DOI 10.31891/2307-5732-2024-331-37
УДК 004.9; 004.8

ГАВРИЛЮК МИРОСЛАВ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0001-5259-7564
e-mail: myroslav.a.havryliuk@lpnu.ua
ГОВДИШ НАЗАРІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0009-0007-9237-6679
e-mail: o.govdish@gmail.com

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ВИКОРИСТАННЯ PNN ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ МЕДИЧНОЇ ДІАГНОСТИКИ В УМОВАХ АНАЛІЗУ МАЛИХ ДАНИХ ВИСОКОЇ РОЗМІРНОСТІ

У цій роботі досліджено ефективність використання різних відстаней у алгоритмі роботи ймовірнісної нейронної мережі для задачі виявлення хвороби Паркінсона за біомедичними показниками голосу у випадку малих даних високої розмірності. Було проведено експериментальне моделювання трьох варіантів реалізації PNN із застосуванням наступних відстаней: Чебишова, мангеттенської, Мінковського, косинусової та канберрської. Результати дослідження продемонстрували різні значення F_1-міри при використанні різних відстаней. Зокрема, використання неевклідових метрик забезпечило суттєве підвищення ефективності аналізу короткого набору даних. Отримані результати свідчать про необхідність правильного підбору цього параметру для різних варіантів реалізацій PNN з метою отримання найвищої точності під час розв'язання задач медичного діагностування.

Ключові слова: ймовірнісна нейронна мережа, малі дані, високорозмірні дані, хвороба Паркінсона, класифікація.

HAVRYLIUK MYROSLAV, HOVDYSH NAZARII
Lviv Polytechnic National University

INVESTIGATION OF THE EFFICIENCY OF USING PNN FOR SOLVING THE PROBLEMS OF MEDICAL DIAGNOSTICS IN THE CONDITIONS OF THE ANALYSIS OF SMALL DATA OF HIGH DIMENSION

Parkinson's disease is one of the illnesses that cause certain difficulties at the stage of diagnosis. Recently, there has been a tendency to increase the popularity of the use of artificial intelligence methods as an auxiliary diagnostic tool. A probabilistic neural network (PNN) has become widely used for solving problems in the field of medicine. Despite the rather high efficiency of its use for certain tasks, some aspects of its functioning remain insufficiently researched in practice. Existing scientific works do not pay due attention to the issue of using the optimal distance as a measure of similarity between objects. Calculating the distance between the current data vector and each reference sample vector is the first step in implementing a PNN. The classification accuracy of a neural network of this type depends on its efficiency. In this work, the effectiveness of using different distances in the algorithm of different implementations of a probabilistic neural network for detecting Parkinson's disease based on biomedical voice indicators in the case of small high-dimensional data was investigated. Experimental modeling of three different variants of PNN implementation was carried out using the following distances: Chebyshov, Manhattan, Minkowski, cosine, and Canberra. The results of the study showed different values of the F_1-measure when applying different distances. It was found that the use of the Euclidean metric in the structure of a probabilistic neural network is not always the best option. In particular, the application of non-Euclidean metrics provided a significant increase in accuracy for the analyzed dataset. This indicates the need for correctly selecting this parameter of the probabilistic neural network to obtain the highest accuracy when solving medical diagnosis problems.

Keywords: probabilistic neural network, small data, high-dimensional data, Parkinson's disease, classification.

Постановка проблеми

Проблема ефективної діагностики завжди є актуальною у галузі медицини. Попри постійний розвиток сфери, методики виявлення багатьох хвороб потребують удосконалення.

Хвороба Паркінсона є одним із захворювань, що викликають певні труднощі на етапі встановлення діагнозу. Традиційно, для обстеження пацієнтів із підозрою на цю недугу використовують такі методи як електроенцефалографія головного мозку, комп'ютерна та магнітно-резонансна томографії та ін. Однак, останнім часом спостерігається тенденція до зростання популярності застосування методів штучного інтелекту як допоміжного діагностичного інструменту.

