УДК 004.85:519.237

ПОРІВНЯЛЬНА ХАРАКТЕРИСТИКА МІР ВІДСТАНІ ТА ПОДІБНОСТІ В ЗАДАЧАХ МАШИННОГО НАВЧАННЯ

Марущак К. П.[1]; Козбур Г. В, к.т.н., доц.[2]
[1,2] Тернопільський національний технічний університет імені Івана Пулюя
[1, 2] Україна, Тернопіль
[1] katiamarushchak@gmail.com; [2] kozbur.galina@gmail.com

Анотація. *У роботі проведено порівняльний аналіз мір відстані (distance measures) та подібності (similarity measures) в задачах машинного навчання. Оцінено їх переваги, обмеження, застосування для різних типів даних, обчислювальну складність. Систематизована у порівняльну таблицю інформація дає змогу полегшити вибір оптимальної міри відстані.*
Ключові слова: *міри відстані; міри подібності; машинне навчання.*

Вступна частина. Міри подібності (similarity measures) та відстані (distance measures) відіграють ключову роль в різних застосуваннях науки про дані для широкого кола задач. Наприклад, в деяких задачах класифікації та регресії необхідно обчислювати відстані між новими елементами та тими, що доступні в навчальному наборі, алгоритми кластеризації, такі як ієрархічна кластеризація і k-середніх, вимагають обчислення мір подібності для групування об’єктів. Міри відстані використовуються для кількісної оцінки схожості або не схожості двох об’єктів на основі їх ознак. Конкретну міру варто розглядати не як вимірювання фізичної відстані, а як ступінь відмінності між об’єктами або вартість переходу з одного стану в інший. Кожен об’єкт можна представити у вигляді n-вимірного вектора.

Метою роботи є систематизація різних мір відстані і подібності, які застосовуються в машинному навчанні, зважаючи на їх часову складність, переваги, обмеження, особливості застосування.

Основна частина. Вибір міри відстані залежить від характеру даних, підходу, області застосування та поставленої задачі. Числові та категорійні ознаки потребують різних підходів до підбору мір відстані. Для кількісних ознак найчастіше застосовують міри відстані, підходи до знаходження яких полягають в обрахунку геометричного або статистичного відхилення між точками [1]. Для якісних ознак найчастіше застосовують міри подібності, що ґрунтуються на перетині множин або знаходженні поелементного збігу.

Застосування деяких мір відстані потребує попереднього масштабування даних, тобто зміни значень вхідних даних так, щоб ознаки мали однаковий діапазон значень, середнє значення, дисперсію чи інші статистичні характеристики. Це зменшує вплив дисбалансу масштабу ознак вхідних даних та збільшує точність результатів моделі через кращу узгодженість ознак [2]. Для даних змішаного типу ефективним є попереднє відображення об'єкта у проміжок $\langle 0, 1 \rangle$.

У таблиці 1 запропоновано систематизацію різних мір відстані, що використовуються в машинному навчанні, враховуючи оцінку часової складності, переваги, обмеження, особливості застосування, залежно від типів даних та підходу.

Таблиця 1 – Систематизація мір відстані за обчислювальною складністю і застосуванням

№	Вид міри відстані	Формула $D(x,y) =$	Перевага міри	Обмеження використання	Часова складність
1.	Евклідова відстань	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$	проста в розумінні та інтерпретації	є неефективною для розріджених даних, даних високої розмірності; потребує попередньої нормалізації даних; чутлива до викидів	$O(n)$
2.	Відстань Мінковського	$\left(\sum_{i=1}^{n}\lvert x_i - y_i\rvert^p\right)^{\frac{1}{p}}$	висока гнучкість через зміну значення параметра p	потребує підбору оптимального значення параметра p (високі значення надмірно підсилюють найбільші відхилень, зменшення параметра дозволяє мінімізувати вплив окремих викидів)	$O(n)$
3.	Косинусна подібність	$\cos(\theta) = \frac{x\,y}{\lVert x\rVert\lVert y\rVert}$	ефективна для розріджених даних високої розмірності; не залежить від масштабу ознак	не враховується величина, лише напрямок; передбачається попередня нормалізація даних для кращої ефективності	$O(n)$
4.	Відстань Махаланобіса	$\sqrt{(x-y)^T\Sigma^{-1}(x-y)}$	ефективно працює для виявлення викидів та на багатовимірних даних; не залежить від масштабу ознак	висока обчислювана складність; кількість спостережень має перевищувати кількість ознак	$O(n^3)$
5.	Відстань Канберри	$\sum_{i=1}^{n}\frac{\lvert x_i - y_i\rvert}{\lvert x_i\rvert + \lvert y_i\rvert}$	висока стійкість до шуму, викидів ознак з великими значеннями	може надмірно підсилювати вплив малих за модулем чисел	$O(n)$
6.	Подібність Жаккара [3]	$\frac{\lvert x \cap y\rvert}{\lvert x \cup y\rvert}$	підходить для бінарних або категоріальних даних; чітко обмежена діапазоном [0, 1]	значення залежить від розміру даних; потребує бінаризації для неперервних даних	$O(n)$
7.	Відстань Хеммінга	$\sum_{i=1}^{n}[x_i \neq y_i]$	оптимальна для двійкових векторів	точки даних мають мати однакову довжину; не враховується величина числової різниці	$O(n)$

У таблиці 1 використано позначення:

$D(x,y)$ – шукана міра відстані між n-вимірними об’єктами x та y;

$x = (x_i),\ y = (y_i),\ i \in \overline{1,\ n}.$

Висновки. Обґрунтований вибір міри відстані чи подібності залежить від природи ознак (кількісні, категорійні) та архітектури даних. Систематизація переваг, обмежень та обчислювальної складності розглянутих мір є основою для оптимізації алгоритмів машинного навчання, мінімізації впливу шуму та підвищення точності моделювання.

Використання інструментів ШІ. Під час підготовки тез було використано інструменти штучного інтелекту для стилістичного редагування тексту. Автори несуть повну відповідальність за зміст роботи.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165-193. https://doi.org/10.1007/s40745-015-0040-1

[2] Levy, A., Shalom, B. R., & Chalamish, M. (2025). A guide to similarity measures and their data science applications. *Journal of Big Data*, 12(1), 188. https://doi.org/10.1186/s40537-025-01227-1

[3] Maheshwari, H. (2021, October 13). *How to decide the perfect distance metric for your machine learning model.* Towards Data Science. https://towardsdatascience.com/how-to-decide-the-perfect-distance-metric-for-your-machine-learning-model-2fa6e5810f11/

Marushchak K. P., Kozbur H. V.

Comparative Analysis of Distance and Similarity Measures in Machine Learning Problems

Abstract. *This paper presents a comparative analysis of distance and similarity measures in machine learning tasks. It evaluates their advantages, limitations, applications to various data types, and computational complexity. The information, organized into a comparative table, facilitates the selection of an optimal distance measure.*

Keywords: *distance measures; similarity measures; machine learning.*