https://doi.org/10.31649/1997-9266-2023-167-2-49-57

УДК 004.852

С. Д. Штовба[1]
М. В. Петричко[2]
М. Ю. Петранова[1]

МЕТРИКА СХОЖОСТІ КАТЕГОРІАЛЬНИХ РОЗПОДІЛІВ, ЩО ВРАХОВУЄ СПОРІДНЕНІСТЬ РІЗНИХ КАТЕГОРІЙ

[1]Донецький національний університет імені Василя Стуса, Вінниця;
[2]Вінницький національний технічний університет

Оцінювання схожості двох об'єктів — це поширена задача в розпізнаванні образів, кластеризації та класифікації. Прикладами таких задач є підбір рецензентів наукових робіт, аналіз схожості текстових документів, ідентифікація поз людей у відеоряді, кластеризація природних ареалів, формування рекомендацій в інтернет-магазинах тощо. У випадку категоріальних атрибутів об'єкти описуються деяким розподілом ступенів належності за категоріями. Метрики схожості таких розподілів зазвичай є суперпозицією схожості об'єктів за кожною категорією. Найчастіше це сума схожості за окремими категоріями. При цьому, кожна категорія розглядається незалежно та ізольовано від інших. В деяких практичних задачах категорії є спорідненими. Тому схожість між об'єктами доцільно розраховувати не лише напряму, як схожість між еквівалентними категоріями, але враховувати і непряму, перехресну схожість через споріднені категорії. Саме така метрика схожості двох категоріальних розподілів, що враховує спорідненість різних категорій, і пропонується у статті. Метрика має дві складові. Перша складова реалізована метрикою Чекановського. Вона визначає пряму схожість розподілів за категоріями як суму перетину розподілів належностей двох об'єктів. Після перетину розподілів залишаються залишки, які і враховуються другою складовою запропонованої метрики. Друга складова метрики є сумою поелементного добутку двох матриць: матриці композиції залишків належності двох категоріальних розподілів та матриці попарної спорідненості категорій. Передбачається, що коефіцієнти спорідненості кожної пари категорій є відомими. Встановлено, що за великої кількості категорій сумарний шумовий внесок від слабо споріднених категорій є значним. Тому запропоновано цей шум фільтрувати і враховувати лише внесок від сильно споріднених категорій.

Ключові слова: категоріальний розподіл, споріднені категорії, метрика схожості, метрика Чекановського, розпізнавання поз, підбір рецензентів, узагальнений розподіл Парето.

Вступ

Оцінювання схожості двох об'єктів — це поширена задача в розпізнаванні образів, кластеризації та класифікації. В цих задачах кожний об'єкт описується вектором атрибутів. Об'єкти можуть задаватися в метричному просторі, тоді кожний атрибут задається на числовій шкалі. Наприклад, в задачі про фішеровські іриси кожна квітка описується чотирма атрибутами, а саме, шириною і довжиною пелюстки та шириною і довжиною чашолистка. Атрибути об'єкта можуть бути і категоріальними, тоді він описується розподілом ступенів належності за категоріями. Таке категоріальне представлення об'єктів часто використовується в задачах класифікації та тематичного моделювання. Для згаданого набору даних результат розпізнавання квітки може бути у формі категоріального розподілу, наприклад, зі ступенем належності 0,7 ірис відноситься до класу *Iris Setosa*, зі ступенем належності 0,1 ірис відноситься до *Iris Virginica* та зі ступенем належності 0,2 — до *Iris Versicolor*.

В залежності від типу опису об'єктів використовують різні метрики схожості об'єктів. Для об'єктів у метричному просторі схожість визначають як величину обернену чи інверсну до відстані між двома точками. Координатами кожної точки є числові значення атрибутів відповідного об'єкта. Чим менше відстань між аналізованими об'єктами, тим вони подібніші. В статті [1] проаналізовано майже 50 різних метрик, найпопулярнішими серед них є частинні випадки метрики

Мінковського — евклідова відстань, манхетенська відстань та метрика Чебишева. Часто використовується також і косинусна метрика, за якою розраховується косинус кута між двома векторами, які виходять з початку координат та прямують до аналізованих об'єктів.

У категоріальному просторі схожість двох об'єктів визначається, зазвичай, як суперпозиція схожості об'єктів за кожною категорією. Найчастіше — це сума схожості за окремими категоріями. При цьому, кожна категорія розглядається незалежно та ізольовано від інших. Є і зворотний підхід, коли спочатку визначають розбіжність об'єктів за кожною категорією, а потім їх агрегують, щоб розрахувати загальну схожість. Один із популярних варіантів такої метрики запропоновано в [2] для розрахунку схожості нечітких множин. В тій статті розбіжність об'єктів визначається через модуль різниці ступенів належності. Усі метрики з оглядової статі [1] та з інших релевантних публікаціях, наприклад, [3], [4] передбачають відсутність спорідненості між категоріями. Але, для деяких практичних задач категорії є спорідненими. Це призводить до того, що схожість між об'єктами слід розраховувати не лише напряму, як схожість між еквівалентними категоріями, але і враховувати непряму, перехресну схожість через споріднені категорії. Розробка такої метрики, яка додатково враховує схожість об'єктів через спорідненні категорії, і є *метою статті.*

Опис об'єктів в просторі споріднених категорії

Розглянемо два приклади змістовних задач, в яких об'єкти описуються в просторі споріднених категорій.

