Вісник Національного технічного університету "ХПІ", 2023, № 1–2 (9–10)
ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.48: 004.94 **DOI: 10.20998/2411-0558.2023.01.04**

В. Д. ДМИТРІЄНКО, д-р техн. наук, проф., НТУ "ХПІ",
С. Ю. ЛЕОНОВ, д-р техн. наук, проф., НТУ "ХПІ",
М. В. МЕЗЕНЦЕВ, канд. техн. наук, доц., НТУ "ХПІ"

НОВІ КОМП‘ЮТЕРНІ КОМПОНЕНТИ ДЛЯ ОЦІНКИ БЛИЗЬКОСТІ ТА РОЗПІЗНАВАННЯ ДВІЙКОВИХ ОБ‘ЄКТІВ, ЩО КОДУЮТЬСЯ СИМВОЛАМИ БІНАРНОГО АЛФАВІТУ

Численні двійкові міри схожості та відстані дозволяють ефективно вирішувати різноманітні завдання розпізнавання, класифікації, оцінки близькості двійкових послідовностей тощо. Оскільки продуктивність запропонованих систем багато в чому залежить від вибору відповідних заходів та відстаней, багато дослідників витратили чимало зусиль, щоб знайти більш ефективні двійкові співвідношення для вирішення зазначених завдань. Численні двійкові співвідношення, особливо бінарні, були запропоновані та досліджені у різних галузях науки. Однак визначити одну або групу відстаней або подібних заходів ефективних при вирішенні будь-яких завдань не вдавалося. У зв'язку з цим було запропоновано низку нових методів, заснованих на використанні для навчання алгоритмів, і на зовнішній схожості співвідношень, що мають різних назв методів і запропонованих різними фахівцями, та введення попередньої обробки вихідних даних з метою визначення вагових коефіцієнтів при їх використанні. Іл.: 2.; Бібліогр.: 15 назв.

Ключевые слова: комп‘ютерні компоненти; розпізнавання; класифікація; двійкові послідовності; відстань; вагові коефіцієнти; машинне навчання.

Постановка проблеми та аналіз літератури. При оцінці близькості та розпізнаванні бінарних двійкових об‘єктів часто використовуються функції близькості Жаккара, Kulzinsky, Dice, Yula і їм подібні, а також відстані Хеммінга, функції подібності за кількостью ознак, що збігаються і т.і. [1 – 5]. Всього відомо понад сотню часто застосовуваних різних функцій близькості та подоби, відстаній для порівняння об‘єктів в геології [3], в біології [6], біометрії та генних мережах [7, 8], метеорології [9], таксономії [10, 11], при розв'язанні задач три- та чотиризначної логіки, задач пропозиційної логіки (логіки висловлювань) [12 – 15], при дослідженнях та вирішенні завдань ідентифікації, розпізнавання тощо. Загальна кількість функцій значно перевищує кількість часто застосовуваних класичних залежностей. При порівнянні пари бінарних об'єктів $J_q(j_{q1}, j_{q2}, \ldots, j_{qn})$, $J_p(j_{p1}, j_{qp2}, \ldots, j_{pn})$ з n якісними ознаками (характеристиками) зазвичай використовують співвідношення табл. 1.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Таблиця 1

Змінні для оцінки близькості бінарних об‘єктів J_q і J_p

	J_p	
J_q	1	0
1	$a = \sum_{k=1}^{n} j_{qk} j_{pk}$	$g = \sum_{k=1}^{n} (1 - j_{pk}) J_{qk}$
0	$f = \sum_{k=1}^{n} (1 - j_{qk}) j_{pk}$	$b = \sum_{k=1}^{n} (1 - j_{pk})(1 - j_{qk})$

За допомогою змінної *a* підраховується кількість ознак, які є в обох об'єктів J_q і J_p, а за допомогою змінної *b* підраховується кількість ознак, яких немає одночасно в наявності в обох об'єктів J_q і J_p. Змінна *g* необхідна для підрахунку числа ознак, які є у об'єкта J_q, але відсутні в об'єкті J_p. За допомогою змінної *f* підраховується кількість якісних ознак, які є у об'єкта J_p, але відсутні в об'єкті J_q. Сума чисел *g* + *f* вказує на кількість бінарних розрядів, у яких відрізняються об'єкти J_q і J_p, тобто. це аналог відстані Хеммінга. Сума (*a* + *b*) визначає кількість однакових бінарних розрядів в об'єктах J_q і J_p.

