УДК 004.622 DOI https://doi.org/10.32782/1995-0519.2022.2.7

КЛАСИФІКАЦІЯ ДАНИХ ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ

Тамара Савчук
кандидат технічних наук,
професор кафедри комп'ютерних наук
Вінницький національний технічний університет, Хмельницьке шосе, 95, Вінниця, Вінницька область, 21000, savchuk_t@vntu.edu.ua
ORCID: 0000-0002-0061-6206

Олександр Пупко
аспірант кафедри комп'ютерних наук
Вінницький національний технічний університет, Хмельницьке шосе, 95, Вінниця, Вінницька область, 21000, salexpro@vntu.edu.ua
ORCID: 0000-0003-3905-704X

Обробка даних у сучасній інформаційній ері має велике значення, оскільки з розвитком технологій збільшується й обсяг інформації, яку важливо обробляти, класифікувати та аналізувати. Для швидкого доступу до даних потрібні різні підходи до їх класифікації, що є актуальним та потребує наукових досліджень у цьому напрямі.

Завдяки стрімкому розвитку нейронних мереж слід звернути увагу на їх імплементацію у структурування та аналіз інформації, що дозволить розробляти ефективні методи, алгоритми та підходи до обробки текстових та числових масивів.

У цій роботі проведено аналіз різних алгоритмів класифікації даних, описано актуальність проведення наукового дослідження в цьому напрямі. Для аналізу було вибрано класифікатори на основі методів статистики, нейронних мереж та засновані на машинному навчанні.

У статті наведено переваги та недоліки кожного з вибраних підходів до обробки інформації, складено порівняльну таблицю та описано особливості їх застосування.

На основі цього наукового дослідження з метою підвищення ефективності класифікації було вибрано класифікатор даних з використанням нейронних мереж, описано алгоритм його побудови та застосування, а також наведено переваги над іншими підходами до аналізу інформації, обґрунтовано доцільність використання нейронних мереж у обробці даних. Сформовано таблицю з результатами класифікації даних з використанням нейронних мереж та статистичного методу.

Надано пропозиції для підвищення ефективності вибраного методу для обробки даних та зроблено відповідні висновки.

Ключові слова: дані, класифікація, нейронні мережі, інформаційні технології, класифікатор, штучний інтелект, обробка даних, машинне навчання, навчання, аналіз, нейрони, аналіз даних, інформація, алгоритми, статистичні методи, статистика, числові дані, текстові дані, текст, структуризація даних, технології, очищення даних, фільтрація даних, масив даних, обсяги даних, моделі даних, ефективність, дослідження, огляд, комп'ютерні науки, класифікація та аналіз, бінарна класифікація, представлення даних, кодування.

Обробка даних у сучасному світі потребує використання більш точних і швидких алгоритмів, позаяк кількість оброблюваної інформації збільшується з кожним днем. Самі інформаційні технології вимагають складної підготовки, великих первісних витрат і наукомісткої техніки. Їх введення повинне починатися зі створення математичного забезпечення, формування інформаційних потоків у системах підготовки фахівців [1].

Такі великі обсяги даних потрібно класифікувати та структурувати задля швидкого доступу до них, тому часто для цього використовують інтелектуальні технології.

Класифікація є однією з найважливіших задач інтелектуального аналізу даних, яка вирішується за допомогою аналітичних моделей, названих класифікаторами [1]. Затребуваність класифікації даних зумовлена порівняльною простотою алгоритмів та методами її реалізації та високою інтерпретованістю результатів порівняно з іншими технологіями аналізу даних.

Натепер розроблено велику кількість різних видів класифікаторів, для побудови яких використовуються як статистичні методи, так і методи машинного навчання [1].

