УДК 004.942, 004.89

СПОСОБИ ВИЗНАЧЕННЯ ПОДІБНОСТІ КАТЕГОРІАЛЬНИХ ВПОРЯДКОВАНИХ ДАНИХ

Кондрук Н. Е. – канд. техн. наук, доцент, доцент кафедри кібернетики і прикладної математики Ужгородського національного університету, Ужгород, Україна.

АНОТАЦІЯ

Актуальність. Розробка ефективних метрик відстані та мір подібності для категоріальних ознак є важливою задачею в аналізі даних, машинному навчанні, теорії прийняття рішень оскільки значна частина властивостей об'єктів описується саме не числовими значеннями. Зазвичай залежність між категоріальними ознаками може бути складнішою, ніж просто їх порівняння за рівністю чи нерівністю. Такі атрибути можуть бути відносно схожими, і для побудови ефективної моделі задачі необхідно врахувати цю подібність під час розрахунку відстані чи міри подібності.

Метою дослідження є підвищення ефективності розв'язання прикладних задач аналізу даних шляхом розробки математичних засобів для визначення подібності об'єктів за категоріальними впорядкованими ознаками.

Методи. Запропоновано відстань на базі зваженої манхетенської відстані та міру подібності для визначення схожості об'єктів за категоріальними впорядкованими ознаками (тобто на множині значень атрибутів можна задати лінійний порядок із шкалами переваг враховуючи предметну область задачі). Доведено, що формула відстані задовольняє аксіомам невід'ємності, симетричності, нерівності трикутника та обмеження з верху, а отже є метрикою відстані в просторі ранжованих категоріальних ознак. Доведено, що міра подібності представлена в досліджені задовольняє аксіомам обмеженості, симетричності, максимальної та мінімальної подібності та описується спадною функцією.

Результати. Розроблений підхід реалізовано на прикладній задачі визначення ступеню схожості об'єктів, які описані впорядкованими категоріальними ознаками.

Висновки. В даному дослідженні розроблено математичні інструменти для визначення подібності структурованих даних, що описуються категоріальними атрибутами, які можна впорядкувати за певним пріоритетом у вигляді рангу із системою переваг. Проаналізовано їх властивості. Проведені експериментальні дослідження показали зручність, «інтуїтивну зрозумілість» логіки проведення обробки даних при розв'язанні прикладних задач. Представлений підхід може забезпечити можливість проводити нові змістовні дослідження аналізу даних. Перспективи подальших досліджень полягають у експериментальному використанні запропонованих інструментів в практичних задачах та вивченні їх ефективності.

КЛЮЧОВІ СЛОВА: метрика відстані, міра подібності, подібність категоріальних даних, впорядковані дані.

НОМЕНКЛАТУРА

$\overline{a_i}\left(a_1^i, a_2^i, \ldots, a_n^i\right)$ – вектор значень (міток) ознак i-го об'єкту;

A_k – ознаки;

a_k^i – значення k-ї ознаки i-го об'єкту;

a_{kl} – l-а мітка k-ї ознаки;

$d(\cdot)$ – відстань;

m – кількість об'єктів (даних);

n – кількість ознак;

O – множина ознак об'єктів;

O_i – i-й об'єкт;

$\overline{r_i}\left(r_1^i, r_2^i, \ldots, r_n^i\right)$ – вектор рангів ознак для i-го об'єкту;

r_{kl} – ранг l-ї мітки k-ї ознаки;

s_k – кількість різних міток (значень) k-ї ознаки;

R^{ORD} – нечітке бінарне відношення, що характеризує схожість векторів впорядковуваних ознак;

$\mu_{R^{ORD}}(\cdot)$ – функція належності нечіткому бінарному відношенню R^{ORD};

Δ_k – розмах рангових шкал для k-ї ознаки.