Різні поєднання змінних *a, b, g, f* дають можливість отримати безліч функцій близькості (афінності) для оцінки бінарних векторів [1 – 7]. Частина цих функцій має вигляд (1) – (17). Аналіз функцій (1) – (17) показує, що значна частина функцій (наприклад, (2) – (5) і (6) – (9)) мають однакову структуру і відрізняються один від одного тільки ваговими коефіцієнтами при змінних *a, b, f* і *g*. Якби ці функції застосовувалися вперше, спочатку вони повинні були для конкретних завдань випробувані за допомогою математичного моделювання, а потім за результатами моделювання виділено найбільш відповідні (ефективні) для виконання конкретного завдання. Однак через тривале використання цих функцій подібності таке моделювання не застосовується, а використовується досвід їх застосування в аналогічних випадках, у конкретних областях (геологія, біологія, біометрія тощо).

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Функції подібності (позначаються *S*) та відстані (позначаються *D*)

$$S_{INTERSECTION} = a, \tag{1}$$

$$S_{Jaccard} = \frac{a}{a+f+g}, \tag{2}$$

$$S_{CZSKANSKI} \frac{2a}{2a+f+g}, \tag{3}$$

$$S_{3W-JACCARD} = \frac{3a}{3a+f+g}, \tag{4}$$

$$S_{SOCAL\&SNKATH-I} = \frac{a}{a+2f+2g}, \tag{5}$$

$$S_{SOCAL\&MICHENER} = \frac{a+b}{a+b+f+g} = \frac{a+b}{n}, \tag{6}$$

$$S_{SOCAL\&SNKATH-II} = \frac{2(a+b)}{2a+f+g+2b}, \tag{7}$$

$$S_{ROGER\&TANIMOTO} = \frac{a+b}{a+2(f+g)+b}, \tag{8}$$

$$S_{GOWER\&EGENDRE} = \frac{a+b}{a+0{,}5(f+g)+b}, \tag{9}$$

$$S_{INNERPRODUCT} = a+b, \tag{10}$$

$$S_{RUSSELL\&RAO} = \frac{a}{a+f+g+b}, \tag{11}$$

$$D_{HAMMING} = D_{CITYBLOCK} = f+g, \tag{12}$$

$$D_{CANBERRA} = \sum_{i=1}^{n} \frac{\left| j_{pi} - j_{qi} \right|}{\left| j_{pi} + j_{qi} \right|}, \tag{13}$$

$$D_{MINKOWSKI} = \left(\sum_{i=1}^{n} \left| j_{pi} - j_{qi} \right|^{1/\lambda} \right)^{\lambda}, \tag{14}$$

$$D_{CHEBJAHEW} = \max_{i=1}^{n} \left| j_{pi} - j_{qi} \right|, \tag{15}$$

$$D_{VART} = \frac{f+g}{4(a+b+f+g)}, \tag{16}$$

$$S_{MICHAEL} \frac{4(a-g)}{(a+b)^2+(b+g)^2} = \frac{4(ab-fg)}{(a+b)^2+(b+g)^2}. \tag{17}$$

Як зазначено вище, функції подібності (афінності) та відстані не вичерпуються співвідношеннями (1) – (17), значна кількість яких і велика кількість публікацій порівняно з різними векторами з бінарним кодуванням інформації, вказують на відсутність єдиної теорії та універсальних функцій для оцінки подібності та відстаней між об'єктами, що описуються за допомогою бінарного алфавіту.

Вибір тієї чи іншої функції подібності або відстані для визначення подібності або розпізнавання об'єктів з якісними ознаками, закодованими за допомогою бінарного алфавіту здійснюється шляхом аналізу обчислювальних експериментів у *n*-мірному просторі ознак, закодованих за допомогою двійкового бінарного алфавіту та суб'єктивних переваг дослідників.

Метою статті є аналіз відомих співвідношень виду (1) – (17) та отримання на основі функцій близькості узагальнюючих властивостей відомих груп співвідношень (1) – (17) та їм подібних, та розробка нового підходу до отримання функцій афінності для бінарних об'єктів шляхом введення вагових коефіцієнтів у безлічі вихідних даних, а потім синтез нейронної мережі, що дозволяє обчислювати відповідні функції подібності та відстані.

