УДК 519.92

М. М. Биков, к. т. н., доц.;
Д. Є. Балховський, асп.;
А. Раїмі, доц.

ПОКРАЩЕНИЙ АЛГОРИТМ КЛАСТЕРНОГО АНАЛІЗУ З ВИКОРИСТАННЯМ ПОТЕНЦІАЛЬНИХ КОДІВ

Запропоновано алгоритм кластеризації великої за розміром вибірки даних, опис якої може бути представлений в різних ознакових просторах, з використанням потенціальних кодів. Його побудова базується на ідеї кластерного аналізу за методом динамічних ядер. Алгоритм передбачає попереднє визначення центрів кластерів і формування в них ядер з декількох вибіркових точок, а потім формування повного кластера шляхом пошуку множини ядер і віднесенню до них вибіркових векторів зображень таким чином, щоб отримати кластери, які мінімізують критерій узгодженості відстаней і їх рангів між точками простору. Алгоритм протестовано на стандартному файлі даних ірисів.

Вступ

Проблема кластеризації даних виникає в задачах ідентифікації станів об'єктів управління і задачах розпізнавання образів. Задача кластеризації даних з їх зображення у вигляді точок n-вимірного простору відповідає проблемі навчання без «учителя». Для випадку $n > 3$ вона є достатньо складною, оскільки людині в таких умовах неможливо уявити геометричні особливості простору даних. Для розв'язання цієї задачі розроблено низку алгоритмів, що базуються на поєднанні евристичних підходів та деякого формального критерію, наприклад, суми середньоквадратичних відхилень точок від центрів кластерів. До них можна віднести алгоритм порогової кластеризації, максимінної відстані, K-внутрішньогрупових середніх, групового голосування, ISODATA та інші [1]. Переваги та недоліки цих алгоритмів не раз обговорювались і добре відомі. Зокрема, до таких недоліків можна віднести залежність типу алгоритму кластеризації від характеру опису даних, а також значні обчислювальні затрати на їх реалізацію.

Аналіз стану досліджень та публікацій

Для подолання вказаних недоліків авторами в роботі [2] було запропоновано алгоритм кластеризації даних з використанням потенціальних кодів (*DRP-codes,* тобто кодів, що зберігають ранги відстаней), який дозволив звузити довільність у виборі початкових центрів кластерів і тим самим скоротити обчислювальні затрати, а також уніфікувати процедуру кластеризації даних, описаних в різних параметричних просторах. Крім того, обчислення рангів відстаней між точками шляхом виконання логічної операції над словами коду також скоротило обчислювальні затрати на реалізацію алгоритму.

Алгоритм кластеризації складається з таких основних етапів:

1) визначення матриці відстаней (близькостей) точок в просторі параметрів;

2) побудова *DRP*-коду точок за матрицею відстаней;

3) визначення початкових центрів кластерів;

4) віднесення всіх точок простору, що залишилися, до знайдених центрів шляхом обчислення їх рангів відстаней до всіх точок кластерів і їх порівняння;

5) переобчислення центрів кластерів з використанням векторів параметрів за формулою

$$\vec{z}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \vec{s}_j , \qquad (1)$$

де N_i — кількість точок $\vec{s}_j$ в i-му кластері.

Недоліком при застосуванні такого алгоритму є велика розрядність *DRP*-коду для кількості точок, більшої 10—11 (наприклад, розрядність для кодування 12 точок дорівнює 66, що не є сумісним з форматами двійкових слів в сучасних комп'ютерах).

Формування цілей статті

Метою даної роботи є розробка алгоритму кластеризації великої за розміром вибірки даних, опис яких може бути представлений в різних ознакових просторах, з використанням потенціальних кодів. Для досягнення цієї мети в роботі використано ідеї кластерного аналізу за методом динамічних ядер [1]. Цей метод передбачає попереднє визначення центрів кластерів і формування в них ядер з декількох вибіркових точок. Проблема кластеризації в даному випадку полягає в необхідності пошуку множини ядер E і віднесення до них вибіркових векторів зображень таким чином, що дозволяє отримати множину кластерів S2, яка мінімізує критерій

$$I(S2, E) = \sum_{\vec{x}_q \in S2_k} \sum_{\vec{z} \in E_k} \left\| \vec{s}_q - \vec{z} \right\|, \tag{2}$$

де $\vec{s}_j$ і $\vec{z}$ — вибіркова точка і центр кластера відповідно, $\left\| \vec{s}_q - \vec{z} \right\|$ — відстань між вказаними точками.