ВСТУП

Важливими інструментами визначення схожості об'єктів в математиці є метрики відстані та міри подібності. Вони забезпечують вимірювання подібності між об'єктами, так що близькі дані вважатимуться подібними, тоді як віддалені – відмінними [1] і як правило відображають близькість в просторі ознак.

Враховуючи постійне збільшенням даних, що отримуються з різних носіїв та їх різнотипність, виявлення, аналіз існуючих зв'язків між ними привертає особливу увагу дослідників.

Подібність, як важливий взаємозв'язок, є числовим показником ступеня схожості двох об'єктів та зазвичай описується як функція відстані із змінними, що визначаються ознаками об'єкта. Це є принциповим і важливим для ефективної аналітики даних у різних областях, таких як аналіз даних, машинне навчання, теорія прийняття рішень, тощо.

Здебільшого в практичних задачах для опису властивостей об'єкта використовують атрибути різної природи. Якщо вони є числовими, то природніше порівнювати їх за рядом «перевірених» класичних метрик Мінковського, частинним випадком якої є метрика Евкліда. Однак, коли ознаки описуються нечисловими (категоріальними) атрибутами, аналіз подібності є набагато складніший. Таким чином,

DOI 10.15588/1607-3274-2023-2-4

задача визначення ступеню схожості або відмінності між такими об'єктами є непростою і відкритою для досліджень. Крім того, поняття подібності може різнитись залежно від конкретної предметної природи даних і повинно передбачати глибоке її розуміння. В ідеалі, поняття подібності визначається експертом галузі, який добре її розуміє [2, 3].

Метою даного дослідження є визначення метрики відстані та міри подібності для категоріальних впорядкованих даних.

Для досягнення мети в роботі необхідно розв'язати наступні задачі:

– визначити метрику відстані та міру подібності між об'єктами, що описуються категоріальними впорядковуваними ознаками;

– описати та обґрунтувати їх основні властивості;

– проілюструвати їх використання при розв'язанні прикладної задачі.

1 ПОСТАНОВКА ЗАДАЧІ

В багатьох методах машинного навчання, інтелектуального аналізу даних та прийняття рішень постає проблема визначення схожості об'єктів, що описані нечисловими ознаками. Тому виникає необхідність розробки ефективних, зрозумілих та обґрунтованих інструментів для її вирішення. В дослідженні розглянуто випадок, коли значення ознак (атрибутів) можна впорядкувати за значимістю, перевагою (певним рангом). Нехай задано нечислові ознаки A_k, $k=\overline{1,n}$ що характеризують деякі об'єкти $O_i, i=\overline{1,m}$. Кожна із A_k може набувати певних значень (міток), які можна проранжувати: $a_{k1} \succ a_{k2} \succ \ldots \succ a_{ks_k}$, тобто можна задати строгий лінійний порядок в просторі значень атрибутів. Кожному значенню a_{kl} ставиться у відповідність деяка числова величина переваги в рангу – r_{kl}, при чому, має виконуватись $r_{k1} > r_{k2} > \ldots > r_{ks_k}$ або $r_{k1} < r_{k2} < \ldots < r_{ks_k}$ для $k=\overline{1,n}$. Тоді, ставиться задача визначити метрику відстані та міру подібності між двома об'єктами O_i та O_j, що характеризуються векторами ознак – $\overline{a_i}\left(a_1^i, a_2^i, \ldots, a_n^i\right)$ та $\overline{a_j}\left(a_1^j, a_2^j, \ldots, a_n^j\right)$.

2 ОГЛЯД ЛІТЕРАТУРИ

Поняття подібності принципово важливе майже в кожній науковій галузі. Кластеризація, класифікація та регресія на основі відстані є базовими методами інтелектуального аналізу даних, які визначають схожість між об'єктами, а отже вибір конкретного показника подібності може виявитись основною причиною успіху або невдачі алгоритму. Міри подібності для категоріальних даних умовно можна розділити на два типи [3]:

1) ті, що моделюються експертами з відповідними знаннями предметних областей разом із експертами машинного навчання, які знають, як кодувати ці знання домену мірами подібності [4, 5].