Узагальнення співвідношень введеням вагових коефіцієнтів. Аналіз співвідношень (2) – (5) і (6) – (9), (11) показує, що кожна з двох груп співвідношень може бути замінена одним виразом з ваговими коефіцієнтами, що змінюються. Перші 4 співвідношення (2) – (5) можна записати так:

$$S = \frac{k_i a}{k_i^1 a + k_{if}^2 f + k_{ig}^3 g},\ i = \overline{1,\ 4},\ j = \overline{1,\ 4},\ g = \overline{1,\ 4}, \qquad (18)$$

де $i = \overline{1,\ 4}$, $k_1 = k_4 = 1$; $k_2 = 2$, $k_3 = 3$; $k_1^1 = k_1^4 = 1$; $k_2^1 = 2$; $k_3^1 = 3$;
$k_{1f}^2 = k_{2f}^2 = k_{3f}^2 = 1$; $k_{4f}^2 = 2$; $k_{1g}^3 = k_{2g}^3 = k_{3g}^3 = 1$; $k_{4g}^3 = 2$; $k_i, k_i^1, k_{if}^2, k_{ig}^3$, $i = \overline{1,\ 4}$ – вагові коефіцієнти.

Аналогічним чином може бути записано співвідношення, де за допомогою одного співвідношення виду (18) може бути отримане співвідношення, за допомогою вагових коефіцієнтів якого можуть бути описані формули перших одинадцяти функцій (1) – (11). Отже, вперше показано, різні функції подібності, запропоновані різними авторами, мають однаковий математичний базис. Це може бути використано для вдосконалення та застосування всієї множини аналізованих методів.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Введення вагових коефіцієнтів при сопоставленні бінарних об‘єктів. Введення вагових коефіцієнтів при сопоставленні об‘єктів J_q $(j_{q1}, j_{q2}, \ldots, j_{qn})$, J_p $(j_{p1}, j_{p2}, \ldots, j_{pn})$ приводить до появи бінарних об‘єктів $J_q(j_{q1}k^1_{q1}, j_{q2}k^1_{q2}, \ldots, j_{qn}k^1_{qn})$, $J_p(j_{p1}k^2_{p1}, j_{p2}k^2_{p2}, \ldots, j_{pn}k^2_{pn})$ та зміни виразів у табл. 1. Зокрема, змінна a обчислюється за допомогою співвідношення $a = \sum_{h=1}^{n} j_{qh} j_{ph} k^1_{qh} k^2_{ph}$, де k^1_{qh}, k^2_{ph} – вагові коефіцієнти. Аналогічно змінюються і вирази для обчислення змінних *b*, *g*, *h*. У обчислювальних експериментах зазвичай застосовувалися значення $0{,}8 \le k^1_{qh}, k^2_{ph} \le 1{,}2$.

Оскільки нейронну мережу Хеммінга можна подати у вигляді узагальненої блок-схеми (рис. 1), то із загального алгоритму функціонування мережі Хеммінга випливає, що мережа, наведена в цій статті, може обчислювати різні міри близькості, зокрема, що задаються співвідношеннями (1) – (17). При реалізації співвідношень (1) – (17) припускаємо, що для отримання нейронних мереж, що описуються цими співвідношеннями, використовуються раніше отримані нейрони та нейронні мережі, що реалізують змінні *a*, *b*, *f*, *g* та опублікованих у ряді робіт авторів, наприклад, у статтях [2, 15].

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

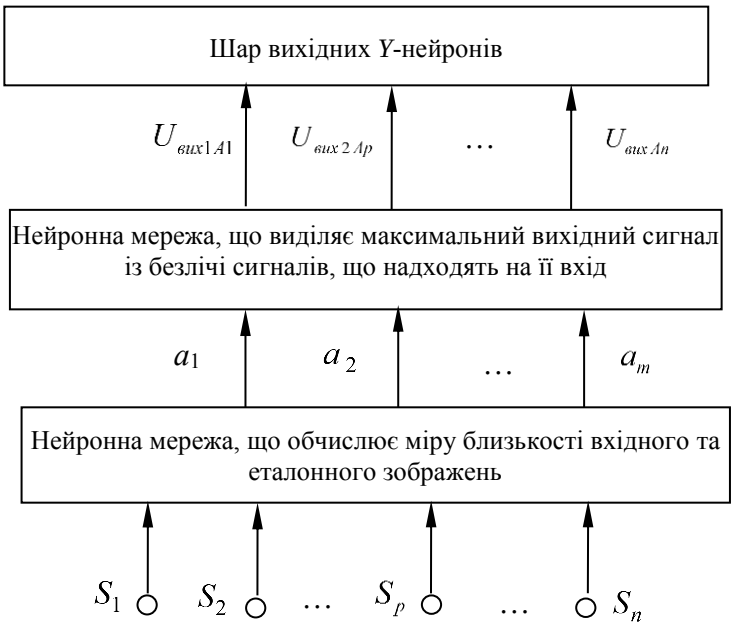

Рис. 1. Узагальнена блок-схема нейронної мережі Хеммінга

На рис. 2 прийняті такі позначення:

– Нейронна мережа, яка обчислює функцію S_{MICAEL}, наведено на рис. 2, де прийняті такі позначення: $\boxed{a}$, $\boxed{b}$, $\boxed{f}$, $\boxed{g}$, – нейронні мережі, обчислюючі допоміжні змінні *a*, *b*, *f*, *g*;

– $\boxed{\sum^d}$ $(d = \overline{1,\ 4})$ – блок алгебраїчного підсумовування;

– $\boxed{\times^k}$ $(k = \overline{1,\ 4})$ – блок добутку.

