DOI: 10.36910/6775-2524-0560-2020-40-25
УДК: 681.3.093:044.3
Яременко Вадим Сергійович, аспірант
https://orcid.org/0000-0001-8557-6938
Материнська Софія Василівна, студент
https://orcid.org/0000-0002-5746-4899
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ВИКОРИСТАННЯ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ ВИЗНАЧЕННЯ НАЯВНОСТІ СЕРДЦЕВО-СУДИННИХ ХВОРОБ ТА ЗАХВОРЮВАНЬ ПЕЧІНКИ ПРИ МАЛИХ НАБОРАХ ДАНИХ

Яременко В.С., Материнська С.В. Використання штучних нейронних мереж для визначення наявності сердцево-судинних хвороб та захворювань печінки при малих наборах даних. В даній роботі проведено аналіз ефективності застосування штучних нейронних мереж для вирішення задачі класифікації для невеликих наборів медичних даних із сфери діагностування. Для дослідження було обрано два набори даних: дані про серцево-судинні захворювання та про хвороби печінки. Отримані результати було порівняно з результатами точності для стандартних методів машинного навчання, що використовуються в задачах класифікації Для проведення дослідження було обрано модель багатошарового перцептрона. Програмним засобом для реалізації став Python, що надає можливість використовувати допоміжні бібліотеки при роботі з методами машинного навчання.
Ключові слова:штучна нейронна мережа, класифікація даних, медичні дані, Python, багатошаровий перцептрон.

Яременко В. С., Материнская С. В. Использование искусственных нейронных сетей для определения наличия сердечно-сосудистых болезней и заболеваний печени при малых наборов данных. В данной работе проведен анализ эффективности применения искусственных нейронных сетей для решения задачи классификации для небольших наборов медицинских данных из сферы диагностики. Для исследования были выбраны два набора данных: данные о сердечно-сосудистых заболеваниях и о болезнях печени.Полученные результаты были сопоставлены с результатами точности для стандартных методов машинного обучения, используемых в задачах классификации Для проведения исследования была выбрана модель многослойного перцептрона. Программным средством для реализации стал Python, что позволяет использовать вспомогательные библиотеки при работе с методами машинного обучения.
Ключевые слова: искусственная нейронная сеть, классификация данных, медицинские данные, Python, многослойный персептрон.

Yaremenko V. S., Materynska S. V. Use of artificial neural networks to determine the presence of cardiovascular diseases and liver diseases in small data sets. In this paper was conducted the analysis of the effectiveness of the use of artificial neural networks to solve the problem of classification for small sets of medical data in the field of diagnosis. Two data sets were selected for the study: data on cardiovascular diseases and on liver diseases. The obtained results were compared with the accuracy results for standard machine learning methods used in classification problems. A multilayer perceptron model was chosen for the study. Python has become a software for implementation, which provides the ability to use auxiliary libraries when working with machine learning methods.
Keywords: artificial neural network, data classification, medical data, Python, multilayer perceptron.

Постановка проблеми. У відкритому доступі знаходяться набори медичних даних, що можна використовувати для вирішення задач класифікації. Специфіка наборів полягає також у їх невеликому розмірі. Проте, поруч із розвитком програмних та апаратних засобів, дану проблему – обробки медичних даних вирішуючи задачу класифікації – всеще вирішують в основному за допомогою стандартних методів машинного навчання з вчителем, отримуючи при цьому не надто ефективні результати. Використання штучних нейронних мереж для задач класифікації медичних даних допоможе виявити, чи можна покращити наявні результати та наскільки ефективним є застосування нейронних мереж для невеликих наборів даних при вирішенні задачі класифікації. Для дослідження обрано два невеликі набори медичних даних: про серцево-судинні захворювання та про хвороби печінки.

Аналіз останніх досліджень і публікацій. Для оцінки стану області з точки зору досліджень, проведених на медичних даних, що вирішують задачу класифікації, було проаналізовано наступні роботи.

Дослідження [2], що аналізувало 453 документи, які представляли методи машинного навчання для визначення хронічних захворювань показали, що дослідники в основному використовували SVM, модель к-найближчого сусіда, метод наївного Баєса(баєсівські мережі), логістичну регресію, random forests та дерева рішень.