Під час розпізнавання відеосцен виникає задача визначення схожості двох об'єктів. Така задача має місце під час аналізу одного відеокадру, коли потрібно вибрати пару схожих об'єктів. Наприклад, вибрати пару людей, які знаходяться у найподібніших позах. Можуть аналізуватися і різні відеокадри. Тоді, задача змістовно інтерпретується як визначення пари акторів з різних кадрів, які знаходяться в подібних позах. Стан кожного актора описується вектором належності до категорій. Наприклад, положення верхньої кінцівки можна описати в просторі 9-ма категоріями: {*Вертикально піднята, Сильно піднята вперед, Сильно піднята вбік, Горизонтально вперед, Горизонтально вбік, Трохи піднята вперед, Трохи піднята вбік, Трохи піднята назад, Опущена}.* Під час аналізу відеосцен не завжди вдається достовірно встановити належність пози актора лише до однієї категорії, наприклад, встановити, що його права рука однозначно відповідає лише категорії «*Трохи піднята назад*». Часто результат аналізу відеосцени представляється у вигляді деякого розподілу належностей за категоріями, наприклад, стан правої руки актора описується як «*Трохи піднята вбік*» з належністю 0,6, *«Трохи піднята назад»* з належністю 0,3 та «*Опущена*» з належністю 0,1. Якщо для актора встановлено, що належність його руки до положення «*Трохи піднята вбік*» становить 0,6, то він буде схожий не лише на актора, у якого рука в такому ж положенні, але деякою мірою він буде схожий і на актора у спорідненних позах з рукою у положеннях «*Горизонтально вбік*», або «*Трохи піднята вперед*», або «*Трохи піднята назад*». При цьому, він зовсім не буде схожим на актора з неспорідненних категорій, до прикладу, з рукою в положенні «*Вертикально піднята*». Але, за відомими метриками під час оцінювання схожості об'єктів усі категорії розглядаються ізольовано, і не враховується їхня спорідненість. Врахування такої спорідненості дозволило би точніше оцінити пози людей, у яких положення різних частин тіла не збігаються, але близькі за розміщенням у просторі.

Розглянемо задачі підбору схожих науковців, наприклад, для рецензування. На підставі наукового доробку кожний науковець може бути категоризований до кількох спеціальностей в рамках деякої системи класифікації наук. Наприклад, науковця *A* віднесено до спеціальності *«Системний аналіз»* зі ступенем належності 0,4 та до спеціальності *«Інформаційні системи та технології»* зі ступенем 0,6. Науковця *B* віднесено до спеціальності *«Системний аналіз»* зі ступенем належності 0,7 та до спеціальності *«Комп'ютерні науки»* зі ступенем 0,3. Науковця *C* віднесено до спеціальності *«Системний аналіз»* зі ступенем належності 0,4 та до спеціальності «*Маркетинг*» зі ступенем 0,6. За будь-якою з відомих метрик схожість між парою вищенаведених науковців буде встановлено лише за їхніми належностями до спільної спеціальності *«Системний аналіз».* Належності до інших категорій не враховуються тому, що вони у науковців не збігаються. Схожість між науковцями *A* та *B* визначається виключно на основі їхніх ступенів належності до категорії *«Системний аналіз»*, які дорівнюють 0,4 та 0,7. Якщо схожість визначати за спільною часткою належності, використовуючи операцію мінімуму, отримуємо, що схожість науковців *A*

та B дорівнює $Fit(A,B)=\min(0,4;\ 0,7)=0,4$. Аналогічно, схожість науковців A та C дорівнює $Fit(A,C)=\min(0,4;\ 0,4)=0,4$, а науковців B та C дорівнює $Fit(B,C)=\min(0,7;\ 0,4)=0,4$. Виходить, що схожість усіх пар науковців однакова. Але предметна область спеціальностей така, що *«Інформаційні системи та технології»* значно ближче до *«Комп'ютерних наук»* ніж до *«Мар*кетингу». Також, «*Комп'ютерні науки*» та «*Інформаційні системи та технології*» значно ближче до «*Системного аналізу*» ніж до «*Маркетингу*». Відповідно схожість науковців A та B має бути вищою ніж схожість науковців A та C чи науковців B та C. Але відомі метрики схожості не враховують спорідненість категорій, тому за ними неможливо врахувати такі особливості.

Пропонована метрика

Позначимо кількість категорій через m. Тоді, об'єкти X та Y, схожість яких будемо оцінювати, опишемо такими розподілами належностей до категорій: $(\mu_1(X),\mu_2(X),...,\mu_m(X))$ та $(\mu_1(Y),\mu_2(Y),...,\mu_m(Y))$. Розподіли вважатимемо нормалізованими, що задовольняють такі умови:

$$\mu_i(X)\in[0;1],\ \mu_i(Y)\in[0;1],\ i=\overline{1,m};$$

$$\sum_{i=\overline{1,m}}\mu_i(X)=1;$$

$$\sum_{i=\overline{1,m}}\mu_i(Y)=1.$$

Задача полягає в тому, щоб для об'єктів X та Y розрахувати показник схожості. Специфіка предмету дослідження полягає в тому, що деякі категорії є спорідненими. Відповідно, слід враховувати не лише схожість за ідентичними категоріями, але і за спорідненими. Нижче пропонується така метрика, яка враховує семантичну спорідненість категорій.