– $\boxed{:}$ – блок поділу.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

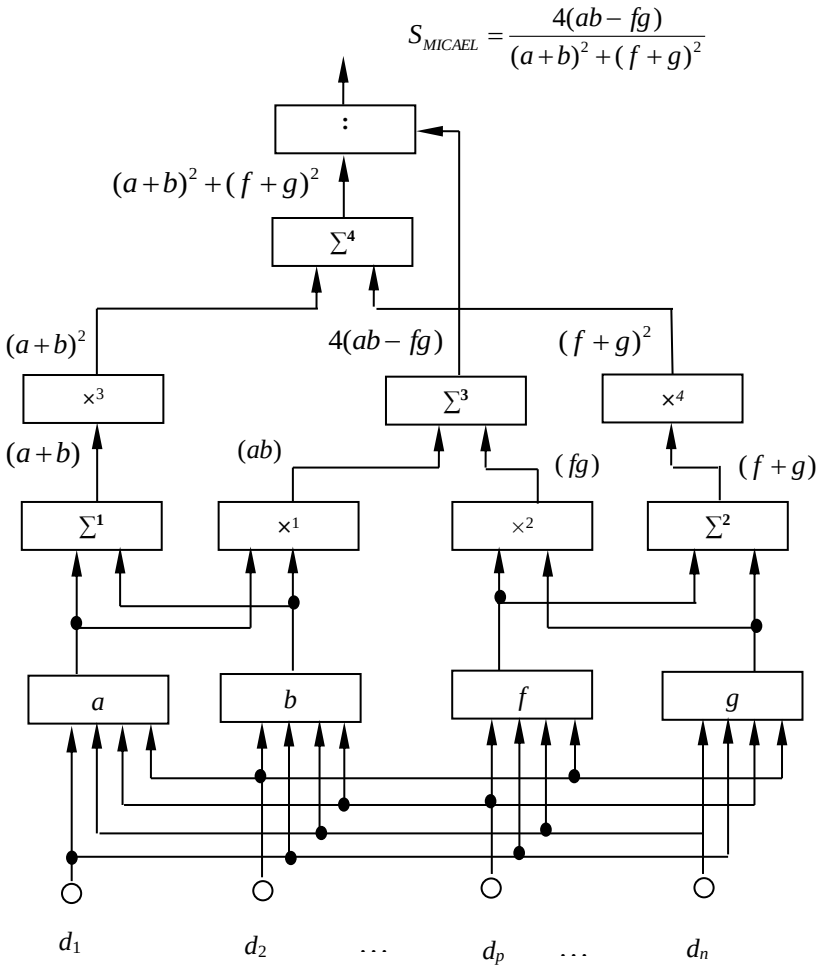

Рис. 2. Нейронна мережа, яка обчислює функцію S_{MICAEL}

Нейронна мережа, яка обчислює функцію S_{MICAEL}, складається з декількох шарів. Перший шар, включає нейронні мережі для обчислення змінних *a*, *b*, *f*, *g*. Побудова цих мереж детально описана у попередніх [2, 15] працях авторів. Другий шар нейронної мережі призначений для обчислення сум (*a* + *b*), (*f* + *g*) та добутків *ab* і *fg*. Третій та четвертий шари – для обчислення співвідношень $(a+b)^2$, $(f+g)^2$, $4(ab-fg)$, $(a+b)^2+(f+g)^2$. Функція подібності S_{MICAEL} обчислюється за допомогою блока поділу.

Висновки. Аналіз класичних співвідношень для оцінки близькості та розпізнавання бінарних двійкових об'єктів показав, що значна частина цих співвідношень відрізняється один від одного лише ваговими коефіцієнтами при відомих функціях *a*, *b*, *g*, *f*. Так, за допомогою співвідношення (18) з різними ваговими коефіцієнтами можуть бути записані чотири співвідношення (2) – (5). Якщо число одночленів у виразі (18) у чисельнику та знаменнику збільшити на одиницю, то за допомогою отриманого виразу можуть бути описані перші 11 співвідношень (1) – (11). Таким чином, вперше показано, що різні функції подібності, запропоновані різними авторами для дослідження, здавалося б абсолютно різних об'єктів, мають загальний математичний базис, що може бути в майбутньому використано для вдосконалення всієї множини методів, що розглядаються.