Підходи до машинного навчання зі вчителем застосовують у найбільшій кількості досліджень із інтеграцією легкого та простого прогнозного моделювання. Тим не менше, в дослідженнях досить ефективно застосовують також глибоке навчання, а саме – штучні нейронні мережі.

У роботі [12] використовували для побудови моделей такі методи, як SVM, "randomforest", дерево рішень, "extratreeclassifier", "gradientboosting" та модель штучної нейронної мережі.

У дослідженні [4] для прогнозування хвороб використовували наступні методи: к-найближчих сусідів та згорткової нейронної мережі. І отримали значно кращі результати при використанні штучної нейронної мережі, ніж з використанням методу к-найближчих сусідів.

У статті [1]для визначення хвороби Паркінсона використовували багатошаровий перцептрон, Баєсівські мережі, "randomforest" та "boostedlogicalregression".

Робота [11] полягала в порівнянні ефективності дерева рішень, методу наївного Баєса, к-найближчих сусідів та застосування штучної нейронної мережі для прогнозування захворювань. В результаті найбільш ефективним виявився метод дерева рішень.

Дослідження [5] використовувало методи опорних векторів, наївного Баєса та дерева рішень. За його результатами дійшли висновку, що метод опорних векторів значно краще оцінює дані, ніж два інші, і є досить ефективним при невеликих об'ємах вхідних даних.

Із дослідження [9], що порівнює роботу алгоритмів дерева рішень, логістичної моделі та "random forest", можна дійти висновку найкраще на конкретному наборі проявила себе логістична модель із найменшою похибкою на тестовому наборі даних.В роботі [6] використовують методи наївного Баєса та опорних векторів.

На основі проаналізованих публікацій можна зробити висновки, що, як і припускалось, в основному проблему класифікації медичних даних вирішують стандартними методами машинного навчання. Короткі результати найчастіше застосовуваних методів представлені у таблиці 1.

Таблиця 1

Частота застосування найпопулярніших методів в проаналізованих публікаціях

Скорочення	Повна назва методу	Кількість застосувань в проаналізованих публікаціях
SVM	Метод опорних векторів	5
NB	Метод наївного Баєса	5
DT	Дерево рішень	5
LR	Логістична регресія	4
RF	"Random forest"	4
ANN	Штучна нейронна мережа	4
KNN	Метод к-найближчих сусідів	3

Для більш детального огляду, зокрема аналізу точності моделей, отриманих різними методами, розглянемо детально деякі роботи.

Що стосується ***даних про серцево-судинні захворювання***, то в дослідженні [9] було застосовано методи логістичної регресії, "randomforest" та дерево рішень з результатами точності – 76%, 80% та 83% відповідно, а в роботі [7] використовували модель ШНМ, а саме багатошаровго перцептрона зі зворотнім поширенням і при розподілі тренувального і тестового наборів як 60:40, отримали точність моделі на тестових даних в 80.17%.

Для набору ***даних про захворювання печінки*** в дослідженні [3]використовували методи к-найближчих сусідів, дерева прийняття рішень, "randomforest" та наївного Баєса, а найкращий отриманий результат 74.2% методом "randomforest". Також результати [8] показують 68% точність

методом логістичної регресії.Згідно з результатами [10] методи SVM, наївного баєса та дерево рішень дали точність 71%, 56% та 66% відповідно.

Окрім проведеного аналізу важливим є коректне проведення кожного з етапів моделювання, які схематично зображені на рис. 1.