2) ті, що моделюються на основі даних навчальної вибірки та дослідження зв'язків між ними [3, 6–9].

Останні міри подібності є дуже корисними в ситуаціях, коли існує багато розмічених даних, але відсутні знання предметної області або коли моделювання мір першого типу є занадто складним.

Однак, використання мір подібності на базі певного підходу машинного навчання, як правило, приводить до деякого описання подібності, що поєднує високу точність та низьку зрозумілість, тому може зменшити довіру користувача до системи [6].

Крім того, як зазначено в [2] не для всіх мір подібності виконуються необхідні аксіоми, що в подальшому збільшує обчислювальні потужності для реалізації алгоритмів, які їх використовують.

В цій роботі описану проблему, пропонується вирішити моделюванням легко зрозумілої та обґрунтованої міри подібності та метрики, що потенційно заслуговує на довіру в порівнянні із мірами подібності подібними до чорної скриньки.

3 МАТЕРІАЛИ І МЕТОДИ

Для кількісної оцінки схожості (відмінності) об'єктів вводяться поняття метрики та міри схожості, які опираються на систему аксіом. Для кожного об'єкта $O_i, i=\overline{1,m}$ закодуємо його вектор ознак $\overline{a_i}\left(a_1^i, a_2^i, \ldots, a_n^i\right)$ відповідним вектором показників рангів $\overline{r_i}\left(r_1^i, r_2^i, \ldots, r_n^i\right)$. Величина $\Delta_k = \left|r_{k1} - r_{ks_k}\right|$, $k=\overline{1,n}$ характеризує розмах рангової шкали для k-ї ознаки. Відстань між об'єктами пропонується визначати наступною формулою, яка є різновидом зваженої манхеттенської метрики:

$$d\left(O_i, O_j\right) = \sum_{k=1}^{n} \frac{\left|r_k^i - r_k^j\right|}{\Delta_k}. \quad (1)$$

Властивості.

1. Аксіома невід'ємності:
$d\left(O_i, O_j\right) \geq 0$, $\forall i, j$, причому, $d\left(O_i, O_j\right) = 0$, коли об'єкти однакові, тобто $i=j$.

2. Аксіома симетричності:
$$d\left(O_i, O_j\right) = d\left(O_j, O_i\right).$$

3. Аксіома нерівності трикутника:
$$d\left(O_i, O_z\right) + d\left(O_z, O_j\right) \geq d\left(O_i, O_j\right).$$

4. Властивість обмеженості зверху:

DOI 10.15588/1607-3274-2023-2-4

$d(O_i, O_j) \le n$, причому максимальне значення досягається, якщо об'єкти приймають крайні граничні значення рангів за всіма ознаками.

Доведення. Перші дві властивості очевидні, так як формула (1) містить модуль різниці рангів об'єктів. Доведемо правило трикутника:

$$d(O_i, O_z) + d(O_z, O_j) \ge d(O_i, O_j),$$

$$\sum_{k=1}^{n} \frac{\left|a_k^i - a_k^z\right|}{\Delta_k} + \sum_{k=1}^{n} \frac{\left|a_k^z - a_k^j\right|}{\Delta_k} \ge \sum_{k=1}^{n} \frac{\left|a_k^i - a_k^j\right|}{\Delta_k},$$

$$\sum_{k=1}^{n} \frac{\left|a_k^i - a_k^z\right|}{\Delta_k} + \sum_{k=1}^{n} \frac{\left|a_k^z - a_k^j\right|}{\Delta_k} - \sum_{k=1}^{n} \frac{\left|a_k^i - a_k^j\right|}{\Delta_k} \ge 0,$$

$$\sum_{k=1}^{n} \left(\frac{\left|a_k^i - a_k^z\right| + \left|a_k^z - a_k^j\right| - \left|a_k^i - a_k^j\right|}{\Delta_k} \right) \ge 0,$$