Широкого застосування для розв'язання задач у галузі медицини набула ймовірнісна нейронна мережа (probabilistic neural network – PNN). Попри досить високу ефективність її використання для певних завдань, деякі аспекти її функціонування залишаються недостатньо дослідженими на практиці.

Аналіз останніх джерел

Використання PNN є досить популярним серед дослідників для розв'язання різноманітних задач зі сфери медицини. Наприклад, у роботі [1] застосовано класичну ймовірнісну нейронну мережу для передбачення одного з ключових показників чоловічого репродуктивного здоров'я. У [2] авторами розв'язувалася задача медичного діагностування на основі використання трьох різних наборів медичних даних. Проте, як показано у цій роботі, використання класичного варіанту реалізації PNN не завжди забезпечує задовільні результати. Через це, у [3] запропоновано новий метод обчислення виходів цієї нейронної мережі, який забезпечує формування повної системи подій із набору ймовірностей належності до кожного із класів задачі. Такий підхід забезпечив суттєве підвищення ефективності діагностування. Незважаючи на це, інтелектуальний аналіз незбалансованих медичних даних із використанням цього варіанту реалізації PNN не показує результатів необхідної точності. У [2] автори запропонували інший варіант

формування вихідного сигналу роботи цієї штучної нейронної мережі. Він враховує як можливість формування повної системи подій, так і можливий нерівномірний розподіл представників кожного класу у наборі даних. Метод показав високу точність роботи під час аналізу низки коротких набору медичних даних.

В роботі [4] автори розробили модель зваженої ймовірнісної нейронної мережі. Вона використовує процедуру аналізу чутливості для обчислення ваг. Ефективність такої мережі продемонстровано для кількох задачах (у тому числі із галузі медицини). У роботі [5] було розроблено гібридний метод на основі комбінованого використання PNN та SVM для визначення якості сплавів, які використовуються при виготовленні біомедичних імплантатів. Усі ці підходи орієнтовано на аналіз коротких наборів даних. Проте у галузі медицини виникають задачі діагностування у випадку аналізу даних великих обсягів. В цьому випадку, PNN стає великою та повільною, що зменшує ефективність її використання. Саме тому, низка наявних досліджень щодо використання PNN в задачах медичної діагностики у випадку аналізу великих наборів даних сфокусована на мінімізації структури PNN для економії часу та обчислювальних ресурсів. Зокрема, робота [6] пропонує два евристичні методи зменшення одного із шарів ймовірнісної нейронної мережі. Один з них використовує процедуру кластеризації K-means, інший – метод опорних векторів (SVM).

Загалом, усі вищезгадані наукові роботи не надають належної уваги питанням використання оптимальної відстані як міри подібності між об'єктами. Саме обчислення відстані між поточним вектором даних та кожним вектором опорної вибірки є першим кроком реалізації PNN. Від його ефективності залежить точність класифікації нейронною мережею цього типу. Таким чином, ця задача є важливою та актуальною і вимагає проведення додаткових експериментальних досліджень.

Метою роботи є дослідження ефективності використання різних відстаней для трьох варіантів реалізації ймовірнісної нейронної мережі в задачах медичного діагностування у випадку аналізу малих даних високої розмірності.

Виклад основного матеріалу

Розглянемо функціонування ймовірнісної нейронної мережі для розв'язання задачі бінарної класифікації. Припустимо, у опорній вибірці об'єктів є k екземплярів класу 1 та m екземплярів класу 2. Позначимо j-й атрибут i-го об'єкту першого класу $X_{i,j}^1$, другого класу – $X_{i,j}^2$. Задача моделі – класифікація вхідного об'єкта X. Тобто необхідно визначити набір ймовірностей належності об'єкта Xдо кожного із визначених класів задачі.