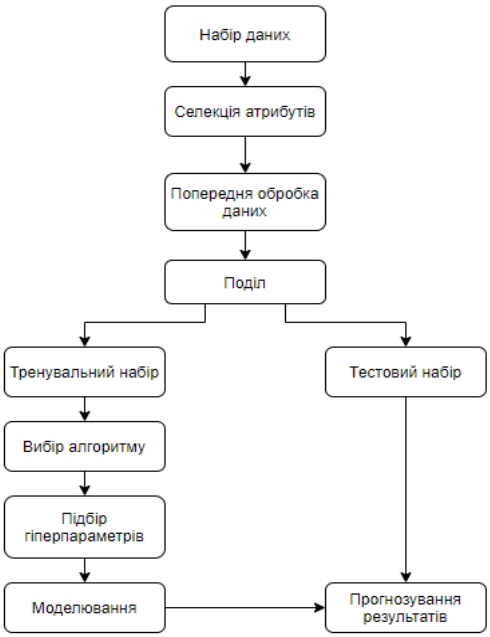

Рис. 1. Алгоритм моделювання за допомогою методів машинного навчання

Якщо коротко описати процес створення моделі, то він буде мати такий вигляд:

1) Збір/вибір набору даних
2) Селекція інформативних атрибутів
3) Обробка набору даних
4) Розподіл на тренувальний та тестовий набори
5) Вибір функції
6) Підбір гіперпараметрів для вибраної функції
7) Навчання моделі на тренувальному наборі
8) Перевірка ефективності на тестовому наборі
9) Застосування отриманої моделі для класифікації нових даних

Зазвичай пункти 5-8 повторюються ітеративно з метою вибору найкращого варіанту для вибраного набору даних.

Мета статті. Метою є побудувати моделі методами машинного навчання з вчителем, що використовуються для вирішення задачі класифікації. Окрім того поставлено ціль покращити точність результатів класифікації медичних даних за допомогою використання штучних нейронних мереж як альтернативного варіанту та порівняти з попередньо отриманими результатами. Для реалізації штучної нейронної мережі використовувати саме модель багатошарового перцептрона зі зворотним поширенням.

Результати дослідження. Для моделювання було застосовано мову Pythonта її допоміжні бібліотеки, зокрема для машинного навчання та візуалізації, а саме sklearn, numpy, pandas, seaborn та matplotlib.

Вхідні дані включають два набори: набір даних про серцево-судинні захворювання, що містить 303 елементи і 13 атрибутів і позначені наявність та відсутність проблем з серцево-судинною системою та набір даних про хвороби печінки, що містить 583 елементи та 10 атрибутів з аналогічними відмітками про наявність проблем з печінкою.

В процесі моделювання для отримання ефективного результату важливо провести всі етапи попередньої обробки даних та самого процесу, оскільки вони безпосередньо впливають на точність отриманих моделей.

У випадку, якщо набір включає велику кількість рис(атрибутів), необхідно провести вибір інформативних за допомогою статистичного аналізу, це дасть змогу побудувати точну модель. В нашому випадку кількість атрибутів була невеликою, тож цей етап був необов'язковим.

Подальші етапи передбачають:

1) перевірку на відсутність значень, що знижує точність моделі: їх можна заміняти на значення за замовчуванням, середні з вибірки, або видаляти рядок запису
2) нормалізацію даних, що покращує ефективність – зведення всіх значень до діапазону 0-1.

Наступний крок – розподіл на тренувальний, валідаційний та тестовий набори даних, адже після побудови моделі на тренувальних даних ми хочемо перевірити її ефективність на тестових даних, які до цього для неї були невідомі таким чином перевіряючи її упередженість. Валідаційний набір використовується для підбору гіперпараметрів, але у випадку невеликих об'ємів вхідних даних, такий поділ здійснювати не доцільно.

Обрані дані було розподілено у співвідношенні 70%:30% відповідно на тренувальний та тестовий набори. Підбір гіперпараметрів здійснено на основі тестового набору, оскільки через невеликий об'єм вхідних даних додаткове виділення валідаційного набору може негативно вплинути на результати. Крім того при розподілі варто враховувати початкове співвідношення класів, що дасть можливість точніше оцінити ефективність моделі.

Після розподілу вхідних даних необхідно вибрати алгоритм, за яким буде навчатись модель. Було обрано такі алгоритми, як метод опорних векторів, логістична регресія, дерево рішень, метод к-найближчих сусідів, "random forest", метод наївного Баєса та штучну нейронну мережу.

