УДК 004.22+004.89

В.М. ТЕРЕЩЕНКО
Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: *vtereshch@gmail.com*.

П.А. ЗАКАЛА
Київський національний університет імені Тараса Шевченка, Київ, Україна,
e-mail: *pzakala@gmail.com*.

ПОШУК БАЗОВОЇ МНОЖИНИ ДЛЯ ЗАДАЧ МАШИННОГО НАВЧАННЯ

Анотація. Розглянуто задачу пошуку базової множини та три способи її розв'язання: геометричний, із застосуванням генетичного алгоритму та на основі нейронних мереж. Проаналізовано ефективність кожного способу та зроблено висновки про шляхи їхнього використання. Особливу увагу приділено підходам на основі нейронних мереж. Проведено порівняльний аналіз різних підходів на основі нейронних мереж, описано їхні сильні та слабкі сторони, а також визначено подальші кроки для розв'язання задачі пошуку базової множини.

Ключові слова: базова множина, дистиляція даних, конденсація даних, геометрична базова множина, генетичні алгоритми.

ВСТУП

Протягом останніх 20-ти років підходи до роботи з даними зазнали суттєвих змін. Це пов'язано з тим, що ціни на зберігання інформації зменшились, а зберігання великих обсягів даних стало широкодоступним. Ще однією причиною цих змін став розвиток обчислювальних можливостей комп'ютерів. Як наслідок, виникли нові методи оброблення та аналізу даних, зокрема з використанням нейронних мереж.

Нейронні мережі зарекомендували себе як ефективний та універсальний метод. Проте їхньою основною проблемою залишається енергоефективність: навчання нейронних мереж потребує значних обчислювальних ресурсів. Одним зі способів розв'язання цієї проблеми є зменшення розміру вибірки, на якій навчають нейронну мережу. Це зменшить час навчання моделі і тим самим збереже ресурс.

Задача пошуку базової множини є методом зменшення обсягу даних для навчання. Базова множина — це вибірка із загального набору даних, яка максимізує інформацію, що зберігається в цих даних. У поєднанні з нейронними мережами цей підхід не лише пришвидшить процес навчання моделей, а й покращить якість самих даних. Аналізуючи базову множину, фахівець зможе краще зрозуміти природу даних, з якими він працює. Базова множина може стати потужним інструментом для аналізу даних, особливо складних, як-от зображення чи текстова інформація.

1. ПОСТАНОВКА ЗАДАЧІ

Базовою множиною (coreset) деякої множини D називають таку зважену множину C, що розв'язок, отриманий на цій множині, є порівнюваним (provably competitive) з розв'язком, отриманим на D. Для задач, що належать класу «навчання без учителя» (unsupervised learning), запропоновано таке означення базової множини [1].

Означення 1 (базова множина). Нехай задано набір даних D та деяку функцію втрат $cost(D, q)$, де $q \in Q$ — розв'язок. Зважену множину C називають