Алгоритмічна реалізація PNN передбачає послідовне виконання наступних кроків:

1. Обчислюються евклідові відстані між вхідним об'єктом та кожним із екземплярів опорної вибірки:

$$R_i^1 = \sqrt{\sum_{j=1}^{n}(X_{i,j}^1 - X_j)^2}, R_i^2 = \sqrt{\sum_{j=1}^{n}(X_{i,j}^2 - X_j)^2} \quad (1)$$

2. На їх основі обчислюються Гауссові відстані:

$$D_i^{1,2} = exp(-\frac{(R_i^{1,2})^2}{\sigma^2}) \quad (2)$$

3. Ймовірність того, що вхідний об'єкт належить до першого класу, обчислюється наступним чином (Алгоритм 1):

$$P_1 = \frac{\sum_{i=1}^{k} D_i^1}{k} \quad (3)$$

4. Аналогічно для другого класу:

$$P_2 = \frac{\sum_{i=1}^{k} D_i^2}{m} \quad (4)$$

5. У підсумку, ймовірнісна нейронна мережа прогнозує клас, до якого належить вхідний об'єкт, за наступним правилом [7]:

$$y^{pred} = \begin{Bmatrix} 0, якщо\, max\{P_c\} = P_1 \\ 1, якщо\, max\{P_c\} = P_2 \end{Bmatrix}, c = 1,2 \quad (5)$$

Існує також спосіб обчислення ймовірностей належності вхідного об'єкта до певного класу, які формують повну систему подій (Алгоритм 2) [3]. Проте, як вже згадувалося вище, цей метод не враховує можливої незбалансованості екземплярів кожного класу у наборі даних:

$$P_1 = \frac{\sum_{i=1}^{k} D_i^1}{\sum_{i=1}^{k} D_i^1 + \sum_{j=1}^{m} D_j^2}, P_2 = \frac{\sum_{j=1}^{m} D_j^2}{\sum_{i=1}^{k} D_i^1 + \sum_{j=1}^{m} D_j^2} \quad (6)$$

Алгоритм 3 [2] натомість комбінує принципи формування вихідного сигналу Алгоритмів 1 та 2:

$$P_1 = \frac{\sum_{i=1}^{k} \frac{D_i^1}{k}}{\sum_{i=1}^{k} \frac{D_i^1}{k} + \sum_{j=1}^{m} \frac{D_j^2}{m}}, P_2 = \frac{\sum_{j=1}^{m} \frac{D_j^2}{m}}{\sum_{i=1}^{k} \frac{D_i^1}{k} + \sum_{j=1}^{m} \frac{D_j^2}{m}} \quad (7)$$

Як видно із (1), класичний алгоритм реалізації ймовірнісної нейронної мережі передбачає обчислення евклідової відстані між об'єктами. Проте ця відстань може не забезпечити достатньої точності. Саме тому в рамках цієї роботи було проведено дослідження та експериментальне моделювання із застосуванням інших відомих типів метрик [8]: Чебишова, мангеттенської, Мінковського, косинусової та канберрської.

Відстань Чебишова визначає дистанцію між об'єктами як максимальну різницю між їх відповідними

координатами:

$$R_i^1 = max_{j=1}^n \left|X_{i,j}^1 - X_j\right|, R_i^2 = max_{j=1}^n \left|X_{i,j}^2 - X_j\right| \quad (8)$$

Мангеттенська відстань є сумою модулів різниць відповідних властивостей об'єктів:

$$R_i^1 = \sum_{j=1}^n \left|X_{i,j}^1 - X_j\right|, R_i^2 = \sum_{j=1}^n \left|X_{i,j}^2 - X_j\right| \quad (9)$$

Відстань Мінковського є узагальненням мангеттенської відстані на довільному евклідовому просторі (експерименти були проведені для m=1,5):

$$R_i^1 = \left(\sum_{j=1}^n \left|X_{i,j}^1 - X_j\right|^m\right)^{\frac{1}{m}}, R_i^2 = \left(\sum_{j=1}^n \left|X_{i,j}^2 - X_j\right|^m\right)^{\frac{1}{m}} \quad (10)$$