Для проведення моделювання деякі з алгоритмів потребують визначення гіперпараметрів. Ми повинні вказати їх значення вручну – алгоритм навчання не вивчає гіперпараметри(вільні параметри) з навчальних даних на відміну від фактичних параметрів моделі. До прикладу проводимо підбір гіперпараметру к для методу к-найближчих сусідів як зображено на рис. 2. Ми підбираємо параметр, при якому алгоритм показує найкращий результат на тестовому наборі. В нашому випадку це к=18.

Оскільки набори даних невеликі за мірками машинного навчання, ми можемо динамічно підбирати гіперпараметри для побудови нейромереж, такі як кількість нейронів в прихованих шарах, функції активації та зворотного поширення. Для навчання було використано моделі багатошарового перцептрона з одним та двома прихованими шарами. У таблиці 2 представлені результати отриманих гіперпараметрів для найефективніших з моделей для кожного із наборів даних.

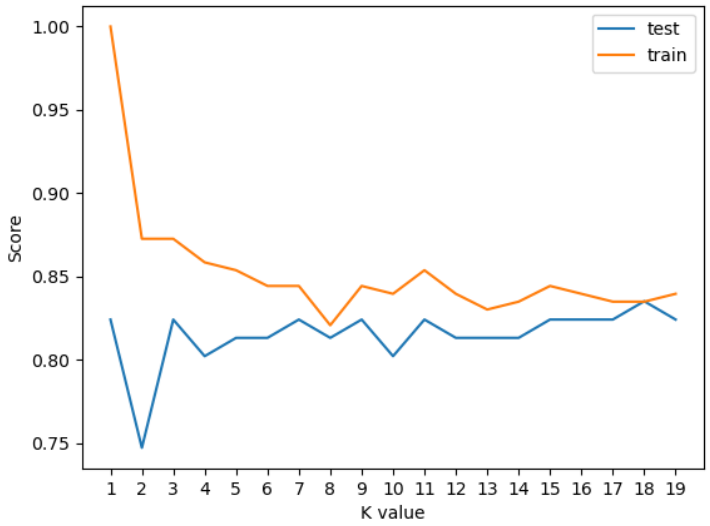

Рис. 2. Підбір параметра к для методу к-найближчих сусідів

Таблиця 2

Підібрані гіперпараметри для багатошарового перцетптрона

Гіперпараметри багатошарового перцептрона	**Серцево-судинні захворювання**	**Хвороби печінки**
Кількість прихованих шарів і нейронів в них	4 – в 1-му шарі 7 – в 2-му шарі	2 – в 1-му шарі 8 – в 2-му шарі
Функція активації	Relu	Tanh
Функція зворотного поширення	lbfgs	lbfgs

Після підбору гіперпараметрів було проведене навчання моделі на тренувальних даних та оцінка точності на тестових. Отримані результати показали, що модель з двома прихованими шарами дає точніші результати, а більша кількість шарів (3 і більше) не покращує точність, а тільки збільшує час навчання. Результати точності для перцептронів з одним та двома прихованими шарами представлені в таблиці 3.

Таблиця 3

Точність моделі багатошарового перцептрона у %

	1 прихований шар	2 приховані шари
Серцево-судинні захворювання	90.11	91.21
Хвороби печінки	74.86	75.43

Провівши навчання і оцінивши точність для кожного із алгоритмів, отримано наступні результати, що відображено в таблиці 4. Можна зрозуміти, що як і було припущено, застосування штучної нейронної мережі дало точніші результати, ніж стандартні методи машинного навчання, що застосовуються для класифікації.

Таблиця 4

Результати точності для кожної з моделей

Повна назва методу	**Отримана точність, %**	
	Серцево-судинні захворювання	Хвороби печінки
Метод опорних векторів	86,83	72
Метод наївного Баєса	85,71	50,29
Дерево рішень	70,33	62,29
Логістична регресія	85,71	73,14
"Random forest"	82,42	68
Штучна нейронна мережа	91,21	75,43
Метод к-найближчих сусідів	83,52	71,43

Висновки. В ході проведеного дослідження було оброблено два набори даних [13][14]. Було виявлено, що модель багатошарового перцептрона зі зворотним поширенням із двома прихованими шарами показала ефективніші результати не тільки в порівнянні із попередньо застосованими методами машинного навчання, а і з результатами попередньо проаналізованих сторонніх досліджень, що можна зрозуміти з таблиці 5.

