АЛГОРИТМ КЛАСИФІКАЦІЇ ДАНИХ ЗА ДОПОМОГОЮ НЕЙРОННОЇ МЕРЕЖІ

К. т. н. П. С. Сафронов[1], к. т. н. О. Ф. Бондаренко[2]

[1]Одеський національний політехнічний університет,
[2]Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Україна
p.s.safronov@gmail.com; bondarenkoaf@gmail.com

Запропоновано алгоритм класифікації даних за допомогою нейронної мережі, який включає три основні процедури: перетворення формату даних, навчання нейронної мережі та тестування нейронної мережі. Перша виконує перетворення масиву вхідних даних до придатного для використання формату. Навчання нейронної мережі виконано на основі методу зворотного поширення помилки. Процедура тестування дозволяє отримати значення вихідного нейрона для даних, які були введені користувачем, порівняти його із значеннями вихідних нейронів, отриманих у процесі навчання нейронної мережі, та прийняти рішення на основі цих значень щодо приналежності об'єкта, який класифікується.

Ключові слова: інтелектуальний аналіз даних, класифікація даних, машинне навчання, нейронна мережа, метод зворотного поширення помилки.

Класифікація — розбиття множини об'єктів або спостережень на заздалегідь задані групи, так звані класи — є однією з найважливіших задач інтелектуального аналізу даних, що застосовується, наприклад, у маркетингу при оцінці кредитоспроможності позичальників, визначенні лояльності клієнтів, розпізнаванні образів, медичній діагностиці та в багатьох інших сферах. Бажаним варіантом вирішення задачі класифікації є зведення складних задач до бінарної класифікації, де число класів обмежене двома. Проте такий підхід не завжди є можливим [1].

Найбільш поширеним способом подання вхідних даних є спосіб, при якому зразок представляється вектором. Компоненти цього вектора являють собою різні характеристики зразка, які впливають на прийняття рішення про те, до якого класу можна віднести даний зразок [2].

Оцінка точності класифікації може проводитися за допомогою крос-перевірки. Точність класифікації тестової множини порівнюється з точністю класифікації навчальної множини. Якщо класифікація тестової множини дає приблизно такі самі результати за точністю, як і класифікація навчальної множини, вважається, що дана модель пройшла крос-перевірку. Поділ на навчальну і тестову множини здійснюється шляхом ділення вибірки у певній пропорції, наприклад: навчальна множина — дві третини даних, тестова — одна третина [3, 4].

Запропонований алгоритм класифікації даних включає три основні процедури: перетворення формату даних, навчання нейронної мережі та тестування нейронної мережі.

Структурна організація вхідних даних зазвичай відповідає текстовому файлу з набором даних, який можна розділити на навчальну і тестову вибірки. Алгоритм процедури перетворення формату даних складається з наступних кроків:

— розділення вмісту файлу вхідних значень на окремі рядки;

— проходження по кожному рядку і розділення кожного рядка на окремі символи;

— формування масиву вихідних значень (якщо кількість нейронів дорівнює 3, він буде містити прихований шар з 3 нейронів; якщо кількість нейронів не дорівнює 3, масив не буде містити прихованого шару);

— перетворення масиву вихідних значень до придатного для використання формату.

Навчання нейронної мережі для будь-якої її структури буде виконуватися на основі даних навчальної вибірки. Структура нейронної мережі визначається кількістю прихованих шарів і числом

нейронів в кожному з них та задається параметрично при створенні нового об'єкту. Процедура навчання являє собою послідовність ітерацій. На кожній ітерації вагові коефіцієнти нейронів підганяються з використанням нових даних з тренувальних зразків. Зміна вагових коефіцієнтів складає суть алгоритму. Кожний крок навчання починається з впливу вхідних сигналів тренувальних зразків. Після цього можуть бути визначені значення вихідних сигналів для всіх нейронів в кожному шарі мережі. На наступному кроці алгоритму вихідний сигнал мережі порівнюється з бажаним вихідним сигналом, який зберігається в тренувальних даних. Різниця між цими двома сигналами називається помилкою вихідного шару мережі. Через те що вихідні значення цих нейронів невідомі, неможливо безпосередньо обчислити сигнал помилки для внутрішніх нейронів. У процесі навчання до кожного масиву вхідних значень, тобто до кожного рядка, застосовується метод зворотного поширення помилки, ідея якого полягає в поширенні сигналу помилки назад на всі нейрони, чиї вихідні сигнали були вхідними для останнього нейрона. Таким чином, для кожного такого елемента отримується значення параметру вихідного нейрона, який буде використовуватися під час тестування мережі.

Алгоритм процедури тестування нейронної мережі складається з наступних кроків:

— формування масиву вагових значень для шару, а також масиву вагових значень для прихованого шару;

— формування масиву вхідних та вихідних значень;

— проходження за допомогою циклу по всіх вагових значеннях з накопиченням суми, кожен складник якої дорівнює значенню ваги, помноженому на вхідне значення;

— застосування до кожної такої суми функції активації та перехід до прихованого шару;

— проходження за допомогою циклу по всіх вагових значеннях прихованого шару з накопиченням суми, кожен складник якої дорівнює значенню ваги прихованого шару, помноженому на вхідне значення, що є результатом виконання попереднього пункту;

— отримання значення вихідного нейрона для даних, які були введені користувачем, та порівняння його із значеннями вихідних нейронів, отриманих у процесі навчання мережі;

— прийняття рішення на основі цих значень щодо приналежності об'єкта, який класифікується.

Використання багатошарової нейронної мережі з прихованим шаром та використання методу зворотного поширення помилки на етапі навчання нейронної мережі дозволило побудувати нейромережевий класифікатор з високою точністю класифікації. Крім того, навмисне спотворення вхідних даних (пропущення деяких вхідних даних) під час тестування нейронної мережі підтвердило стійкість запропонованого алгоритму до порушень вхідних передумов.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Başarslan M. S., Argun İ. D. Classification Of a bank data set on various data mining platforms // 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).— Turkey, Istanbul.— 2018.— P. 1–4.— doi: 10.1109/EBBT.2018.8391441.

2. Gu X., Liu L., Li J., Huang Y., Lin J. Data Classification based on Artificial Neural Networks // 2008 International Conference on Apperceiving Computing and Intelligence Analysis.— China, Chengdu.— 2008.— P. 223–226.— doi: 10.1109/ICACIA.2008.4770010.

3. Qu Z. Application of data mining in classification analysis of safety accidents based on alternate covering neural network // 2009 International Conference on Future BioMedical Information Engineering (FBIE). — China, Sanya.— 2009.— P. 144–147.— doi: 10.1109/FBIE.2009.5405861.

4. Jan T., Sajeev A. S. M. Boosted Probabilistic Neural Network for IoT Data Classification // 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing.— Greece, Athens.— 2018.— P. 408–411.— doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00082.

P. S. Safronov, O. F. Bondarenko

Data classification algorithm using neural networks

The authors propose a data classification algorithm using a neural network. The algorithm includes three main procedures: data format conversion, neural network training, and neural network testing. The data format conversion procedure converts the input data array to a usable format. The training of the neural network is based on the backpropagation. The testing procedure allows obtaining the values of the output neuron for the data entered by the user, compare it with the values of the output neurons obtained during neural network training, and make decisions based on these values about the class of the object being classified.

Keywords: data mining, data classification, machine learning, neural network, backpropagation method.