Необхідність використання в аналізі даних великого числа різноманітних методів класифіка-

ції зумовлена тим, що вирішувані за її допомогою завдання можуть мати свої особливості, пов'язані, наприклад, з числом класів (бінарна класифікація або з декількома класами) або з поданням вихідних даних – їх обсягом, розмірністю та якістю, що вимагає вибору адекватного класифікатора. Тому вибір класифікатора, відповідного особливостям розв'язуваної задачі аналізу, є важливим фактором отримання правильного рішення [1].

Класифікатори, засновані на машинному навчанні, називаються метричними [2]. Як правило, вони простіші в реалізації та використанні, ніж параметричні, а їхні результати зручніші для інтерпретації і розуміння. Але при цьому метричні класифікатори є наближеними моделями і забезпечують рішення тільки в обмеженому числі практично значущих випадків, як правило, неточних та неєдиних, що зменшує їх ефективність у вирішенні задачі класифікації [2].

Класифікатори на основі методів статистики мають достатню математичну обґрунтованість, проте вони є складними у використанні і вимагають знання ймовірного розподілу вихідних даних і оцінки їх параметрів, мають фіксовану структуру моделі, оцінюють тільки ймовірність приналежності об'єкта класу, що також звужує спектр використання цього класифікатора в задачах класифікації даних [2].

Альтернативою двох вищезгаданих підходів до класифікації даних є використання нейронних мереж.

Нейронні мережі є непараметричними моделями, що не вимагають припущень про ймовірнісний розподіл даних, але при цьому і не використовують характеристики відстаней. Це робить їх універсальними класифікаторами, дозволяючи отримувати результати навіть у випадках, коли параметричні і метричні класифікатори не забезпечують прийнятного рішення [2]. Недоліком при цьому буде необхідність у потужному апаратному забезпеченні, що має забезпечити прийнятну швидкодію.

Однак з метою усунення виявлених недоліків параметричних і метричних класифікаторів, використовуючи спеціальні способи представлення даних, можна адаптувати нейронні мережі для роботи з категоріальними даними, тобто задавати на вході і формувати на виході категоріальні значення. Для цього категоріальні ознаки відповідним чином кодуються за допомогою числових значень [2].

Переваги та недоліки кожного виду класифікатора наведені в таблиці 1.

Тоді процес формування класифікаційної моделі даних на основі нейронних мереж включатиме такі етапи, як:

– попередня обробка та очищення даних;

– вибір кількості використовуваних ознак;

– визначення кількості зв'язків між нейронами.

Попередня обробка та очищення даних – відбір ознак, які є значущими з точки зору відмінності класів. Об'єкти предметної сфери можуть описуватися великим числом ознак, але не всі вони дозволяють коректно розрізняти об'єкти різних класів. Наприклад, якщо об'єкти різних класів мають приблизно однаковий розмір, то використання «габаритних» ознак не має сенсу. Використання ознак, значення яких є випадковими і такими, що не відображають закономірностей розподілу об'єктів по класах, також є недоцільним.

Крім цього, важливу роль відіграє вибір кількості використовуваних ознак. З одного боку, чим більше ознак застосовується у разі побудови класифікатора, тим більше інформації використовується для поділу класів. Але при цьому зростають обчислювальні витрати і вимоги до розміру нейронної мережі. Зниження кількості використовуваних ознак погіршують роздільність класів. Наприклад, може скластися ситуація, коли у об'єктів різних класів виявляться однакові значення ознак і може виникнути розбіжність.

Для побудови ефективно працюючого класифікатора важливо правильно визначити кількість зв'язків між нейронами, які налаштовуються в процесі навчання і обробляють вхідні дані під час її роботи. З одного боку, якщо ваг у мережі буде мало, то вона не зможе реалізовувати складні функції поділу класів. З іншого боку, збільшення числа зв'язків призводить до зростання інформаційної ємності моделі (ваги працюють як елементи пам'яті) [3].