дана рівність завжди виконується, бо виконується:

$$|x - y| + |y - z| \ge |x - z|, \quad \forall x, y, z.$$

Якщо два об'єкти діаметрально різняться за певною ознакою (приймаючи крайні граничні значення в рангу), то величина $\left|a_k^i - a_k^j\right| = \Delta_k$. Якщо ж потрібно знайти відстань між об'єктами, які приймають діаметрально протилежні значення за всіма ознаками, то згідно (1), вона буде рівна *n*, а для всіх інших випадків менша за *n*.

Висновок. Отже, формула (1) визначає метрику відстані в просторі категоріальних ранжованих ознак.

В багатьох методах кластерного аналізу для визначення схожості між даними використовується поняття міри подібності. Міра подібності має задовольняти дещо іншим властивостям, ніж метрика відстані. Тому на основі (1) пропонується побудувати міру подібності:

$$\mu_{R^{ORD}}(O_i, O_j) = \frac{1}{e-1}\left(e^{-\frac{1}{n}\sum_{k=1}^{n} \frac{\left|a_k^i - a_k^j\right|}{\Delta_k} + 1} - 1 \right). \qquad (2)$$

Подібність об'єктів характеризується нечітким бінарним відношенням R^{ORD} на множині об'єктів $O = \left\{ O_i \middle| i = \overline{1,m} \right\}$ із функцією належності $\mu_{R^{\text{ORD}}} : O^2 \to [0,1]$. Чим більше значення величини $\mu_{R^{\text{ORD}}}$ ближче до 1, тим в більшому ступені об'єкти O_i та O_j будуть подібними за категоріальними ранжованими ознаками.

Дослідимо властивості міри подібності (2).

1. Аксіома обмеження:

$$0 \le \mu_{R^{ORD}} \le 1.$$

2. Аксіома симетричності:

$$\mu_{R^{ORD}}(O_i, O_j) = \mu_{R^{ORD}}(O_j, O_i).$$

3. Аксіома максимальної подібності:

$$\mu_{R^{ORD}}(O_i, O_j) = 1 \Leftrightarrow O_i = O_j.$$

4. Аксіома мінімальної подібності:

$\mu_{R^{ORD}}(O_i, O_j) = 0 \Leftrightarrow O_i$, O_j – гранично різняться.

5. Функція (2) є спадною.

Доведення: $\mu_{R^{ORD}}$ є функціоналом від $d(O_i, O_j)$:

$$\mu_{R^{ORD}}(O_i, O_j) = \frac{1}{e-1}\left(e^{-\frac{1}{n} d(O_i, O_j) + 1} - 1 \right), \qquad (3)$$

$$\mu_{R^{ORD}} = \frac{1}{e-1}\left(e^{-x+1} - 1 \right), \qquad (4)$$

де $x = \frac{1}{n} d(O_i, O_j)$. Враховуючи властивість 4 для (1) – $x \in [0; 1]$.

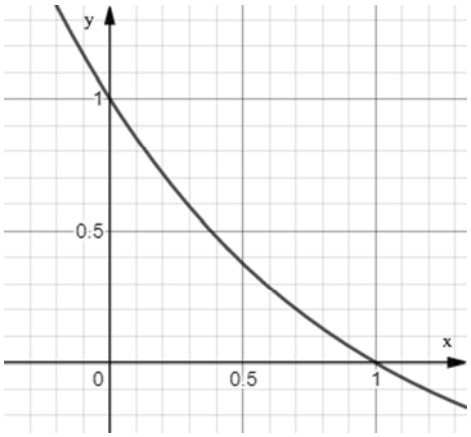

Рисунок 1 – Графік функції (4)

Із властивостей 1 і 4 для (1) та, що функція (4) є монотонно спадною (рис. 1) слідує, що:

– якщо об'єкти однакові, то $d(O_i, O_j) = 0$ і $\mu_{R^{ORD}}(O_i, O_j) = \frac{1}{e-1}\left(e^{-\frac{1}{n}\cdot 0 + 1} - 1 \right) = 1$;

– якщо об'єкти діаметрально різняться, то $d(O_i, O_j) = n$ і $\mu_{R^{ORD}}(O_i, O_j) = \frac{1}{e-1}\left(e^{-\frac{1}{n}\cdot n + 1} - 1 \right) = 0$;

– в інших випадках $0 < \mu_{R^{ORD}} < 1$.

DOI 10.15588/1607-3274-2023-2-4

Отже, аксіоми 1, 3, 4 доведено.

Аксіома симетричності випливає із 2-ї властивості для (1) і (3).

4 ЕКСПЕРИМЕНТИ

Нехай задано об'єкти, що описують осіб-клієнтів і визначені впорядкованими категоріальними ознаками. Ознака A_1 – «Вік» може набувати таких значень {«юний», «молодий», «середній», «похилий»}. Ознака A_2 – «Освіта» може приймати значення {«середня», «професійно-технічна», «фахова передвища», «вища/бакалавр», «вища/магістр»}. Ознака A_3 – «Місце проживання» може приймати значення {«село», «селище міського типу», «мале місто», «середнє місто», «велике місто», «місто-мільйонник»}. Нехай задано три об'єкти O_1, O_2, O_3, що описані відповідними значеннями ознак.

Таблиця 1 – Значення ознак об'єктів

Об'єкт	Вік	Освіта	Місце проживання
O_1	Середній	Вища/бакалавр	Сер. місто
O_2	Юний	Середня	Село
O_3	Похилий	Вища/магістр	Місто-м.

Закодуємо значення ознак призначивши величини рангів (табл. 2).

Таблиця 2 – Ранги значень (міток) ознак

Ознака A_1		Ознака A_2		Ознака A_3	
значення	ранг	значення	ранг	значення	ранг
Юний	21	Середня	1	Село	1
Молодий	34	Проф.тех.	3	С.М.Т.	2
Середній	52	Фах.перед	4	М. місто	3
Похилий	75	Вища/бак	6	С. місто	4
		Вища/маг	7	В. місто	5
				Місто-м.	6

Для прикладу, розрахуємо відстані між об'єктами за формулою (1) та міру подібності за (3).

$$d(O_1,O_3)=\frac{|52-75|}{54}+\frac{|6-7|}{6}+\frac{|4-6|}{5}=0{,}99,$$

$$\mu_{R^{ORD}}(O_1,O_3)=\frac{1}{e-1}\left(e^{-\frac{1}{3}\cdot 0{,}99+1}-1\right)=0{,}55.$$

5 РЕЗУЛЬТАТИ

Розрахуємо відстань (1) та міру подібності (4) між об'єктами O_1, O_2, O_3 (табл. 3, табл. 4).

Таблиця 3 – Числові значення відстані (1) між об'єктами

$d(O_i, O_j)$	O_1	O_2	O_3
O_1	0	2	0,99
O_2	2	0	3
O_3	0,99	3	0

Таблиця 4 – Числові значення міри подібності (4) між об'єктами

$\mu_{R^{ORD}}(O_1,O_2)$	O_1	O_2	O_3
O_1	1	0,23	0,56
O_2	0,23	1	0
O_3	0,56	0	1

Із отриманих результатів слідує, що метрику відстані можна інтерпретувати як ненормовану величину визначення подібності між об'єктами, а міру подібності – стандартизовану із проміжку [0; 1].