Ряд експериментів з вихідними даними показав, що введення нових вагових коефіцієнтів у відповідні співвідношення може бути ефективно виконано на стадії підготовки вихідних даних.

References:

1. Dmitrienko V.D., Zakovorotnyy A.Yu., and Leonov S.Yu. (2020), "Neural networks for determining affinity functions", *2020 Intenational Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Ankara, Turkey, 2020, pp. 647-652.

2. Dmitrienko V.D., Leonov S.Yu. and Zakovorotnyy A.Yu. (2020), "Computer components for proximity estimation and binary object recognition under uncertainty", *Herald of the National Technical University "KhPI". Series of "Informatics and Modeling"*. – Kharkov: NTU "KhPI". – 2020. – № 2 (4). – P. 58 – 76.

3. Michalski R.S., Stepp R.S., and Dilay (1981), "A recent advance in data analysis: chustering objects into classes characterized by cojunctive concepts", *Invited chapter in the book Progress in Pattern Recognition*, Vol. 1, North-Holland Publishing Company, Amsterdam-NewYork-Oxford, pp. 33-49.

4. Choi Send-Seok, Cha Sung-Hyuk, Tappert Charters C. (2010), "A survey of Binary Similarity and Distance Measures", *Systemics, Cybernetics and Informatics*, Vol. 8, pp. 43-48.

5. Hubalek Z. (1982), "Coefficients oj Association and Similarity, Based on Binary (Presence-Absence) Data", *An Evaluation, Biological Reviews*, Vol. 57, No. 4, pp. 669-689.

6. Willett P. (2003), Similarity-based approaches to virtual screening", *Biochemical Society Transactions*, Vol. 31, pp. 603-606.

7. Michael H. (1976), Binary Coefficients: A theoretical and empirical study", *Mathermatical Geolology*, Vol. 8, No. 2.

8. Dunn G., Everitt B.S. (1982), "An Introduction to Mathematical Taxonomy", *Cambridge University Pree*.

9. Fincly J.P. (1924), 'Tornado prediction", *The American Meteorological Journal*, 1, pp. 25-88.

10. Goodman L.A., Kruskal W.H (1963), "Measures of association for cross classification III. Approximate samling theary", *Journal of the American Statistical Association*, 58, pp. 310-364.

11. Sneath P.H.A., Socal R.R. (1973), "Numerical Taxonomy: The Principles and Practice of Numerical Classification", *W.H. Freeman and Company*, San-Francisco.

12. Tomova N.E. (2009), "The emergence of three-valued logics: logical and philosophical analysis*", Bulletin of Moskow University. Series 7, Philosophy*, Moskow, MSU, pp. 68-74.
13." Four-valued semantics for relevant logics (and some their rivals)", *Journal of Philosophical Logic*, 24 (2), pp. 139-160.
14. Burdyuk V.Ya. (2011), "Propositional trust logic", *Journal Cybernetics and Systems Analysis*, 3, pp. 182-187.
15. Dmitrienko V.D., Leonov S.Yu. (2019), "A neural network using a scalar product and defining several solutions" // Herald of the NATIONAL Tachnical University "KhPI". Series "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2019. – No. 28 (1353). – P. 68-82.

Статтю представив д.т.н., проф. Національного технічного університету "Харківський політехнічний інститут" В.І. Носков.

Поступила (received) 06.08.2023

Dmitrienko Valerii, Dr. Tech. Sci., Professor
National Technical University "Kharkiv Polytechnic Institute"
Str. Kirpicheva, 2, Kharkiv, Ukraine, 61002
Tel.: +38 (057) 707-61-98, e-mail: valdmitrienko@gmail.com
ORCID ID: 0000-0003-2523-595X

Leonov Sergey, Dr. Tech. Sci., Professor
National Technical University "Kharkiv Politechnical Institute"
Str. Kirpichova, 2, Kharkov, Ukraine, 61002
Tel.: (099) 911-911-3, e-mail: serleomail@gmail.com
ORCID ID 0000-0001-8139-0458

Mezentsev Mykola, Cand. Tech. Sci., Docent
National Technical University "Kharkiv Polytechnic Institute"
Str. Kirpicheva, 2, Kharkiv, Ukraine, 61002
Tel.: +38 (098) 859-88-98, e-mail: mykola.mezentsev@khpi.edu.ua
ORCID ID: 0000-0001-7834-2797