Таблиця 5

Порівняння отриманих результатів для обох наборів вхідних даних

	Нейронна мережа	Стандартні методи машинного навчання		Результати отримані зі сторонніх досліджень	
Серцево-судинні захворювання	91.21	86.83%	SVM	83%	DT
Хвороби печінки	75.43	73.14%	LR	74.2%	RF

Таким чином показники ефективності моделі були підвищені за допомогою нейронних мереж на 4.38% – від 86.83% для методу опорних векторів до 91.21% для перцептрона з 2 прихованими шарами – для даних про серцево-судинні захворювання і на 2.29% – від 73.14% для логістичної регресії до 75.43% для аналогічної штучної нейронної мережі. А в порівнянні з точністю проаналізованих досліджень – точність моделі для першого набору збільшилась на 8,2%, а для другого – на 1,2%.

Отримані результати доводять ефективність застосування штучних нейронних мереж для вирішення задач класифікації, зокрема що стосується невеликих наборів вхідних даних, що продемонстровано на практиці.

Перспективи подальших досліджень. Тема ефективності класифікації малих об'ємів вхідних даних, залишається актуальною, тож подальші дослідження на основі даного можуть проводитись у різних площинах. Вони можуть включати в себе аналіз інших моделей нейронних мереж та доцільності їх застосування для вирішення даної проблеми, яка стосується саме задачі класифікації медичних даних. Окрім того, можливе проведення узагальнення для аналізу наборів, що подібні невеликими розмірами для визначення продуктивного підходу для покращення точності моделей, побудованих на їх основі.

References.

1. Challa, K. N. R., Pagolu, V. S., Panda, G., & Majhi, B. (2016, October). An improved approach for prediction of Parkinson's disease using machine learning techniques. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)* (pp. 1446-1451). IEEE.
2. Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. *Journal of Personalized Medicine*, *10*(2), 21. URL: https://www.mdpi.com/2075-4426/10/2/21/htm (Last accessed: 17.03.2020).
3. Muthuselvan, S., Rajapraksh, S., Somasundaram, K., & Karthik, K. (2018). *Classification of Liver Patient Dataset Using Machine Learning Algorithms. International Journal of Engineering & Technology, 7(3.34), 323.* doi:10.14419/ijet.v7i3.34.19217
4. Dahiwade, D., Patle, G., & Meshram, E. (2019, March). Designing Disease Prediction Model Using Machine Learning Approach. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1211-1215). IEEE.
5. Kanchan, B. D., & Kishor, M. M. (2016, December). Study of machine learning algorithms for special disease prediction using principal of component analysis. In *2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC)* (pp. 5-10). IEEE.
6. Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems (IJAIS)*, *3*(7), 25-30.
7. Khemphila, A., & Boonjing, V. (2011, August). Heart disease classification using neural network and feature selection. In *2011 21st International Conference on Systems Engineering* (pp. 406-409). IEEE.
8. Indian Liver Patients - Logistic Predictions. URL: https://rpubs.com/bpoulin-CUNY/338004. (Last accessed: 17.03.2020).
9. Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137.
10. Kumar, K., Sreedevi, M., & Padmanabha Reddy, Y. C. A. (2018). Survey on machine learning algorithms for liver disease diagnosis and prediction. *Int. J. Eng. Technol.(IJET (UAE))*, *7*(18), 99-102.
11. Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. *International Journal of Engineering*, *1*(8), 1-4.
12. Wroge, T. J., Özkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., & Ghomi, R. H. (2018, December). Parkinson's disease diagnosis using machine learning and voice. In *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1-7). IEEE.
13. Indian Liver Patient Dataset. 2012. URL: https://www.kaggle.com/jeevannagaraj/indian-liver-patient-dataset (Last accessed: 17.03.2020)
14. Heart Disease Data Set / A.Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano. 1988. URL : https://archive.ics.uci.edu/ml/datasets/Heart+Disease. (Last accessed: 17.03.2020)