УДК 519.254

М.Г. Сидорова

Дніпропетровський національний університет імені Олеся Гончара

ЗАСТОСУВАННЯ АНСАМБЛІВ АЛГОРИТМІВ ДЛЯ ПІДВИЩЕННЯ СТІЙКОСТІ РЕЗУЛЬТАТІВ КЛАСТЕРИЗАЦІЇ

Здійснено огляд існуючих підходів застосування ансамблів алгоритмів у кластерному аналізі. Запропоновано інформаційну технологію підвищення стійкості результатів кластеризації даних медичного обстеження пацієнтів.

Ключові слова: *кластерний аналіз, ансамблі алгоритмів, інформаційна технологія.*

Осуществлен обзор существующих подходов применения ансамблей алгоритмов в кластерном анализе. Предложена информационная технология повышения устойчивости результатов кластеризации данных медицинского обследования пациентов.

Ключевые слова: *кластерный анализ, ансамбли алгоритмов, информационная технология.*

A review of existing cluster ensembles techniques has been conducted. The information technology of enhance the stability of the clustering results of medical examinations of patients has been offered.

Keywords: *cluster analysis, cluster ensembles, information technology.*

Вступ. На даний момент серед великої кількості підходів до розв'язання задачі кластеризації не існує універсальних методів. Кожен підхід має свої переваги та недоліки. Результат суттєво залежить від структури досліджуваних даних, вибору системи ознак, мір близькості, способів формалізації уявлень про схожість об'єктів та кластерів. Випадковий необгрунтований вибір методу може призвести до того, що отримане ним розбиття буде зовсім відмінним від природної, притаманної досліджуваним даним кластерної структури. Тому актуальними проблемами кластерного аналізу є оцінювання результатів для пошуку розбиття, що найкраще відповідає структурі даних, та розробка колективних методів кластерного аналізу (побудова ансамблю алгоритмів), що дозволить отримувати найбільш узгоджені та стійкі розв'язки.

Застосування ансамблів алгоритмів у кластерному аналізі є досить актуальним напрямом досліджень, оскільки на основі даного підходу може бути вирішено багато задач, таких як підвищення точності та стійкості результатів, зменшення простору ознак, кластеризація різнотипних даних, розпаралелювання обчислень та ін.

Аналіз публікацій. У загальному вигляді ідею застосування ансамблю алгоритмів для отримання колективного розв'язку задачі кластеризації можна подати у вигляді схеми (див. рисунок).

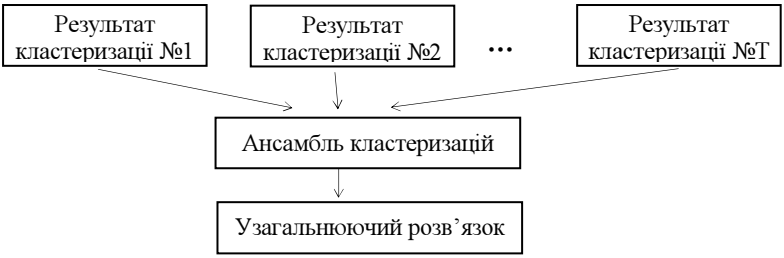

Рис. Загальна ідея ансамблевого підходу

Таким чином, можна виділити основні етапи ансамблевого підходу до кластеризації:

1) отримання індивідуальних розв'язків;

2) групування результатів (ансамбль кластеризацій);

3) визначення узагальненого (результуючого) розв'язку задачі.

Отримання індивідуальних розв'язків. На цьому етапі отримують набір $G = \{G_1, G_2, \ldots, G_T\}$ різних розв'язків кластеризації для подальшого їх об'єднання, де G_t, $t = \overline{1,T}$ – варіант розбиття (угрупування) об'єктів на K_t кластерів. Залежно від задачі способи отримання індивідуальних розв'язків можуть бути такими: застосування до вихідних даних різних алгоритмів кластеризації; застосування одного алгоритму з різними значеннями параметрів (у тому числі кількістю кластерів) та/або початкових значень (центрів, степенів приналежності тощо); розбиття простору ознак на підгрупи та почергове застосування алгоритму кластеризації до даних, що характеризуються кожною підгрупою ознак, та ін.

Групування результатів та визначення узагальненого розв'язку. Серед найбільш популярних (найчастіше вживаних у літературі та на практиці) можна виділити такі підходи:

1. *Прямий підхід*, або перевизначення кластерів. Даний підхід передбачає визначення відповідності між позначеннями кластерів різних індивідуальних розв'язків та застосування алгоритму голосування для отримання результуючого розбиття [1–4]. Найчастіше для перевизначення кластерів застосовують угорський алгоритм. Слід зазначити, що кількість кластерів у кожному вихідному розбитті та узагальненому розв'язку має бути однаковою. Ще одним недоліком даного підходу є значна складність.