У результаті, коли число зв'язків у мережі перевищить число прикладів навчальної вибірки, мережа буде не апроксимувати залежності в даних, а просто запам'ятає і буде відтворювати комбінації «вхід–вихід» з навчальних прикладів. Такий класифікатор буде ефективно працювати на навчальних вибірках даних і видавати довільні відповіді на нових, які не використовувались у процесі навчання. Іншими словами, мережа не отримає узагальнюючу здатність, і використовувати на практиці побудований на її основі класифікатор буде недоцільно.

Для визначення кількості зв'язків між нейронами застосовують два підходи – конструктивний

Таблиця 1

Переваги та недоліки вибраних видів класифікаторів

Вид класифікатора	Переваги	Недоліки
Класифікатори, що використовують методи статистики	Достатня математична обґрунтованість	Складні у використанні і вимагають знання ймовірного розподілу вихідних даних і оцінки його параметрів, мають фіксовану структуру моделі, оцінюють тільки ймовірність приналежності об'єкта класу.
Класифікатори з використанням нейронних мереж	Нейронна мережа є самонавчальною моделлю, робота якої практично не вимагає втручання користувача. Нейромережі є універсальними апроксиматорами, що застосовуються до будь-якої безперервної функції з прийнятною точністю. Такі класифікатори базуються на нелінійних моделях, що дозволяють ефективно вирішувати завдання класифікації навіть за відсутності лінійної роздільності класів.	Конфігурація мережі, що апроксимує функцію поділу класів у просторі ознак, заздалегідь невідома. Тому доводиться підбирати її експериментально або використовувати досвід аналогічних рішень.
Класифікатори з використанням машинного навчання	Не вимагають оцінки параметрів розподілу вихідних даних, а міра збіжності в них формалізується за допомогою функції відстані (зазвичай евклідова). Такі класифікатори називаються метричними. Вони простіші в реалізації і використанні, ніж параметричні, а їх результати зручніші для інтерпретації і розуміння.	Забезпечують рішення тільки в обмеженому числі практично значущих випадків, можуть дати неточне або не єдине рішення.

і деструктивний [3]. Перший полягає в тому, що спочатку ініціалізується мережа мінімального розміру, і потім її поступово збільшують до забезпечення необхідної точності. При цьому після кожного збільшення мережі її заново навчають. Також є так званий метод каскадної кореляції, за якого після закінчення кожної епохи навчання відбувається коригування архітектури мережі з метою мінімізації помилки.

У разі деструктивного підходу спочатку ініціалізується мережа завищеного розміру, потім з неї видаляються нейрони і зв'язки, які мають найменший вплив на точність класифікатора, з урахуванням того, що число прикладів у навчальній множині має бути більшим кількості ваг мережі, що налаштовуються.

Слід зазначити, що для побудови класифікатора використовується нейронна мережа прямого поширення, в якій сигнали поширюються в одному напрямку, починаючи від вхідного шару нейронів через приховані шари до вихідного шару, і на вихідних нейронах отримується результат опрацювання сигналу. Такий тип нейронної мережі підходить для задачі класифікації найбільше, оскільки найкраще працює з обробкою даних для їх подальшої структуризації.

Побудова класифікатора даних на основі нейронної мережі включає такі кроки.

1. Підготовка даних.

1.1. Розробити базу даних із прикладів, характерних для такого завдання.

1.2. Розбити всю сукупність даних на дві множини: навчальну і тестову (можлива розбивка на 3 множини: навчальну, тестову і валідаційну).

2. Попередня обробка даних.

2.1. Провести відбір ознак, які є значущими з точки зору завдання класифікації.

2.2. Виконати трансформацію і, за необхідності, очищення даних (нормалізацію, видалення аномалій). В результаті бажано отримати лінійно розподілений по класах простір множини прикладів [4].

2.3. Вибрати систему кодування вихідних значень.

3. Конструювання, навчання і оцінка якості мережі.

3.1. Вибрати топологію мережі: кількість шарів, число нейронів у шарах і т. д.