Схожість двох об'єктів X та Y пропонується визначити таким чином:

$$Fit(X,Y)=F(X,Y)+\Delta F(X,Y), \tag{1}$$

де $F(X,Y)$ — доданок, що оцінює безпосередню (пряму) схожість об'єктів X та Y за категоріями; $\Delta F(X,Y)$ — доданок, що враховує схожість об'єктів X та Y через споріднені категорії.

Перший доданок в формулі (1) розрахуємо за спрощеним варіантом метрики Чекановського для випадку, коли ступені належності знаходяться у діапазоні $[0,1]$ і розподіли пронормовані. Розрахункова формула така:

$$F(X,Y)=\sum_{i=\overline{1,m}}\min(\mu_i(X),\mu_i(Y)), \tag{2}$$

де $\mu_i(X)$ — ступінь належності об`єкта $X$ до $i$-ї категорії, $i=\overline{1,m}$; $\mu_i(Y)$ — ступінь належності об`єкта Y до i-ї категорії, $i=\overline{1,m}$.

Формулу (2) можна інтрепретувати як суму належностей перетину нечітких множин X та Y. В формулі (2) вважається, що загальна схожість двох об'єктів є сумою їхніх схожостей за кожною категорією. Схожість за категорією визначається як мінімум належностей обох об'єктів до цієї категорії. Таким чином, у метрику схожості (2) один з об'єктів вносить усе значення ступеня належності до категорії, а інший — лише частину.

Після застосування формули (2) отримуємо такі залишки належності:

$$r_i(X)=\max(0,\mu_i(X)-\mu_i(Y));$$

$$r_i(Y)=\max(0,\mu_i(Y)-\mu_i(X)),\ i=\overline{1,m}.$$

Врахуємо внесок залишків у схожість двох об'єктів через спорідненість категорій. Вважатимемо, що інформація про попарну спорідненість категорій подана у формі такого бінарного відношення:

$$\mathbf{K}=\|k_{ij}\|,$$

де $k_{ij} \in [0,1]$ — коефіцієнт спорідненості *i*-ї та *j*-ї категорій, $i = \overline{1,m}$, $j = \overline{1,m}$.

Чим подібніші категорії, тим вищий коефіцієнт спорідненості. Відношення спорідненості є симетричним та рефлексивним $k_{ij} = k_{ji}$ та $k_{ii} = 1$ відповідно.

Композицію залишків представимо такою матрицею:

$$\mathbf{E} = \left\| e_{ij} \right\|,$$

де $e_{ij} = \min\left(r_i(X), r_j(Y)\right)$, $i = \overline{1,m}$, $j = \overline{1,m}$.

Внесок залишків у метрику (1) через попарну спорідненість категорій розрахуємо так:

$$\Delta F(X,Y) = \sum_{i=1,m} \sum_{j=1,m} \left(e_{ij} \cdot k_{ij}\right). \tag{3}$$

Приклад. Задано 2 об'єкти з такими належностями до категорій $\{A, B, C, D\}$:

$X = (0{,}5 \quad 0{,}2 \quad 0{,}1 \quad 0{,}2)$ та $Y = (0{,}7 \quad 0{,}1 \quad 0{,}2 \quad 0)$. Спорідненість категорій описана такою матрицею:

$$\mathbf{K} = \begin{Vmatrix} 1{,}0 & 0{,}5 & 0{,}0 & 0{,}0 \\ 0{,}5 & 1{,}0 & 0{,}1 & 0{,}0 \\ 0{,}0 & 0{,}1 & 1{,}0 & 0{,}3 \\ 0{,}0 & 0{,}0 & 0{,}3 & 1{,}0 \end{Vmatrix}.$$

Розрахуємо схожість об'єктів *X* та *Y* за запропонованою метрикою (1).

Для розрахунку першого доданку метрики схожості (1) виконаємо перетин двох розподілів, який схематично подано на рис. 1. На рисунку заштриховано спільні частини розподілів за кожною категорією.

Числове значення першого доданку у формулі (1) є таким:

$$F(X,Y) = \min(0{,}5,\ 0{,}7) + \min(0{,}2,\ 0{,}1) + \min(0{,}1,\ 0{,}2) + \min(0{,}2,\ 0) = 0{,}5 + 0{,}1 + 0{,}1 + 0 = 0{,}7.$$

Залишки після перетину становлять: $e(X) = (0 \quad 0{,}1 \quad 0 \quad 0{,}2)$ та $e(Y) = (0{,}2 \quad 0 \quad 0{,}1 \quad 0)$.