З огляду на те, що з метою уніфікації алгоритмів кластеризації точок в різних ознакових просторах (детерміністичному, ймовірнісному, нечіткому, логічному та ін.), замість відстаней між точками використовуються *DRP*-коди їх рангів, то замість критерію (2) в роботі використовується критерій

$$I(S2, E) = \sqrt{\left(d_{ij} - d_{ij}\right)^2} \Big/ \sqrt{d_{ij}^2} \,. \tag{3}$$

Цей критерій в роботах [3, 4] названий стресом (напруженістю), оскільки він показує, наскільки сильно повинні бути напруженими відстані d_{ij} між точками $\vec{s}_i$ і $\vec{s}_j$, щоб стати лінійною функцією рангів відстаней $R\left(d_{ij}\right)$. Це дає можливість, базуючись на роботах з неметричного масштабування і узагальнених методів опису даних [3] і [5] відповідно, зробити висновок про збіжність розроблюваного алгоритму.

Основна частина

Розглянемо спочатку математичні основи *DRP*-кодів. Для цього введемо ряд визначень. Нехай $\{C_i(s)\}$, $i = \overline{1, k}$ означає множину станів об'єкта або системи, де $C_i(s)$ представляє один із цих станів і є конкатенацією деяких елементів $s_j \in S$, $j = \overline{1, m}$; k — кардинальне число (потужність) множини станів, а m — число елементів у множині S, в загальному випадку $k >> m$. Вважаємо, що s_j зображений точкою або кластером n-вимірного параметричного простору. Практичними прикладами елемента s_j можуть бути звуки або фонотипи мови, примітиви зображень, вершини карт знань, нейрони нейронної мережі, образи всередині бази даних, елементи деякої системи чи об'єкта. Для зручності називатимемо елементи символами, а стани — рядками. Оскільки через вплив навколишніх завад спотворюється описовий простір символів, то рішення про стани системи (їх ідентифікація) прийматимемо за правилом мінімуму відстані

$$d\left[C_i(s), C_j(s) \right] = \min \rightarrow C_i(s) = C_j(s); \tag{4}$$

$$d\left[C_i(s), C_j(s) \right] = \sum_l d\left(s_l^i, s_l^j\right), \tag{5}$$

де $C_i = \left(s_1^i, s_2^i, \ldots, s_l^i, \ldots, s_q^i\right)$; $C_j = \left(s_1^j, s_2^j, \ldots, s_l^j, \ldots, s_q^j\right)$; $l = \overline{1, q}$; q — довжина рядка.

Опис символів у різних параметричних просторах (детерміністичному, імовірнісному, наближеному, нечіткому та ін.) породжує різноманіття ступеней подібності у них, а, отже, і алгоритмів обчислення відстаней $d\left(s_l^i, s_l^j\right)$ між символами s_l^i, s_l^j в формулі (2).

У загальновідомому підході до кодування символи s_j в запам'ятовувальному пристрої представлено у вигляді двійкових кодів, побудованих тільки з урахуванням вимоги їх розрізнення. Щоб зберігати інформацію про просторову конфігурацію символів, необхідно, крім m кодів символів, запам'ятовувати ще і $m(m-1)/2$ кодів відстаней між різними парами символів у параметричному просторі. Процедура знаходження відстаней між рядками символів потребує арифметичного підсумовування відстаней між парами символів еталонного рядка і рядка, що розпізнається. Очевидно, що загальноприйнятий підхід до кодування вимагає додаткових витрат пам'яті на зберігання $m(m-1)/2$ кодів відстаней, а також значно обмежує швидкість класифікації через втрати часу на прочитування з пам'яті цих відстаней.

Тому пропонуємо такий спосіб двійкового зображення символів рядків, за якого інформація про відстані між ними містилася б в їх кодах. При цьому, як буде далі показано, для збереження адекватності ідентифікації простір двійкових кодів повинен бути ізоморфним простору кодованих символів з точністю до рангів відстаней. Такі коди можна назвати DRP-кодами (distance rank preserving codes — кодами, що зберігають ранги відстаней), або потенціальними (за аналогією з полем електричних зарядів, в якому сила взаємодії між ними залежить від їх величин). В подальшому для зручності освітлення задачі кластеризації символи s_j будемо називати точками *n*-вимірного простору.