3.2. Вибрати активаційну функцію нейронів (наприклад, логістичну, гіпертангенс і ін.).

3.3. Вибрати алгоритм навчання мережі.

3.4. Оцінити якість роботи мережі на основі валідації множини навчальних даних та оцінки

результату оптимізації архітектури (зменшення ваг, проріджування простору ознак).

3.5. Вибрати варіант мережі, який забезпечує найкращу здатність до узагальнення і оцінити якість роботи по тестовій вибірці.

4. Використання і діагностика.

4.1. З'ясувати ступінь впливу різних чинників на прийняте рішення (евристичний підхід).

4.2. Переконатися, що мережа забезпечує необхідну точність класифікації (кількість неправильно розпізнаних прикладів невелика).

4.3. За необхідності повернутися до кроку 2, змінивши спосіб представлення прикладів або змінивши базу даних [5].

Схема алгоритму побудови класифікатора даних з використанням нейронної мережі представлена на рисунку 1.

З метою дослідження ефективності використання нейронних мереж для класифікації даних було проведено низку експериментальних досліджень, кожне з яких полягало у визначенні точності та швидкодії залежно від кількості даних та набору характеристик [6], в одному випадку у разі використання нейронних мереж, а в другому – з використанням статистичного методу. Отримані результати представлено у таблиці 2.

Переваги використання нейронних мереж порівняно зі статистичним методом класифікації даних представлено в таблиці 3.

Таким чином, запропонований підхід до класифікації даних дозволить підвищити точність з 50% до 90% за рахунок використання нейронної мережі та оптимально підібраних до неї характеристик.

Рис. 1. Схема алгоритму побудови класифікатора даних з використанням нейронної мережі

Таблиця 2.

Результати класифікації даних з використанням нейронних мереж та статистичного методу

Кількість даних	Набір характе-ристик	Точність вибраного підходу		Швидкість роботи програми, сек.	
		Нейронні мережі	Статисти-чний метод	Нейронні мережі	Статисти-чний метод
50 000	2	55%	50%	380	440
50 000	4	75%	55%	450	510
50 000	8	90%	60%	800	860
100 000	8	88%	65%	1400	1600

Таблиця 3

Переваги використання нейронних мереж порівняно зі статистичними методами класифікації даних

Класифікатори, що використовують методи статистики	Класифікатори з використанням нейронних мереж
Складні у використанні	Прості в розробці, не вимагають втручання користувача, універсальні апроксиматори
Фіксована структура	Нелінійна структура
Низька швидкодія	Висока швидкодія
Точність 50–60% (оскільки оцінюється тільки ймовірність приналежності того чи іншого об’єкта відповідному класу)	Точність від 50% до 90% залежно від кількості заданих характеристик та вхідних даних.

ЛІТЕРАТУРА

1. Столяр Р. Інформаційні технології у сучасному світі. *Sophus науковий клуб* : вебсайт. URL: http://sophus.at.ua/publ/2013_12_19_20_kampodilsk/sekcija_7_2013_12_19_20/informacijni_tekhnologiji_v_suchasnomu_sviti/49-1-0-863. (дата звернення: 17.08.2022).

2. Штучна нейронна мережа. *Вікіпедія: вільна енциклопедія.* URL: https://uk.wikipedia.org/wiki/Штучна_нейронна_мережа (дата звернення: 17.08.2022).

3. Damas M., Salmeron M., Diaz A., Ortega J., Prieto A., Olivares G. Genetic algorithms and neuro-dynamic programming: application to water supply networks. *Proceedings of 2000 Congress on Evolutionary Computation*. 2000. Vol. 1. P. 7–14 .

4. Bozinovski S. A self-learning system using secondary reinforcement. *Proceedings of the Sixth European Meeting on Cybernetics and Systems Research*. 1982. P. 397–402.

5. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. *Data Mining* : вебсайт. URL: https://ranalytics.github.io/data-mining/076-NN.html (дата звернення: 17.08.2022).