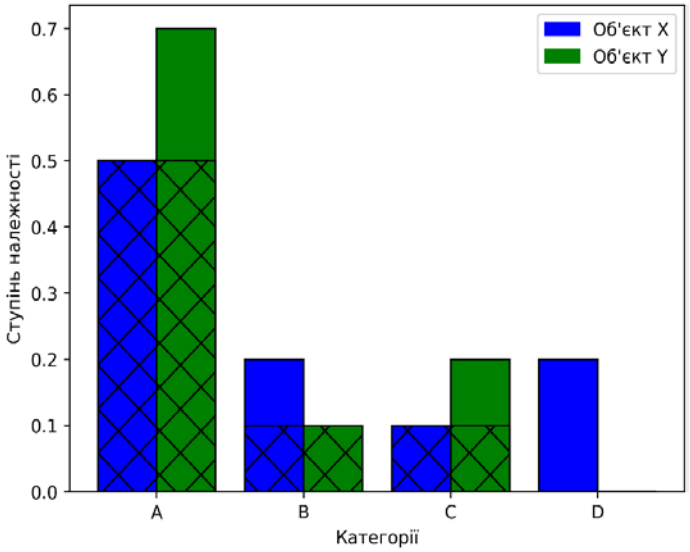

Рис. 1. Перетин двох категоріальних розподілів для розрахунку $Fit(X,Y)$

Композиція залишків дорівнює

$$\mathbf{E} = \begin{Vmatrix} 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}1 & 0{,}0 & 0{,}1 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}2 & 0{,}0 & 0{,}1 & 0{,}0 \end{Vmatrix}.$$

Виконавши поелементний добуток матриць **E** та **K**, отримуємо таку матрицю внесків через споріднені категорії:

$$\begin{Vmatrix} 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}05 & 0{,}0 & 0{,}01 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}03 & 0{,}0 \end{Vmatrix}.$$

З цієї матриці видно, що внесок від врахування спорідненості другої та першої категорій становить 0,05, внесок від врахування спорідненості другої та третьої категорій становить 0,01, а внесок від врахування спорідненості четвертої та третьої категорій становить 0,03. Внесок через спорідненість інших категорій є нульовим. Сумарний внесок від усіх споріднених категорій становить

$$\Delta F(X,Y) = 0{,}05 + 0{,}01 + 0{,}03 = 0{,}09.$$

Підсумкове значення схожості об'єктів *X* та *Y* за формулою (1) дорівнює $F(X,Y) = 0{,}7 + 0{,}09 = 0{,}79$.

Експериментальні дослідження запропонованої метрики

В вищенаведеному прикладі врахування спорідненості категорій збільшило метрику схожості на 0,09, що становить 13 % від початкового значення, отриманого за метрикою Чекановського. Проведемо обчислювальні експерименти, щоб встановити наскільки чутлива запропонована метрика до врахування спорідненості категорій.

Виконаємо 20 серій експериментів. В кожній серії досліджуються розподіли з однаковою кількістю категорій m. Кількість категорій від серії до серії збільшується від 3 до 60 з кроком 3. В кожній серії розраховується схожість 3000 пар об'єктів *X* та *Y*. Належності об'єктів *X* і *Y* та матрицю спорідненості категорій згенеруємо випадково з використанням узагальненого розподілу Парето. Вибір цього розподілу зумовлено тим, що в практичних задачах тематичного моделювання розподіл схожості категорій є подібними до паретівського (див., наприклад, [5], [6]). Для кожного експерименту спочатку випадковим чином згенеруємо параметри розподілів. Параметр форми k згенеруємо з діапазону $[0{,}15,\ 0{,}5]$, параметр масштабу σ згенеруємо з діапазону $[0{,}001,\ 0{,}01]$, зміщення вважатимемо нульовим — $\theta = 0$. У кожному експерименті за синтезованими розподілами випадковим чином згенеруємо координати векторів *X* і *Y* та матрицю спорідненості категорій **K**. В матриці **K** максимальне значення елементів обмежимо на рівні 0,4; вектори *X* і *Y* нормалізуємо на 1. Далі розраховуємо схожість між *X* та *Y* за запропонованою метрикою (1).

Результати експериментів у формі коробчастих діаграм показано на рис. 2. З рисунка видно, що значення схожості потраплять в діапазон $[0,\ 1]$. Зі збільшенням категорій розкид результатів знижується. Медіана розподілів за $m > 3$ не залежить від кількості категорій.