6 ОБГОВОРЕННЯ

Отже, отримані результати демонструють приклади використання метрики відстані (1) та міри подібності (2) для визначення подібності, схожості об'єктів, що описуються впорядкованими ознаками. Важливим моментом є призначення числових характеристик рангів. Система рангів може бути задана, як послідовні натуральні числа, так і містити скачки. Все залежить від конкретних особливостей прикладних задач. Так, за основу рівнів рангів ознаки A_1 взято середні значення вікових рівнів визначених в [10]; для A_2 – ранжування, в якому кожен рівень освіти виділений скачком; для ознаки A_3 – послідовне ранжування. Якщо в прикладній задачі кожен рівень певної ознаки має суттєво впливати на розв'язок, то рекомендується їх підсилювати системою рангів із скачками. Так, ранги значень ознаки A_3 можна було визначити у відповідності до кількості населення відповідного типу населеного пункту, але для задачі оцінки схожості об'єктів-клієнтів це не є значущим.

В метриці відстані (1) кожний її доданок є нормованою величиною і приймає значення із проміжку [0; 1], тому «вклади» кожної ознаки у відстань є співрозмірними. З іншого боку, кожна складова (1) характеризуватиме відсоток відмінності об'єктів за відповідною ознакою.

Міра подібності (4) має хорошу чутливість в околі нуля (рис. 1), що дозволить використовувати її для кластеризації із визначенням порогів [11]. Також, областю застосування відстані та міри можуть бути методи класифікації, що використовують метрику відстані (наприклад, *k*NN), методи кластерного аналізу даних (наприклад, *k*-means, метод кластеризації заснований на нечітких бінарних відношеннях); деякі задачі прийняття рішень.

ВИСНОВКИ

Вирішується проблема розробки математичного апарату для аналізу схожості об'єктів, що описуються впорядковуваними категоріальними ознаками. Дана праця є продовженням досліджень вивчення застосування різних мір подібності до розв'язання різних практичних задач, де виникає потреба визначення структури даних та аналізу прихованих взаємозв'язків між ними.

Наукова новизна отриманих результатів полягає в тому, що запропоновані відстань (1) та міра

DOI 10.15588/1607-3274-2023-2-4

подібності (3). Доведено, що (1) є різновидом манхетенської зваженої відстані і є метрикою в просторі впорядкованих категоріальних ознак, а (4) задовольняє аксіомам, що визначають міру подібності.

Продемонстровано можливе використання (1) та (3) на конкретних прикладах.

Практичне значення отриманих результатів полягає в можливості застосування відстані d та міри $\mu_{R^{ORD}}$ при розв'язанні різних задачах аналізу даних та машинного навчання, які описуються впорядкованими категоріальними ознаками та деяких класів задач прийняття рішень.

ПОДЯКИ

Роботу виконано в рамках держбюджетної науково-дослідної теми Ужгородського національного університету «Методи обчислювального інтелекту для обробки і аналізу даних» (номер державної реєстрації 0121U109279).