6. Bottaci L. Artificial Neural Networks Applied to Outcome Prediction for Colorectal Cancer Patients in Separate Institutions. *The Lancet*. 1997. P. 469–472.

CLASSIFICATION OF DATA BY NEURAL NETWORKS

Tamara Savchuk
Candidate of Technical Sciences, Professor of the Department of Computer Science
Vinnytsia National Technical University, Khmelnytske highway, 95, Vinnytsia, Vinnytsia Region, 21000, savchuk_t@vntu.edu.ua
ORCID: 0000-0002-0061-6206

Oleksandr Pupko
Postgraduate Student of the Department of Computer Science
Vinnytsia National Technical University, Khmelnytske highway, 95, Vinnytsia, Vinnytsia Region, 21000, salexpro@vntu.edu.ua
ORCID: 0000-0003-3905-704X

Data processing is important in the modern information age. As the development of technology increases the amount of information that is important to process, classify and analyze. Rapid access to data requires different approaches to their classification, which is relevant and requires research in this area.

Due to the rapid development of neural networks, attention should be paid to their implementation in the structuring and analysis of information, which will develop effective methods, algorithms, and approaches to the processing of text and numerical arrays.

In the given work the review of various algorithms for the classification of data is carried out, the urgency of carrying out scientific research in this direction is described. Classifiers based on statistical methods, neural networks, and based on machine learning were selected for analysis.

The article presents the advantages and disadvantages each of the selected approaches to information processing, compiles a comparative table and describes the features of their application.

Based on this research to increase the efficiency of classification, the data classifier using neural networks was chosen, the algorithm of its construction and application is described, as well as the advantages over other approaches to information analysis, the feasibility of using neural networks in the data processing. A table with the results of data classification using neural networks and statistical method is formed.

Suggestions for improving the efficiency of the chosen method for data processing are presented and the corresponding conclusions are made.

Key words: data, classification, neural networks, information technology, classifier, artificial intelligence, data processing, machine learning, learning, analysis, neurons, data analysis, information, algorithms, statistical methods, statistics, numerical data, text data, text, data structuring, technology, data cleaning, data filtering, data array, data volumes, data models, efficiency, research, review, computer science, classification and analysis, binary classification, data representation, coding.

REFERENCES

1. Stolyar, R. (2013). Informatsiini tekhnolohii v suchasnomu sviti [Information technologies in the modern world]. *Sophus*. Retrieved from: http://sophus.at.ua/publ/2013_12_19_20_kampodilsk/sekcija_7_2013_12_19_20/informacijni_tekhnologiji_v_suchasnomu_sviti/49-1-0-863 (Last accessed: 17.08.2022) [in Ukrainian].

2. Wikimedia Foundation. (2022, May 15). Artificial Neural Network. Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Artificial_neural_network (Last accessed: 17.08.2022).

3. Damas, M., Salmeron, M., Diaz, A., Ortega, J., Prieto, A., & Olivares, G. (n.d.). Genetic algorithms and neuro-dynamic programming: Application to water supply networks. *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512)*. Retrieved from: https://doi.org/10.1109/cec.2000.870269.

4. Bozinovski, S. (1982). A self-learning system using secondary reinforcement. Cybernetics and Systems Research, 397–402.

5. Klassifikacija, regressija i drugie algoritmy Data Mining s ispol'zovaniem R [Classification, Regression and Other Data Mining Algorithms Using R]. (n.d.). Retrieved from: https://ranalytics.github.io/data-mining/076-NN.html (Last accessed: 17.08.2022).

6. Bottaci, L., Drew, P.J., Hartley, J.E., Hadfield, M.B., Farouk, R., Lee, P.W., MacIntyre, I., Duthie, G.S., & Monson, J. (1997). Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. The Lancet, 350, 469–472.

Стаття надійшла 15.05.2022