Косинусову відстань обчислюється наступним чином:

$$R_i^1 = 1 - \frac{\sum_{j=1}^n X_{i,j}^1 \cdot X_j}{\left\|X_{i,j}^1\right\| \cdot \|X_j\|}, R_i^2 = 1 - \frac{\sum_{j=1}^n X_{i,j}^2 \cdot X_j}{\left\|X_{i,j}^2\right\| \cdot \|X_j\|} \quad (11)$$

Канберрська відстань:

$$R_i^1 = \sum_{j=1}^n \frac{\left|X_{i,j}^1 - X_j\right|}{\left|X_{i,j}^1\right| + |X_j|}, R_i^2 = \sum_{j=1}^n \frac{\left|X_{i,j}^2 - X_j\right|}{\left|X_{i,j}^2\right| + |X_j|} \quad (12)$$

Дослідження проводилось на основі набору даних, що містить об'єкти, атрибутами яких є біомедичні показники голосу як людей із хворобою Паркінсона (147 записів), так і здорових (48). 31 людини (23 мають хворобу Паркінсона). Цільовим атрибутом є категорійна змінна "статус": значення "1" означає наявність хвороби у відповідної особи, тоді як "0" – її відсутність (задача бінарної класифікації). Цей набір є прикладом малих даних високої розмірності через обмежену кількість об'єктів (195) та велику кількість атрибутів (22). У роботі використано кілька показників оцінювання ефективності роботи PNN. Проте, досліджуваний набір даних є незбалансованим, тому однією із найважливіших метрик є F_1-міра.

Набір даних було попередньо нормалізовано за допомогою робастного масштабування, яке доцільно використовувати у разі великого впливу викидів. Експерименти проводились із використанням засобів мови програмування Python та її бібліотек. Для оптимізації параметра σ було застосовано метод диференціальної еволюції [9]. Результати моделювання трьох різних варіантів реалізації PNN (Алгоритми 1, 2 та 3) при використанні різних відстаней на першому кроці алгоритмів подано в таблиці 1.

Таблиця 1

Результати моделювання

Алгоритм	Відстань	Точність	Влучність	Повнота	F_1-міра	Оптимальне значення σ
Алгоритм 1	Евклідова	0,897	0,938	0,938	0,938	0,079
	Мангеттенська	0,923	0,968	0,938	0,952	2,667
	Косинусова	0,795	0,9	0,844	0,871	0,008
	Чебишова	0,949	0,969	0,969	0,969	0,037
	Мінковського	0,923	0,968	0,938	0,952	0,376
	Канберрська	0,769	0,96	0,75	0,842	0,312
Алгоритм 2	Евклідова	0,923	0,914	1	0,955	0,568
	Мангеттенська	0,949	0,941	1	0,97	2,860
	Косинусова	0,897	0,889	1	0,941	0,450
	Чебишова	0,949	0,969	0,969	0,969	0,037
	Мінковського	0,949	0,941	1	0,97	1,011
	Канберрська	0,795	0,962	0,781	0,862	2,309
Алгоритм 3	Евклідова	0,821	0,821	1	0,901	0,363
	Мангеттенська	0,795	0,816	0,969	0,886	1,061
	Косинусова	0,795	0,833	0,938	0,882	0,049
	Чебишова	0,821	0,821	1	0,901	0,170
	Мінковського	0,821	0,821	1	0,901	0,579
	Канберрська	0,385	0,722	0,406	0,52	0,571

Як видно із таблиці 1, у деяких випадках використання інших метрик (а не евклідової) забезпечує суттєво вище значення F_1-міри. Наприклад, для Алгоритму 1 найвищу ефективність аналізу вищеописаного набору даних досягнуто при застосуванні відстані Чебишова. При цьому, використання інших двох метрик (мангеттенської та Мінковського) також показує кращі результати, ніж використання евклідової.

Для Алгоритму 2 найкращі значення F_1-міри продемонструвало застосування двох відстаней –

Мінковського та мангеттенської (цей показник є максимальним у рамках всього дослідження). Результати також показали, що використання метрики Чебишова у цьому варіанті реалізації PNN є більш ефективним, ніж використання евклідової метрики. Для Алгоритму 3 найкращі результати показало застосування наступних відстаней: евклідової, Чебишова та Мінковського.

