УДК 004.353
Н. Яворська, А. Яворський
(Тернопільський національний технічний університет імені Івана Пулюя)

ПОРІВНЯННЯ ДЕРЕВ РІШЕНЬ ТА НЕЙРОННИХ МЕРЕЖ ДЛЯ КЛАСИФІКАЦІЇ ТЕКСТІВ В ЗАДАЧАХ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

Зберігання інформації в письмовому вигляді на паперових носіях перестає бути актуальним. Інформація переходить на жорсткі диски комп'ютерів і в осередки пам'яті. Обсяги інформації ростуть і вимагають класифікації, як приклад можна взяти електронні сховища текстів. Автоматична класифікація з використанням комп'ютерів проводиться за допомогою алгоритмів кластеризації.

Розглянемо два методи класифікації – дерева рішень та нейронні мережі. Дані методи відносяться до методів машинного навчання, тому для їх використання необхідна низка текстів з уже відомими класами. Наявні дані поділяють на дві групи – навчальну і тестову вибірку. Перша використовується для побудови моделі, друга – для тестування (оцінці ефективності). У разі нейронних мереж іноді виділяють ще й третю, валідаційну групу, яка використовується для оцінки якості проміжних результатів (певної нейронної мережі, якщо проводиться перебір різних мереж).

Деревом рішень називають представлений у вигляді ациклічного графа план, за яким здійснюється класифікація об'єктів, описаних набором атрибутів. У задачі класифікації текстів об'єктами є тексти (фрагменти текстів), цільовим атрибутом – клас текстів (стиль, жанр і т.п.). Як атрибути, що описують текст, будемо використовувати кількісні ознаки тексту. Дерево на основі певних даних будується однозначно. Кожен вузол дерева містить умову розгалуження по одному з атрибутів. У тому випадку, коли атрибути, що описують об'єкт, номінальні («так – ні», «задовільно – добре – відмінно» і т.п.), вузол має стільки розгалужень, скільки значень має зворотній атрибут. Листя дерева містять значення цільового атрибута (класи, які брали участь в навчанні). Слідуючи по навченому дереву відповідно до значень атрибутів довільного об'єкта, ми опинимося в одному з листів дерева. Значення цього листа визначить значення цільового атрибута (клас) об'єкта.

Нейронні мережі – вид математичних моделей, які будуються за принципом організації мереж нервових клітин мозку. В основі їх побудови лежить ідея про те, що як завгодно складні процеси можна моделювати досить простими елементами (нейронами), а вся основна функціональність нейронної мережі, що складається з набору нейронів, забезпечується зв'язками між нейронами. Кожен нейрон являє собою акумулятор, він будує зважену суму своїх входів (вхід множиться на відповідний йому вагу) і потім пропускає цю величину через порогову функцію. Таким чином, виходить його вихідне значення. Для класифікації текстів на входи мережі пропонується подавати кількісні ознаки текстів. Розмірність вхідного шару в такому випадку буде дорівнює числу ознак.

На задачі класифікації текстів за кількісними ознаками нейронні мережі помітно перевершують по ефективності дерева рішень. Але при цьому їх настройка відбувається значно довше (в 10 і більше разів), ніж виробляється побудова дерев рішень. Крім того, дерева рішень дають відповідь, яка легко інтерпретується у вигляді набору правил вибору того чи іншого класу, а нейронні мережі – лише інформацію про ступінь приналежності до класів. Тому вибір того чи іншого методу залежить від завдання: якщо необхідна висока точність класифікації текстів, то слід застосовувати нейронні мережі, якщо потрібна наочна і швидка відповідь – дерева рішень.