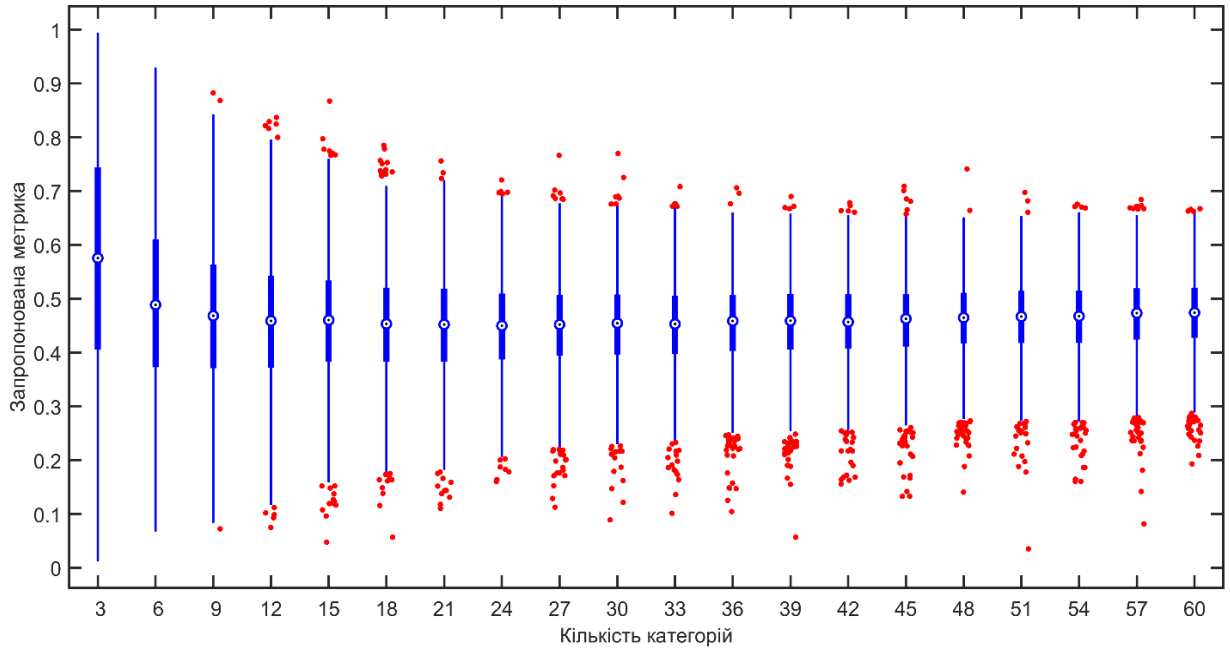

Рис. 2. Розподіли схожості за запропонованою метрикою

На рис. 3 показано коробчасті діаграми розподілу доданка $\Delta F(X,Y)$, що враховує схожість об'єктів *X* та *Y* через споріднені категорії. Медіана цієї величини збільшується від 0 до 0,06 зі збільшенням кількості категорій. Зростає також і розкид розподілів, при цьому кількість викидів спадає. Зростання медіани може свідчити про те, що зі збільшенням кількості категорій значний внесок у $\Delta F(X,Y)$ вносить велика кількість дрібних зв'язків між спорідненими категоріями. Зі збільшенням категорій кількість пар споріднених категорій зростає квадратично. Внесок багатьох пар споріднених категорій є незначним — шумовим, але сума їхніх внесків виявляється великою. Зменшення кількості викидів також є негативним чинником. Необхідність нової метрики обумовлена в першу чергу потребою виявлення специфічних випадків, з сильними перехресним впливом через споріднені категорії. А зменшення кількості викидів свідчить, що шумові спорідненості ускладнюють виявлення таких пар об'єктів.

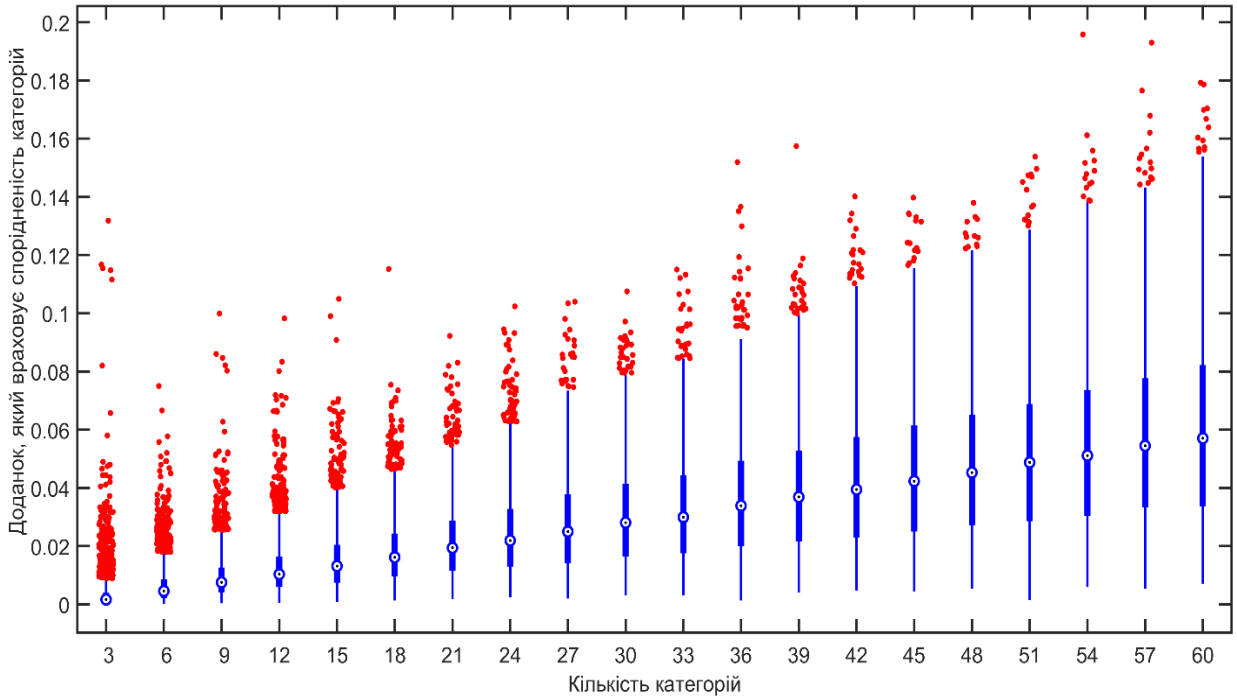

Рис. 3. Розподіли внеску споріднених категорій

Відфільтруємо шумовий внесок споріднених категорій. Вважатимемо, спорідненість шумовою, якщо коефіцієнт спорідненості менший за 0,04. Коробчасті діаграми розподілу внеску шумової спорідненості показано на рис. 4. З нього видно, що медіана шумового внеску лінійно зростає і досягає значення 0,045 за 60-ти категорій. Щодо відносного внеску шуму (рис. 5), то його медіана перевищує 10 % у серіях експериментів за великої кількості категорій — 57 та 60. В кожній серії експериментів мають місце численні випадки, коли шумовий внесок перевищує 15 %. Шумовий внесок перевищує 30 % лише у 5 випадках з 60000.

