УДК 004.852 DOI: https://doi.org/10.30837/IYF.CVSAMM.2024.087

ОСОБЛИВОСТІ СУЧАСНИХ МЕТОДІВ МАШИННОГО НАВЧАННЯ ЩОДО ВИРІШЕННЯ ЗАДАЧІ ПРОГНОЗУВАННЯ ДАНИХ

Меженна І.Д.
Науковий керівник – к.т.н., доц. Творошенко І.С.
Харківський національний університет радіоелектроніки, каф. ІНФ,
м. Харків, Україна
e-mail: iryna.mezhenna@nure.ua

This work is devoted to the process of analyzing and identifying the features of machine learning methods in solving data prediction problems. The relevance of the forecasting task in the modern world is described, and the importance of choosing the right method is mentioned. The following tools were considered: quantile regression forests, recurrent neural networks and support vector method. The areas in which the tools listed are commonly used are described. Problems that can be encountered when working with it. Some recommendations are provided that may help to improve the quality of the models.

Прогнозування даних – це процес використання статистичних даних для встановлення тенденцій та патернів з метою передбачення майбутніх значень або подій. У сучасному світі прогнозування є актуальним і важливим напрямком з численними застосуваннями в різних галузях, таких як фінанси, медицина, транспорт, маркетинг тощо. Для вирішення задач цього класу існує багато інструментів, які варіюються від стандартних статистичних пакетів до спеціалізованих програм, але одним із провідних підходів, особливо в контексті складних завдань або великих обсягів даних, є машинне навчання.

Проте, зважаючи на постійні дослідження та розвиток в галузі штучного інтелекту, інколи буває складно підібрати той метод, який буде доцільно використовувати для вирішення конкретної задачі [1–4]. Метою даної роботи є проведення аналізу та виділення особливостей сучасних методів машинного навчання щодо вирішення задачі прогнозування даних.

На сьогоднішній день існує широкий спектр методів машинного навчання, які використовуються для прогнозування даних. Розглянемо детальніше найпопулярніші з них.

Ліс з квантильною регресією – це метод машинного навчання, який поєднує в собі властивості випадкового лісу і квантильної регресії. Випадковий ліс – це ансамбль дерев рішень, які навчаються на випадкових підвибірках навчальних даних. Квантильна регресія, з іншого боку, спрямована на оцінку квантилів розподілу вихідної змінної. Кожне дерево будує свій власний прогноз і використовується як частина схеми пропозицій для створення кінцевого прогнозу. Підсумкове прогнозування ґрунтується не на якомусь

окремому дереві, а лісі в цілому. Ліс з квантильною регресією став популярним методом в областях, де важлива точність прогнозів в усіх частинах розподілу вихідної змінної, і де потрібна стійкість до аномальних значень або шуму в даних. Виділимо основні особливості методу, які потрібно враховувати під час роботи з ним.

Модель лісу слід навчати на великій кількості об'єктів, не менше кількох сотен, для отримання найкращого результату. Цей інструмент не підійде для дуже маленьких наборів даних.

Зазвичай, кількість дерев дорівнює 100, але це число не керується даними. Його необхідно збільшувати при складних відносинах між незалежними змінними, розміром набору та змінною для прогнозування. Рекомендується збільшити кількість дерев у лісі не менше, ніж в 3 рази, хоча б до 500 дерев, щоб найкращим чином оцінити продуктивність моделі.

Час виконання дуже чутливий до кількості змінних, що використовуються у кожному дереві. При застосуванні меншої кількості змінних зменшується ймовірність перенавчання, проте тоді може виникнути необхідність збільшення дерев (для поліпшення продуктивності моделі).

Ліс з квантильною регресією може погано спрацьовувати при спробі прогнозувати незалежні змінні, що знаходяться поза діапазоном незалежних змінних, які використовувалися для навчання. Якщо прогнозоване значення на основі незалежних змінних є набагато вищим або нижчим за діапазон початкового навчального набору, модель оцінюватиме значення як таке, що перебуває поруч із найвищим або найнижчим значенням у початковому наборі даних.

Ще одним потужним інструментом для прогнозування даних є рекурентні нейронні мережі. Ця модель глибокого навчання добре підходить для завдань обробки послідовностей даних і може бути використана для прогнозування часових рядів з різних сфер життя. Розглянемо й інші особливості цього інструменту.

Рекурентна нейронна мережа обробляє дані послідовно, і це обмежує її ефективність при обробці великої кількості текстів. Типова модель може успішно проводити аналіз кількох речень, але для створення резюме за цілою сторінкою тексту їй знадобляться величезні обчислювальні потужності, великий обсяг пам'яті та багато часу.

Також модель може зіштовхнутися з такими проблемами як вибухаючий або зникаючий градієнт під час опрацювання довгих послідовностей. Перший варіант означає ситуацію, в якій градієнт експоненціально зростає аж до повної втрати стабільності. Коли градієнт стає нескінченно великим, модель втрачає передбачуваність і починають виникати проблеми з продуктивністю, такі як перенавчання. Зникнення градієнта означає, що мережа втратила можливість ефективно навчатися за запропонованими даними, що призводить до недонавчання.

Наступний, не менш ефективний та популярний, інструмент – метод опорних векторів для регресії. Це варіант методу опорних векторів, який використовується для вирішення задач прогнозування даних у вигляді неперервної змінної. Основне завдання – знайти лінію чи поверхню, що найкраще підходить до точок даних, при цьому допускаючи деяке відхилення, тобто помилку.

Метод працює ефективніше у високорозмірних просторах: чим більша кількість ознак, тим точніше передбачення. Проте варто пам'ятати про збільшення при цьому обчислювальних витрат.

Також інструмент має декілька гіперпараметрів, таких як параметр м'якої межі, параметр ядра та параметр епсилон. Неправильне налаштування цих гіперпараметрів може призвести до перенавчання або недонавчання моделі. Варто експериментувати для досягнення кращих результатів.

Більш того, метод опорних векторів потребує нормалізації даних. Цей процес може включати в себе обробку пропущених значень, згладжування або видалення викидів, а також масштабування даних до узгодженого діапазону. Попередня обробка допомагає підвищити точність моделі.

Згідно описаного вище, можна зробити висновок, що для вирішення задач прогнозування даних існує багато різноманітних інструментів, з певними перевагами та недоліками. Висвітлені результати дослідження підвищать розуміння особливостей сучасних методів. Це допоможе уникнути помилок при застосуванні та полегшить процес вибору оптимального підходу для вирішення конкретної задачі прогнозування даних.

Список використаних джерел:

1. Gorokhovatskyi V., Tvoroshenko I., and Yakovleva O. (2024) Transforming image descriptions as a set of descriptors to construct classification features, Indonesian Journal of Electrical Engineering and Computer Science, 33 (1), pp. 113–125.

2. Gorokhovatskyi V., Tvoroshenko I. (2023) Identification of visual objects by the search request. Int. scientific symp. «Intelligent Solutions-S». Computational intelligence. Decision making theory: proceedings of the international symposium, September 28, 2023, Kyiv-Uzhorod, Ukraine, 25–27.

3. Творошенко, И. С. (2010). Анализ процессов принятия решений в интеллектуальных системах. Системи обробки інформації, (2), 248–253.

4. Tvoroshenko I., Gorokhovatskyi V., Kobylin O., and Tvoroshenko A. (2023) Application of deep learning methods for recognizing and classifying culinary dishes in images, International Journal of Academic and Applied Research, 7(9), pp. 57–70.