2. *Графовий підхід*. Основною ідеєю даного підходу є перетворення результатів індивідуальних кластеризацій у граф (гіперграф) та застосування алгоритмів розділення графа (гіперграфа) для отримання результуючого розбиття. У роботі [5] запропоновано три графово-ансамблевих алгоритми: CSPA (The cluster-based similarity partitioning algorithm), HGPA (Hypergraph partitioning algorithm) та MCLA (Meta-clustering algorithm). Робота [6] присвячена методу HBGF (Hybrid Bipartite Graph Formulation), що полягає у побудові двостороннього графа, вершинами якого є як об'єкти вихідної вибірки, так і отримані кластери. Також графові підходи розглядаються у роботах [3; 7; 8].

3. *Матричний підхід*. Даний підхід є найбільш поширеним у застосуванні на практиці [3; 9; 10]. У загальному випадку складається із двох етапів: об'єднання результатів індивідуальних кластеризацій у вигляді матриці та визначення на її основі результуючого розбиття шляхом застосування будь-якого методу кластерного аналізу. Існує багато варіацій та різновидів даного підходу залежно від способу побудови узагальненої матриці та вибору визначального алгоритму кластеризації. У роботі [11] проведено експериментальне порівняння 24 методів даної групи на 24 різних наборах даних.

4. *Імовірнісний підхід*. Даний підхід ґрунтується на імовірнісних моделях, статистичних властивостях вихідних даних та результатів індивідуальних кластеризацій. У роботі [8] пропонується представляти результати набору окремих кластеризацій у вигляді нових атрибутів, що характеризують об'єкти вихідних даних. Задача знаходження узагальненого розв'язку розглядається як задача відновлення суміші поліноміальних розподілів та вирішується за допомогою EM-алгоритму. Ще один метод імовірнісного ансамблевого підходу BCE (Bayesian Cluster Ensembles) запропоновано у роботі [12].

У роботі [13] подається огляд основних проблем, задач та методів ансамблевого підходу кластеризації.

Досить мало робіт присвячено ансамблям нечітких методів кластеризації. Проте в останні роки все більше уваги приділяється даному напряму [7; 14; 15].

Постановка задачі. Запропонувати інформаційну технологію кластеризації із застосуванням ансамблевого підходу для підвищення стійкості результатів кластерного аналізу даних медичного обстеження пацієнтів, хворих на серцеву недостатність. Дані представлені у вигляді матриці $X = \{x_{ij}; i = \overline{1, N}, j = \overline{1, p}\}$, де N – кількість пацієнтів, p – кількість досліджуваних ознак, x_{ij} – значення j-ї ознаки, що спостерігається в i-го пацієнта, дійсне число.

Основний матеріал. Для вирішення поставленої задачі пропонується обчислювальна технологія, що складається з таких етапів:

1. Попередня обробка даних. Перш ніж застосовувати алгоритми кластерного аналізу, для підвищення їхньої точності необхідно провести попередню обробку даних, що полягає у відборі інформативних ознак та стандартизації даних.

2. Визначення набору індивідуальних роз'язків задачі кластеризації. Застосовуючи різні методи кластерного аналізу до вихідних даних, отримуємо набір угрупувань (індивідуальних розв'язків) $G = \{G_1, G_2, \ldots, G_T\}$, де $G_t = \{g_1^{(t)}, g_2^{(t)}, \ldots, g_{K_t}^{(t)}\}$, $t = \overline{1,T}$, $g_i^{(t)} = \{x_l\}$, $i = \overline{1,K_t}$, $l = \overline{1,N_i^{(t)}}$, $x_l = \{x_{lj}\}, j = \overline{1,p}$, K_t – кількість кластерів у t-му угрупуванні, $N_i^{(t)}$ – кількість об'єктів у і-му кластері t-го угрупування. $\sum_{i=1}^{K_t} N_i^{(t)} = N$, $\bigcup_{i=1}^{K_t} g_i^{(t)} = X$, $g_i^{(t)} \cap g_j^{(t)} = \varnothing$, $i, j = \overline{1,K_t}, i \neq j$.

Для того щоб набір індивідуальних роз'язків містив найбільш повну інформацію про угрупування, притаманне досліджуваним даним, пропонується застосовувати різноманітні алгоритми кластерного аналізу, а саме: алгоритми швидкої ієрархічної агломеративної кластеризації (одиничного зв'язку (ближнього сусіда), повного зв'язку (дальнього сусіда), середнього зв'язку, центрального та Уорда), графовий метод найкоротшого незамкненого шляху, енетичний алгоритм, методи розділової кластеризації К-середніх у варіантах Болла–Холла та Мак-Кіна, а також метод Forel [16; 17].