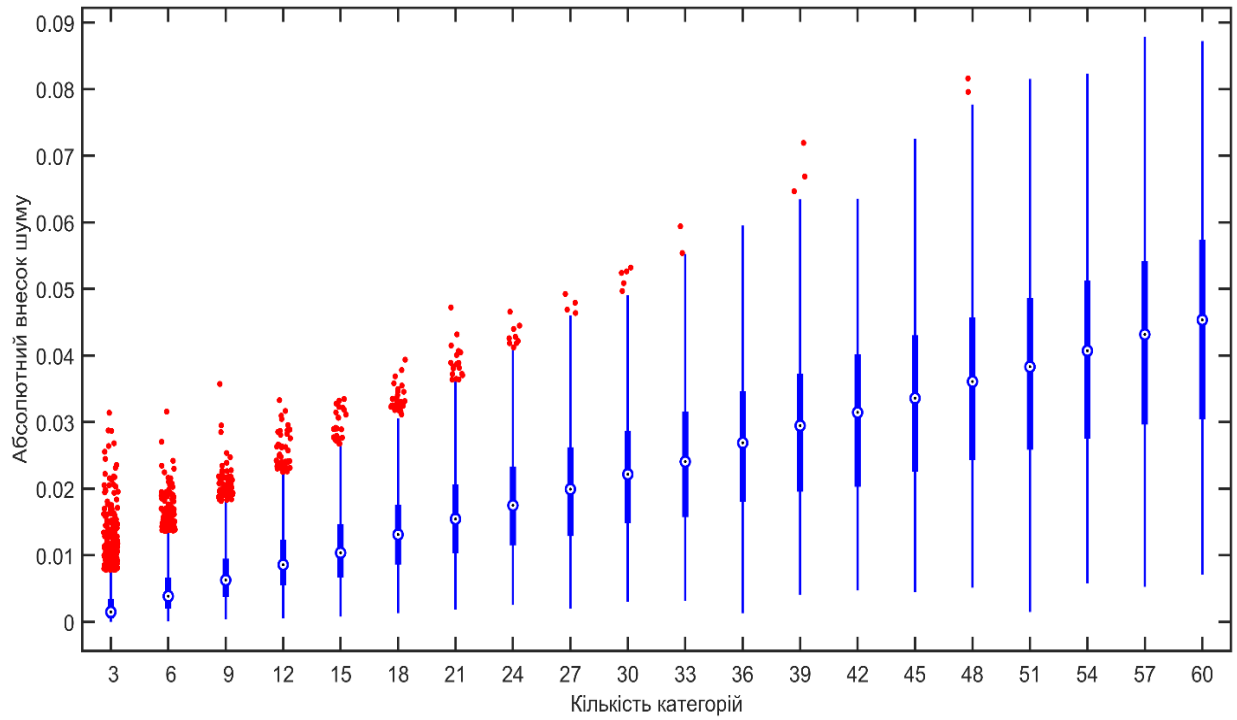

Рис. 4. Розподіли шумового внеску від врахування споріднених категорій

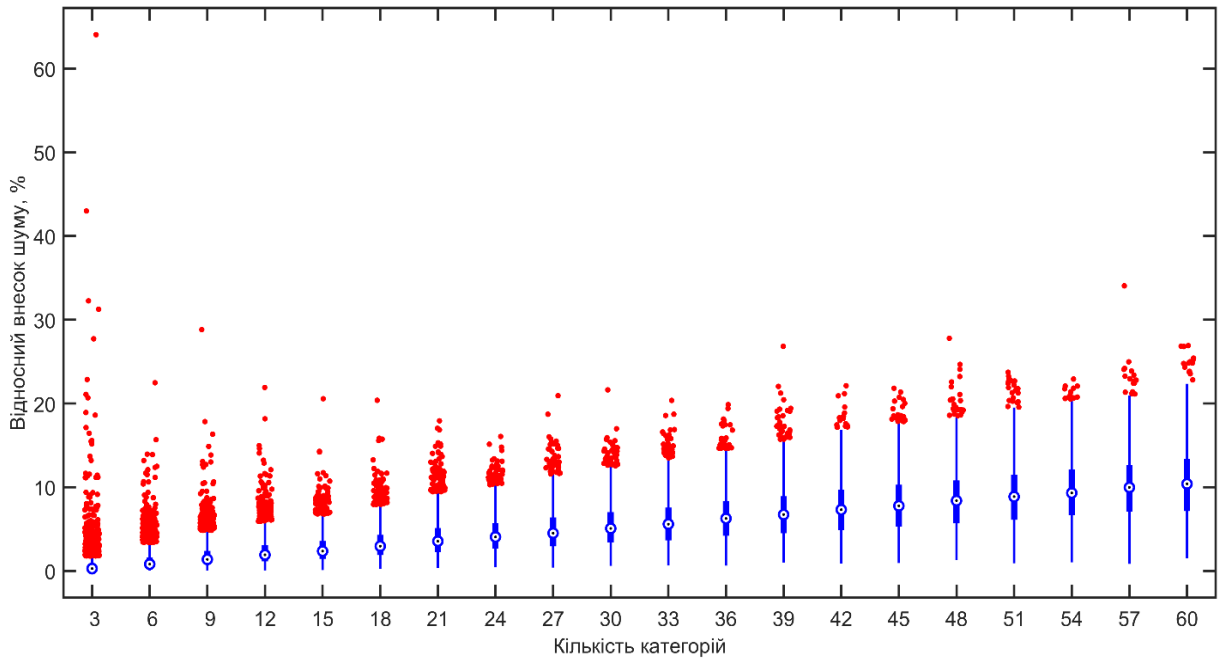

Рис. 5. Розподіли відносної величини шумового внеску від споріднених категорій

Після вилучення шумової складової коробчасті діаграми розподілу внеску споріднених категорій показано на рис. 6. Медіана цієї величини зі збільшенням кількості категорій збільшується лише з 0 до 0,01. Це в 6 разів менше ніж без вилучення шуму. При цьому спостерігається досить велика кількість викидів, що вказує на те, що метрика дозволить ідентифікувати випадки сильної взаємодії через споріднені категорії.

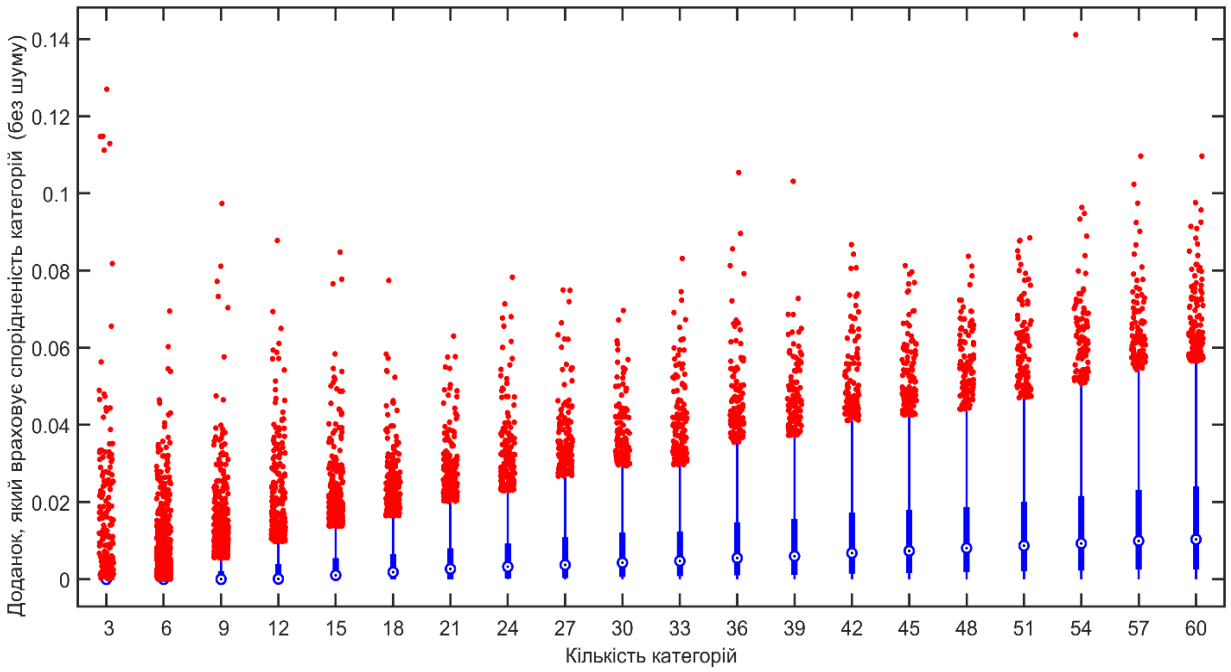

Рис. 6. Розподіли внеску споріднених категорій після фільтрації шуму

Висновки

Запропонована нова метрика схожості категоріальних розподілів, яка враховує спорідненість категорій. Метрика має дві складові. Перша складова реалізована метрикою Чекановського. Вона визначає пряму схожість розподілів за категоріями як суму перетину розподілів належностей двох об’єктів. Друга складова метрики враховує схожість об’єктів через споріднені категорії. Передбачається, що коефіцієнти спорідненості кожної пари категорій є відомими.

Обчислювальні експерименти показали, що у випадку паретовських розподілів належностей об'єктів за категоріями та паретовських розподілів коефіцієнтів спорідненості категорій, запропонована метрика приймає значення з інтервалу [0, 1]. Встановлено, що зі збільшенням кількості категорій сильно зростає внесок у запропоновану метрику доданка, який враховує спорідненість категорій. Це обумовлено тим, що квадратично зростає кількість дрібних зв'язків між спорідненими категоріями, кожний з яких додає деякий внесок у значення метрики. І хоча внесок від багатьох пар споріднених категорій є незначним — шумовим, та сума внесків виявляється великою. Для усунення цього недоліку запропоновано не враховувати шумову спорідненість категорій. Простий фільтр з пороговим обмеженням коефіцієнта спорідненості на рівні 0,04 усунув цей недолік. Після такої фільтрації шуму розподіли другого доданку метрики, який враховує спорідненість категорій, мають значну кількість викидів. Значна кількість викидів спостерігається в усіх серіях експериментів, як з малою кількістю категорій, так і з великою. Цей факт вказує на те, що запропонована метрика дозволяє легко ідентифікувати пари об'єктів, схожість яких визначається значною мірою через належність до споріднених категорій.