ЛІТЕРАТУРА

1. Suárez J. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges / J. Suárez, S. García, F. Herrera // Neurocomputing. – 2021. – №425. – P. 300–322. DOI: 10.1016/j.neucom.2020.08.017
2. Learning similarity measures from data / [B. Mathisen, A. Aamodt, K. Bach, H. Langseth] // Progress in Artificial Intelligence. – 2019. – №9. – P. 1–15. DOI: 9. 10.1007/s13748-019-00201-2.
3. DISC: Data-Intensive Similarity Measure for Categorical Data / [A. Desai, H. Singh, V. Pudi et al.]. // Advances in Knowledge Discovery and Data Mining. – 2011. – №6635. – P. 469–481. DOI: 10.1007/978-3-642-20847-8_39
4. Cunningham P. A Taxonomy of Similarity Mechanisms for Case-Based Reasoning / Padraig Cunningham. // IEEE Transactions on Knowledge and Data Engineering. – 2009. – №21. – P. 1532–1543. DOI: 10.1109/TKDE.2008.227.
5. Nikpour N. Bayesian-Supported Retrieval in BNCreek: A Knowledge-Intensive Case-Based Reasoning System / N. Nikpour, A. Aamodt, K. Bach. // Case-Based Reasoning Research and Development. – 2018. – №11156. – P. 323–338. DOI: 10.1007/978-3-030-01081-2_22
6. Top-Down Induction of Similarity Measures Using Similarity / [T. Gabel, E. Godehardt, E. Hüllermeier, M. Minor] // Case-Based Reasoning Research and Development. – 2015. – №9343. – P. 149–164. DOI: 10.1007/978-3-319-24586-7_11
7. Hoffer E. Deep Metric Learning Using Triplet Network / E. Hoffer, N. Ailon. // Similarity-Based Pattern Recognition. – 2014. – №9370. – P. 84–92. DOI: 10.1007/978-3-319-24261-3_7
8. A method for k-means-like clustering of categorical data / [T. Nguyen, T. Dinh, S. Sriboonchitta, V. Huynh] // Journal of Ambient Intelligence and Humanized Computing. – 2019. – P. 1–11. DOI: 10.1007/s12652-019-01445-5.
9. Learning similarity measures from data / [B. Mathisen, A. Aamodt, K. Bach, H. Langseth] // Progress in Artificial Intelligence. – 2020. – №9. – P. 129–143. DOI: 10.1007/s13748-019-00201-2
10. Dyussenbayev A. Age Periods Of Human Life / A. Dyussenbayev // Advances in Social Sciences Research Journal. – 2017. – №4. – P. 258–263. DOI:10.14738/assrj.46.2924
11. Kondruk N. Clustering method based on fuzzy binary relation / N. Kondruk // Eastern-European Journal of Enterprise Technologies. – 2017. – № 2(4). – P. 10–16. DOI: 10.15587/1729–4061.2017.94961
12. Kondruk, N.E., Malyar, M.M. Analysis of Cluster Structures by Different Similarity Measures / N. E. Kondruk, M. M. Malyar, // Cybernetics and Systems Analysis. – 2021. – №57. – P. 436–441. https://doi.org/10.1007/s10559-021-00368-4.
13. Кондрук Н. Е. Використання довжинної міри подібності в задачах кластеризації / Н. Е. Кондрук // Радіоелектроніка, інформатика, управління. – 2018. – №3 (46). – С. 98–105. DOI: 10.15588/1607-3274-2018-3-11.

Стаття надійшла до редакції 15.04.2023.
Після доробки 28.05.2023.

UDC 004.942, 004.89

METHODS FOR DETERMINING SIMILARITY OF CATEGORICAL ORDERED DATA

Kondruk N. E. – PhD, Associate professor, Associate Professor of Department of Cybernetics and Applied Mathematics, Uzhhorod National University, Uzhhorod, Ukraine.

ABSTRACT

Context. The development of effective distance metrics and similarity measures for categorical features is an important task in data analysis, machine learning, and decision theory since a significant portion of object properties is described by non-numerical values. Typically, the dependence between categorical features may be more complex than simply comparing them for equality or inequality. Such attributes can be relatively similar, and to construct an effective model, it is necessary to consider this similarity when calculating distance or similarity measures.

Objective. The aim of the study is to improve the efficiency of solving practical data analysis problems by developing mathematical tools for determining the similarity of objects based on categorical ordered features.

Method. A distance based on weighted Manhattan distance and a similarity measure for determining the similarity of objects based on categorical ordinal features (i.e. a linear order with scales of preference considering the problem domain can be specified on the attribute value set) are proposed. It is proven that the distance formula satisfies the axioms of non-negativity, symmetry, triangle inequality, and upper bound, and therefore is a distance metric in the space of ranked categorical features. It is also proven that the similarity measure presented in the study satisfies the axioms of boundedness, symmetry, maximum and minimum similarity, and is described by a decreasing function.

