УДК 004.852

С. Д. Штовба[1]
М. В. Петричко[2]

Метрика для виявлення схожих об'єктів з урахуванням спорідненості категорій

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

Анотація
Оцінювання схожості двох об'єктів – це поширена задача в розпізнаванні образів, кластеризації та класифікації. У випадку категоріальних атрибутів об'єкти описуються деяким розподілом ступенів належності за категоріями. Метрики схожості таких розподілів зазвичай являють собою суперпозицію 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.
Ключові слова: категоріальний розподіл, споріднені категорії, метрика схожості, метрика Чекановського, підбір рецензентів.

Abstract
Estimating a level of similarity of two objects is a common problem in pattern recognition, clustering and classification. 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 akin. 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 akin categories. It is such similarity metric of two categorical distributions that accounts for the kinship of different categories is proposed in this paper.
Keywords: categorical distribution, kinship categories, similarity metric, Czekanowski metric, reviewer recommendation.

Оцінювання схожості двох об'єктів – це поширена задача в розпізнаванні образів, кластеризації та класифікації. В цих задачах кожен об'єкт описується вектором атрибутів. Об'єкти можуть задаватися в метричному просторі, тоді кожен атрибут задається на числовій шкалі. Наприклад, в задачі про фішеровські ірис кожна квітка описується чотирма атрибутами, а саме, шириною і довжиною пелюстки та шириною і довжиною чашолистика. Атрибути об'єкту 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 датасету 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'єкту. Чим менше відстань між аналізованими об'єктами, тим вони більш схожі. В статті [1] проаналізовано майже 50 різних

метрик, найбільш популярними 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 зворотній 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 задачу підбору схожих науковців, наприклад, для рецензування. На підставі наукового доробку кожен 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 наведених вище науковців буде встановлено лише за їх належностями до спільної спеціальності *«Системний аналіз»*. Належності до інших категорій не враховуються тому, що вони у науковців не співпадають. Схожість між науковцями A та B визначається виключно на основі їх 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 *«Маркетингу»*. Також, *«Комп'ютерні науки»* значно ближче до *«Інформаційні системи та технології»*, 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:

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

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

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

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 є такою:

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

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 їх схожостей за кожною категорією. Схожість за категорією визначається як мінімум належностей обох об'єктів до цієї категорії. Таким чином, у метрику схожості (2) один із об'єктів вносить усе значення ступеня належності до категорії, а у другий – 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:

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

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

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

Композицію залишків представимо такою матрицею:

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

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

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

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

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) зробимо перетин двох розподілів (рис. 1). Числове значення першого доданку 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 $e(Y)=(0.2 \quad 0 \quad 0.1 \quad 0)$. 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: $\Delta F(X,Y)=0.05+0.01+0.03=0.09$. Результуюче значення схожості об'єктів X та Y за формулою (1) дорівнює $F(X,Y)=0.7+0.09=0.79$.

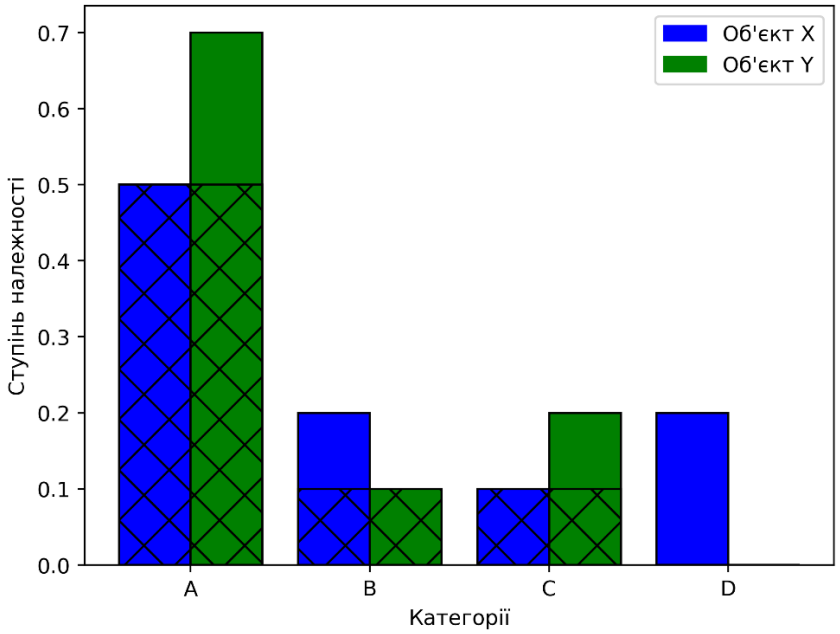

Рис. 1. Перетин двох категоріальних розподілів для розрахунку $Fit(X,Y)$

Висновки

Запропонована нова метрика схожості категоріальних розподілів, яка враховує спорідненість категорій. Метрика має дві складових. 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.

Запропонована метрика може використовуватися для задач класифікації, кластеризації, категоризації та тематичного моделювання, в яких під час оцінювання схожості двох об'єктів необхідно враховувати їх належність до споріднених категорій. Такими задачами можуть бути підбір рецензентів наукових робіт, аналіз схожості текстових документів, 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. doi: 10.1109/ICME.2006.262443.
2. Wang, Wen-June. "New similarity measures on fuzzy sets and on elements." Fuzzy sets and systems 85.3 (1997): 305-309. https://doi.org/10.1016/0165-0114(95)00365-7
3. Cha, Sung-Hyuk. “Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions.” (2007).
4. Jie Yu, Qi Tian, J. Amores and N. Sebe, "Toward Robust Distance Metric Analysis for Similarity Estimation," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, pp. 316-322, doi: 10.1109/CVPR.2006.310.

Сергій Дмитрович Штовба *– д.т.н., професор кафедри інформаційних технологій, Донецький національний університет імені Василя Стуса, м. Вінниця, e-mail: s.shtovba@donnu.edu.ua.*

Микола Володимирович Петричко *– аспірант, факультету інтелектуальних інформаційних технологій та автоматизації Вінницького національного технічного університету, м. Вінниця, e-mail: mpetrychko@vntu.edu.au.*

Shtovba Serhiy *— Professor, Information Technologies Department, Vasyl Stus’ Donetsk National University, Vinnytsia, e-mail: s.shtovba@donnu.edu.ua.*

Petrychko Mykola *— PhD student, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, email: mpetrychko@vntu.edu.ua.*