Таким чином, застосування евклідової метрики у структурі ймовірнісної нейронної мережі не завжди є оптимальним варіантом. Результати дослідження демонструють, що використання інших відстаней може бути виправданим і доцільним, тому необхідно проводити додаткові дослідження під час використання PNN для правильного підбору цього параметру.

Висновки

У цій роботі було досліджено ефективність використання різних відстаней у алгоритмі роботи ймовірнісної нейронної мережі для задачі виявлення хвороби Паркінсона за біомедичними показниками голосу у випадку малих даних високої розмірності. Було проведено експериментальне моделювання трьох варіантів реалізації PNN із застосуванням наступних відстаней: Чебишова, мангеттенської, Мінковського, косинусової та канберрської.

Результати дослідження продемонстрували різні значення F_1-міри при використанні різних відстаней. Зокрема, використання неевклідових метрик забезпечило суттєве підвищення ефективності аналізу короткого набору даних. Отримані результати свідчать про необхідність правильного підбору цього параметру для різних варіантів реалізацій PNN з метою отримання найвищої точності під час розв'язання задач медичного діагностування.

Література

1. Izonin I., Tkachenko R., Ryvak L., Zub K., Rashkevych M., Pavlyuk O. Addressing Medical Diagnostics Issues: Essential Aspects of the PNN-based Approach. Proceedings of the 3rd International Conference on Informatics & Data-Driven Medicine, Växjö, Sweden, November 19–21, 2020. P. 209–218.

2. Izonin I., Tkachenko R., Greguš M. I-PNN: an improved probabilistic neural network for binary classification of imbalanced medical data. Database and Expert Systems Applications. DEXA 2022. Lecture Notes in Computer Science, vol 13427. Springer, Cham. P. 147–157. DOI: https://doi.org/10.1007/978-3-031-12426-6_12.

3. Duriagina Z.A., Tkachenko R.O., Trostianchyn A.M., Lemishka I.A., Kovalchuk A.M., Kulyk V.V., Kovbasyuk T.M. Determination of the best microstructure and titanium alloy powders properties using neural network. Journal of Achievements in Materials and Manufacturing Engineering. 2018. 87 (1). P. 25–31. DOI: https://doi.org/10.5604/01.3001.0012.0736.

4. Kusy M., Kowalski P.A. Weighted probabilistic neural network. Information Sciences. 2018. 430. P. 65–76. DOI: https://doi.org/10.1016/j.ins.2017.11.036.

5. Izonin I., Tkachenko R., Gregus M., Duriagina Z., Shakhovska N. PNN-SVM approach of Ti-based powder's properties evaluation for biomedical implants production. Computers, Materials & Continua. 2022. 71 (3). P. 5933–5947. DOI: https://doi.org/10.32604/cmc.2022.022582.

6. Kusy M., Kluska J. Assessment of prediction ability for reduced probabilistic neural network in data classification problem. Soft Computing. 2017. 21. P. 199–212. DOI: https://doi.org/10.1007/s00500-016-2382-9.

7. Havryliuk M., Hovdysh N., Tolstyak Y., Chopyak V., Kustra N. Investigation of PNN Optimization Methods to Improve Classification Performance in Transplantation Medicine. Proceedings of the 6th International Conference on Informatics & Data-Driven Medicine, Bratislava, Slovakia, November 17-19, 2023. P. 338–345.

8. Basystiuk O., Melnykova N. Multimodal approaches for natural language processing in medical data. Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine. Lyon, France, November 18–20, 2022. P. 246–252.

9. Basystiuk O., Melnykova N., Rybchak Z. Machine Learning Methods and Tools for Facial Recognition Based on Multimodal Approach. Proceedings of the Modern Machine Learning Technologies and Data Science Workshop (MoMLeT&DS 2023), Lviv, Ukraine, June 3, 2023. P. 161–170.