпізнаних тестових векторів (з 200)	Точність діагностики
1	Евклідова відстань	172	28	86%
2	Квадрат евклідової відстані	160	40	80%
3	Зважена евклідова відстань	185	15	92,5%
4	Манхетенська відстань	181	19	90,5%
5	Відстань Махаланобіса	170	30	85%
6	Косинус	174	26	87%
7	Відстань Мінковського			
	при r = 2 p= 3	165	35	82,5%
	при r = 2 p= 4	160	40	80%
	при r = 3 p= 2	170	30	85%
	при r = 3 p= 4	170	30	85%
	при r = 4 p= 2	168	32	84%
	при r = 4 p= 3	171	29	85,5%
8	Відстань Чебишева	162	38	81%

Як видно з представлених результатів експериментів, вибір метрики істотно впливає на точність кластеризації та, як з цього випливає, на точність діагностики системи медичного діагностування за показниками загального аналізу крові на основі нейронної мережі Кохонена. Згідно з отриманими даними, найвища точність діагностування – 92,5% - досягається при використання зваженої Евклідової метрики.

Висновки

У даній статті був проведений огляд відомих метрик та експериментально досліджено точність роботи системи медичного діагностування за аналізом крові на основі нейронної мережі Кохонена при використанні різних метрик. У даній задачі метрика використовується для визначення відстані між вектором вхідних значень, що являє собою набір показників загального аналізу крові, та вектором значень кожного визначеного мережею центру кластеру, який відповідає певному діагнозу пацієнта. Встановлено, що тип метрики впливає на точність кластеризації. Для запропонованої системи медичного діагностування з використанням показників аналізу крові за допомогою нейронної мережі Кохонена на основі експериментальних досліджень було встановлено, що найвища точність забезпечується при використанні зваженої Евклідової відстані (92,5%). Тому в подальшому при реалізації системи медичного діагностування за аналізом крові на основі нейронної мережі Кохонена варто використовувати зважену Евклідову відстань.

Список літератури

1. Колесницький О.К., Журавська Ю.О. Застосування нейронної мережі Кохонена для медичного діагностування пацієнтів за аналізом крові // Інформаційні технології та комп'ютерна інженерія. – 2014 - №1. – с.5-10.

2. Колесницький О.К., Журавська Ю.О. Експериментальні дослідження системи медичного діагностування пацієнтів за аналізом крові на основі нейронної мережі Кохонена // Оптико-електронні інформаційно-енергетичні технології. – 2014 - №2. – с.104-109.

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

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

5. Нейронная сеть Кохонена [Електронний ресурс]. Режим доступу: http://www.machinelearning.ru/wiki/ index.php?title=Нейронная_сеть_Кохонена – Назва з екрану.

6. Мера расстояния [Електронний ресурс]. Режим доступу - http://www.aiportal.ru/articles/autoclassification/measure-distance.html – Назва з екрану.

7. Обзор алгоритмов кластеризации данных [Електронний ресурс]. Режим доступу - http://habrahabr.ru/post/101338/

8. Меры расстояния - переменные с интервальной шкалой [Електронний ресурс]. Режим доступу - http://www.datuapstrade.lv/rus/spss/section_20/5/ – Назва з екрану.

9. Мера близости и расстояние между объектами [Електронний ресурс]. Режим доступу - http://ecocyb.narod.ru/513/MSM/msm2_2.htm – Назва з екрану.

10. Расчет парных расстояний между объектами исходного множества данных [Електронний ресурс] . Режим доступу - http://matlab.exponenta.ru/statist/book2/14/pdist.php – Назва з екрану.

Відомості про авторів

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

Журавська Юлія Олександрівна – студентка групи 1КН-14м, Вінницький національний технічний університет, м. Вінниця.