3. Оцінка якості отриманих результатів для виключення з подальшого аналізу неякісних розв'язків. Існують різноманітні функціонали та індекси якості, які дозволяють порівнювати отримані різними методами розбиття за певним обраним критерієм. Найуживанішими є сума внутрішньокластерних дисперсій за всіма ознаками, сума квадратів відстаней до центрів класів, сума внутрішньокластерних відстаней, відношення середньої внутрішньокластерної і середньої міжкластерної відстаней, індекси Каліфського – Гарабача, Данна, Беджека – Данна та ін. [17]. Для отримання багатокритеріальної оцінки пропонується агрегувати значення індексів якості за допомогою колективних методів прийняття рішень, наприклад Борда, Коупленда та ін.

4. Побудова ансамблю кластеризацій. Для отримання стійкого угрупування об'єднуємо результати індивідуальних кластеризацій у матрицю узгодженості таким чином:

а) створюємо матрицю $S=\{s_{ij}\}$; $i,j=\overline{1,N}$ та ініціалізуємо її нулями: $s_{ij}=0;\ i,j=\overline{1,N}$;

б) розглядаємо по черзі угрупування з набору індивідуальних кластеризацій $G_t : t=\overline{1,T}$. Якщо i-й та j-й об'єкти у t-му угрупуванні належать до одного кластера, то s_{ij} збільшуємо на одиницю: $s_{ij}=s_{ij}+1$, інакше значення s_{ij} залишаємо без змін;

в) зводимо елементи матриці подібності до одиничної шкали: $s_{ij}=\frac{s_{ij}}{T};\ i,j=\overline{1,N}$. Після такого перетворення s_{ij} набувають значень на відрізку від 0 до 1;

г) здійснюємо перетворення елементів матриці: $s_{ij}=1-s_{ij}$; $i,j=\overline{1,N}$;

д) визначення узагальнюючого розв'язку. Підсумкове розбиття $G'=\{g_1,g_2,\dots,g_{K'}\}$, $\bigcup_{i=1}^{K'} g_i = X$, $g_i \cap g_j = \varnothing$, $i,j=\overline{1,K'}, i\neq j$ можна отримати, застосовуючи до матриці S алгоритми кластерного аналізу, за вихідну інформацію використовують матрицю відстаней між об'єктами (наприклад, ієрархічні або графові методи).

Для визначення нечіткого розбиття на кластери, тобто коли кожен об'єкт відноситься до кожного кластера з певним степенем приналежності $\mu_{li} \in [0,1]$, $i = \overline{1,N}$, $l = \overline{1,K'}$, $\sum_{l=1}^{K'} \mu_{li} = 1$, слід до матриці S застосовувати нечіткі методи кластерного аналізу, які за вихідну інформацію використовують матрицю відстаней між об'єктами (наприклад, Уіндхема, Дева–Сена).

Розглянемо результати застосування запропонованої технології до даних медичного обстеження хворих на серцеву недостатність. Обчислювальні схеми усіх складових частин технології реалізовано в авторському програмному забезпеченні «Medisa».

Досліджувані дані були отримані за допомогою Допплер-Ехокардіографії та зібрані в Українському державному науково-дослідному інституті медико-соціальних проблем інвалідності. У 394 пацієнтів замірялися такі показники: кінцево-діастолічний розмір лівого шлуночка, кінцево-систолічний розмір лівого шлуночка, кінцево-діастолічний об'єм, кінцево-систолічний об'єм, фракція викиду та систолічний тиск легеневої артерії. Метою аналізу було розподілення пацієнтів на п'ять груп, що відповідають стадіям серцевої недостатності та відсутності хвороби (див. табл.).

Таблиця

Результати кластеризації

Метод	Індекс Каліфського – Гарабача	Індекс Данна	Індекс Беджека – Данна
Швидкий ієрархічний повного зв'язку (дальнього сусіда)	936,2	0,04	0,61
Швидкий ієрархічний Уорда	1129,1	0,05	0,57
К-середніх Болла – Холла	860,8	0,02	0,51
К-середніх Мак-Кіна	1081,7	0,01	0,64
Forel	865,8	0,06	0,57
Генетичний	773,54	0,01	0,04
Узагальнений розв'язок (швидкий ієрархічний метод середнього зв'язку)	1040,5	0,04	0,64

У таблиці представлено оцінки якості отриманих результатів після відсіювання неякісних угрупувань. Чим більше значення індекса, тим якісніший результат. В останньому рядку наведено оцінки

узагальненого розв'язку, отриманого швидким ієрархічним методом середнього зв'язку. Як бачимо, даний результат є, можливо, не найкращий, проте один із найякісніших, а головн – стійкий. Таким чином, застосовуючи ансамблевий підхід, ми зменшуємо ризик отримання неякісного розбиття в умовах невизначеності.