Код B, який зберігає ранги відстаней (DRP-код), є відображення $i \to B_i$ множини $M = \{1, 2, \ldots, m\}$ в множину $\{0,1\}^n$ двійкових послідовностей довжини n таке, що

$$\underset{i,j}{\forall}\left(R\left(d_{ij}\right) = r \Rightarrow R\left(h_{ij}\right) = r\right); \quad r = \overline{1, m_r}; \quad i, j \in M. \tag{6}$$

У виразі (6) $R\left(d_{ij}\right)$ — ранг відстаней d_{ij} між точками *i* та *j* в просторі ознак; $R\left(h_{ij}\right)$ — ранг відстані h_{ij} в просторі двійкових кодів; *r* — ціле число, конкретне значення рангу; m_r — максимальна величина рангу.

Ранговою конфігурацією простору *m* точок називають множину $(m-1)$-елементних підмножин, елементами цих підмножин є ранги відстаней, інцидентних одній і тій же точці.

Можливість побудови повного DPR-коду, тобто здатного відобразити в двійковому вигляді будь-яку рангову конфігурацію, для потреб задачі кластеризації, доведена в роботі [5]. У даному розгляді ранг відстані $R\left(d_{ij}\right)$ між двійковими словами DPR-коду знаходиться за допомогою операції логічного множення AND.

В покращеному алгоритмі на першому кроці генератор випадкових чисел вибирає 10—11 точок з усього масиву, і для них виконуються пункти 1—5 алгоритму, розробленому в [2]. Для вибраних точок будується індексна функція. На наступних кроках кожен раз вибираються нові 10—11 точок, для яких повторюються пункти 1—5, а потім визначаються ранги відстаней кожного нового центра до ядра (2—3-х точок) кожного попередньо знайденого кластера. Після цього обчислюється сума рангів до кожного ядра і вибирається мінімальна з них. Якщо вона більша половини типової, то встановлюється новий центр кластера і поновлюється індексна функція класів. Наступні кроки повторюються допоки не будуть переглянуті всі точки простору.

Описана процедура повторюється ітеративно згідно з градієнтним методом, на кожній ітерації обчислюється критерій (3). Вибирається така структура кластерів, яка відповідає мінімальному значенню критерію напруженості.

Запропонований алгоритм було реалізовано в програмному середовищі MATLAB. Його тестування на еталонному файлі «iris.dat» показали однакові за точністю результати, порівняно з алгоритмом ISODATA, але більшу в 3,7 рази швидкість роботи. До переваг розробленого алгоритму

слід віднести також можливість його застосування до даних, представлених різними ознаковими просторами.

Висновки

Розроблено алгоритм кластеризації великої за розміром вибірки даних з використанням потенціальних кодів. Однією з переваг алгоритму є те, що він може бути застосований до даних, представлених в різних ознакових просторах. Його побудова базується на ідеї кластерного аналізу за методом динамічних ядер. Алгоритм передбачає попереднє визначення центрів кластерів і формування в них ядер з декількох вибіркових точок, а потім формування повного кластера пошуком множини ядер і віднесенням до них вибіркових векторів зображень таким чином, щоб отримати кластери, які мінімізують критерій узгодженості відстаней і їх рангів між точками простору. Тестування алгоритму на еталонному файлі ірисів «iris.dat» показали однакові за точністю результати порівняно з алгоритмом ISODATA, але більшу в 3,7 рази швидкість роботи.

СПИСОК ЛІТЕРАТУРИ

1. Ту Дж., Гонсалес Р. Принципы распознавания образов. — М.: Мир, 1978. — 411 с.

2. Биков Н. М., Кузьмін І. В., Яковенко А. І. Кластеризація даних з використанням потенціальних кодів // Вісник Вінницького політехнічного інституту. — 2001. — № 6. — С. 61—64.

3. Shepard B. N. The analysis of proximities: Multidimensional Scaling with an unknown Distance Funstion // Psychometrika. — 1962. — Vol. 27. — № 2. — P. 125—140.

4. Kruskal J. B. Nonmetric multidimensional Scaling: A Numerical Method // Psychometrika. — 1964. — Vol. 36. — № 2. — P. 115—129.

5. Bykov N. M., Bykova K. N. Unified method of knowledge representation in evolutionary artificial intelligence systems // Proceedings of SPIE. — 2003. — Vol. 5098. — P. 244—253.

Биков Микола Максимович — професор, ***Балховський Дмитро Євгенович*** — аспірант.

Кафедра комп'ютеризованих систем управління, Вінницький національний технічний університет;

Раїмі Абдурахман — доцент кафедри інформатики.

Університет Анта-Діоп, Дакар, Сенегал