Запропонована метрика може використовуватися для задач класифікації, кластеризації, категоризації та тематичного моделювання, в яких під час оцінювання схожості двох об'єктів необхідно враховувати їхню належність до споріднених категорій. Такими задачами можуть бути підбір рецензентів наукових робіт, аналіз схожості текстових документів, ідентифікація поз людей у відеоряді, кластеризація природних ареалів, формування рекомендацій в інтернет-магазинах тощо.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] N. Sebe, J. Yu, Q. Tian, and J. Amores, "A New Study on Distance Metrics as Similarity Measurement," in *2006 IEEE International Conference on Multimedia and Expo*, Toronto, Ont., 2006, pp. 533-536. https://doi.org/10.1109/ICME.2006.262443 .

[2] Wang Wen-June, "New similarity measures on fuzzy sets and on elements," *Fuzzy sets and systems*, no. 85.3, pp. 305-309, 1997. https://doi.org/10.1016/0165-0114(95)00365-7 .

[3] Cha Sung-Hyuk. "Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions," *International journal of mathematical models and methods in applied sciences*, no. 1.4, pp. 300-307, 2007.

[4] Jie Yu, Qi Tian, J. Amores, and N. Sebe, "Toward Robust Distance Metric Analysis for Similarity Estimation," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2006, pp. 316-322, https://doi.org/10.1109/CVPR.2006.310 .

[5] S. Shtovba, and M. Petrychko, "An Algorithm for Topic Modeling of Researchers Taking Into Account Their Interests in Google Scholar Profiles," in *CEUR Workshop Proceedings*, vol. 2864 "Proceedings of the Fourth International Workshop on Computer Modeling and Intelligent Systems", pp. 299-311, 2021. https://doi.org/10.32782/cmis/2864-26 .

[6] S. Shtovba, and M. Petrychko, "Jaccard Index-Based Assessing the Similarity of Research Fields in Dimensions," *CEUR Workshop Proceedings*, vol. 2533 "Proceedings of the First International Workshop on Digital Content & Smart Multimedia", pp. 117-128, 2019.

Рекомендована кафедрою комп'ютерних систем управління ВНТУ

Стаття надійшла до редакції 7.03.2023

Штовба Сергій Дмитрович — д-р техн. наук, професор, професор кафедри інформаційних технологій, e-mail: s.shtovba@donnu.edu.ua ;

Петранова Марина Юріївна — канд. фіз.-мат. наук, молодший науковий співробітник науково-дослідної лабораторії вивчення проблем штучного інтелекту, e-mail: m.petranova@donnu.edu.ua .

Донецький національний університет імені Василя Стуса, Вінниця;

Петричко Микола Володимирович — аспірант кафедри комп'ютерних систем управління, e-mail: mpetrychko@vntu.edu.ua .

Вінницький національний технічний університет, Вінниця

S. D. Shtovba[1]
M.V. Petrychko[2]
M. Yu. Petranova[1]

A Similarity Metric of Categorical Distributions that Accounts for the Kinship of Different Categories

[1] Vasyl′ Stus Donetsk National University, Vinnytsia;
[2]Vinnytsia National Technical University

Estimating a level of similarity of two objects is a common problem in pattern recognition, clustering and classification. Among these problems can be reviewer recommendation, similar text documents analysis, human pose detection in video, species distribution clustering, recommendation in internet-shops etc. In case of categorical attributes an object is described as a distribution of membership degrees over categories. Similarity metrics of such distributions are usually defined as a superposition of objects' similarities for each category. Most often it is a sum of similarities in separate categories. In addition to that each category is considered independently and in isolation from the others. Some practical problems have categories that are kinship. Therefore, it is expedient to consider objects' similarity not only directly, as a similarity between equivalent categories, but it is also necessary to consider an indirect similarity, cross-similarity through kinship categories. It is such similarity metric of two categorical distributions that accounts for the kinship of different categories is proposed in this paper. The metric has two components. The first component is defined as Czekanowski metric. It defines a direct similarity of categorical distributions as a sum of intersection of distributions' membership degrees of two objects. After the intersection the residuals are accounted for in the second component of the metric. The second metric's component is defined as element-wise product of two matrices: matrix of residuals composition from membership degrees of two categorical distributions and matrix of categories' paired kinship. It is assumed that kinship indices for each pair of categories are known. As a result, with a large number of categories the overall noisy contribution from weakly kinship categories is prominent. Therefore, it is proposed to filter the noise and account only for contribution from strongly kinship categories.

Keywords: categorical distribution, kinship categories, similarity metric, Czekanowski metric, pose detection, reviewer recommendation, generalized Pareto distribution.

Shtovba Serhiy D. — Dr. Sc. (Eng.), Professor, Professor of the Chair of Information Technologies, e-mail: s.shtovba@donnu.edu.ua ;

Petrychko Mykola V. — Post-Graduate Student, of the Chair of Computer Control Systems, e-mail: mpetrychko@vntu.edu.ua ;

Petranova Maryna Yu. — Cand. Sc. (Eng.), Junior Researcher of the Laboratory of Artificial Intelligence Problems, e-mail: m.petranova@donnu.edu.ua