Для більш детального аналізу кластерної структури досліджуваних даних був отриманий нечіткий узагальнений розв'язок методом Уіндхема, що демонструє степені приналежності об'єктів до кожного з кластерів.

Висновки. У даній роботі проведено огляд існуючих підходів застосування ансамблів алгоритмів у кластерному аналізі. Запропоновано інформаційну технологію підвищення стійкості результатів та отримання нечіткого узагальненого розв'язку кластеризації даних медичного обстеження пацієнтів, хворих на серцеву недостатність. Розроблено обчислювальні схеми та програмне забезпечення, проведено практичну апробацію на реальних даних. Дана технологія може бути застосована і в інших предметних галузях.

Бібліографічні посилання

1. **Dudoit S.** Bagging to improve the accuracy of a clustering procedure / S. Dudoit, J. Fridlyand // Bioinformatics oxford university. – 2003. – Vol. 19, No. 9. – P. 1090–1099.
2. **Fischer B.** Bagging for path-based clustering / B. Fischer, J. M. Buhmann // IEEE Transaction on Pattern Analysis and Machine Intelligence. –2003. – Vol. 25, No.11. – 7 p.
3. **Бирюков А. С.** Решение задач кластерного анализа коллективами алгоритмов / А.С. Бирюков, В.В. Рязанов, А.С. Шмаков // Журнал вычислительной математики и математической физики. – 2008. – Т. 48, № 1. – С. 176–192.
4. **Topchy A.** Adaptive clustering ensembles / A. Topchy, B. Minaei Bidgoli, A. K. Jain, W. Punch Proceeding International Conference on Pattern Recognition (ICPR), Cambridge, UK, 2004. – P. 272–275.
5. **Strehl A.** Cluster ensembles – a knowledge reuse framework for combining multiple partitions / A. Strehl, J. Ghosh // Journal on Machine Learning Research. – Feb. 2002. – P. 583–617.
6. **Fern X. Z.** Solving cluster ensemble problems by bipartite graph partitioning / X. Z. Fern, C. E. Brodley // Proceedings of the 21 st International Conference on Machine Learning, Canada, 2004.
7. **Ahmadzadeh M. A.** Graph Based Approach for Clustering Ensemble of Fuzzy Partitions / M. Ahmadzadeh, Z. Golestan, J. Vahidi, B. Shirazi // Journal of mathematics and computer Science –2013. – Vol. 6. – P. 154–165.

8. **Topchy A.** A mixture model for clustering ensembles / A. Topchy, A.Jain, W. Punch // SIAM International Conference on Data Mining, 2004. – P. 379–390.
9. **Fred A.** Combining Multiple Clusterings Using Evidence Accumulation / A. Fred, A. Jain // IEEE Transaction on Pattern Analysis and Machine Intelligence –2005. – Vol. 27, No. 6. – P. 835–850.
10. **Pestunov I. A.** Ensemble of Clustering Algorithms for Large Datasets / I. A. Pestunov, V. B. Berikov, E. A. Kulikova, S. A. Rylov // Optoelectronics, Instrumentation and Data Processing. – 2011, Vol. 47, No. 3. – P. 245–252.
11. **Kuncheva L. I.** Experimental comparison of cluster ensemble methods / L. I. Kuncheva, S. T. Hadjitodorov, L. P. Todorova, B. Sofia // Proc. 9th International Conference on Information Fusion, July 2006. – 7 p.
12. **Wang H.** Bayesian Clustering Ensemble / H. Wang, H. Shan, A. Banerjee // SIAM international conference on data mining (SDM). Sparks, Nevada, USA. 29April – 2 May, 2009. – P. 211–222.
13. **Reza G.** A Survey: Clustering Ensembles Techniques / G. Reza, S. Md. Nasir, I. Hamidah, M. Norwati // Proceedings of World Academy of Science, Engineering And Technology. – February 2009 – Vol. 26. – P. 636–645.
14. **Li Taoying** Fuzzy Clustering Ensemble with Selection of Number of Clusters / Taoying Li, Yan Chen // Journal Of Computers. – July 2010. – Vol. 5, №. 7. – P. 1112–1119.
15. **Avogadri R.** Fuzzy ensemble clustering based on random projections for DNA microarray data analysis / R. Avogadri, G. Valentini // Artificial Intelligence in Medicine, 2008. – P. 173–183.
16. **Мандель И. Д.** Кластерный анализ / И. Д. Мандель. – М., 1988. – 176 с.
17. **Айвазян С. А.** Классификация многомерных наблюдений / С. А. Айвазян, З. И. Бежаева, О. В. Староверов. – М., 1974. – 240 с.

Надійшла до редколегії 26.06.2013 р.