DOI 10.15588/1607-3274-2023-2-4

Results. The developed approach has been implemented in an applied problem of determining the degree of similarity between objects described by ordered categorical features.

Conclusions. In this study, mathematical tools were developed to determine similarity between structured data described by categorical attributes that can be ordered based on a specific priority in the form of a ranking system with preferences. Their properties were analyzed. Experimental studies have shown the convenience and "intuitive understanding" of the logic of data processing in solving practical problems. The proposed approach can provide the opportunity to conduct new meaningful research in data analysis. Prospects for further research lie in the experimental use of the proposed tools in practical tasks and in studying their effectiveness.

KEYWORDS: distance metric, similarity measure, categorical data similarity, ordered data.

REFERENCES

1. Suárez J., García S., Herrera F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges, *Neurocomputing*, 2021, № 425, pp. 300–322. DOI: 10.1016/j.neucom.2020.08.017
2. Mathisen B., Aamodt A., Bach K., Langseth H. Learning similarity measures from data, *Progress in Artificial Intelligence,* 2019, № 9, pp. 1–15. DOI: 9. 10.1007/s13748-019-00201-2.
3. Desai A., Singh H., Pudi V. et al. DISC: Data-Intensive Similarity Measure for Categorical Data, *Advances in Knowledge Discovery and Data Mining,* 2011, №6635, pp. 469–481. DOI: 10.1007/978-3-642-20847-8_39
4. Cunningham P. A Taxonomy of Similarity Mechanisms for Case-Based Reasoning, *IEEE Transactions on Knowledge and Data Engineering*, 2009, №21, pp. 1532–1543. DOI: 10.1109/TKDE.2008.227.
5. Nikpour N., Aamodt A., Bach K. Bayesian-Supported Retrieval in BNCreek: A Knowledge-Intensive Case-Based Reasoning System, *Case-Based Reasoning Research and Development*, 2018, №11156, pp. 323–338. DOI: 10.1007/978-3-030-01081-2_22
6. Gabel T., Godehardt E., Hüllermeier E., Minor M. Top-Down Induction of Similarity Measures Using Similarity, *Case-Based Reasoning Research and Development*, 2015, №9343, pp. 149–164. DOI: 10.1007/978-3-319-24586-7_11
7. Hoffer E., Ailon N. Deep Metric Learning Using Triplet Network, *Similarity-Based Pattern Recognition*, 2014, №9370, pp. 84–92. DOI: 10.1007/978-3-319-24261-3_7
8. Nguyen T., Dinh T., Sriboonchitta S., Huynh V. A method for k-means-like clustering of categorical data, *Journal of Ambient Intelligence and Humanized Computing,* 2019, pp. 1–11. DOI: 10.1007/s12652-019-01445-5.
9. Mathisen B., Aamodt A., Bach K., Langseth H. Learning similarity measures from data, *Progress in Artificial Intelligence,* 2020, №9, pp. 129–143. DOI: 10.1007/s13748-019-00201-2
10. Dyussenbayev A. Age Periods Of Human Life, *Advances in Social Sciences Research Journal,* 2017, №4, pp. 258–263. DOI:10.14738/assrj.46.2924
11. Kondruk N. Clustering method based on fuzzy binary relation, *Eastern-European Journal of Enterprise Technologies*, 2017, № 2(4), pp. 10–16. DOI: 10.15587/1729–4061.2017.94961
12. Kondruk, N. E., Malyar M. M. Analysis of Cluster Structures by Different Similarity Measures, *Cybernetics and Systems Analysis*, 2021, №57, pp. 436–441. https://doi.org/10.1007/s10559-021-00368-4.
13. Kondruk N. E. Vykorystannja dovzhynnoi' miry podibnosti v zadachah klasteryzacii', *Radio Electronics, Computer Science, Control,* 2018, №3 (46), pp. 98–105. DOI: 10.15588/1607-3274-2018-3-11.

DOI 10.15588/1607-3274-2023-2-4