МОДЕЛЮВАННЯ ЦІЛЬОВОЇ ФУНКЦІЇ ТА ПІДКЛАСИ РОЗВ'ЯЗНИХ ЗАДАЧ У ПРОБЛЕМІ КЛАСТЕРИЗАЦІЇ

Надія ТИМОФІЄВА
МННЦ ІТіС НАН та МОН України 03022, Київ, просп. Ак. Глушкова, 40,
тел.: (044) 502 63 32, E-mail: TymNad@gmail.com

Abstract

The method of modelling of the objective function in the problem of clusterization is described with the use of theory of combinatorial optimization. Some characteristic properties of this problem, the account of which allowed to generalize the concept of complication of its decision, are considered. It is shown that a presence in her of subclass of solvable problems depends on of a structure of input data, on the chosen estimation of similarity of objects and on the structure of argument of objective function, which laying out of n-element partitioning of sets into subsets. Introduction of subclass of solvable problems after a structure of input data conducted with the use of method of modelling their structure by the functions of natural argument, one the functions is combinatorial. It is shown that after the structure of argument of objective function of parallel of calculations in the cluster problem also taken to the solvable problems.

1. Вступ

Проблема кластеризації є надзвичайно розгалуженою і має місце у найрізноманітніших галузях людської діяльності [1–4]. В залежності від галузі в літературі ця проблема ще відома як кластер-аналіз, кластерний аналіз, розпізнавання без учителя (самонавчання), стратифікація, таксономія, автоматична класифікація тощо. В подальшому користуємося терміном кластерний аналіз для визначення теорії, а щодо розв'язання певних індивідуальних задач – кластеризація.

У розпізнаванні мовних сигналів проблема кластеризації виникає при структуризації бібліотеки еталонних сигналів, при об'єднанні в однорідні кластери подібні фонеми. Останню задачу ще називають самонавчання або таксономія [3,4]. Її ототожнюють із задачею про купу каміння. У [5] запропоновано сегментацію майже періодичного сигналу шляхом розв'язання задачі кластеризації.

Незважаючи на те, що проблемою кластеризації займаються не одне десятиліття, її точної математичної постановки ще не розроблено. Як правило, формальну постановку задачі кластеризації проводять у термінах математичної статистики або з використанням термінології теорії бінарних відношень.

Нижче сформулюємо задачу кластеризації в термінах теорії комбінаторної оптимізації. Використавши запропоновану математичну постановку, уведемо поняття складності розв'язання цієї задачі.

1. Деякі характерні властивості проблеми кластеризації

Кластеризація – спосіб групування однорідних об'єктів з метою виділення кластерів або "згустків" цих об'єктів, тобто виділення таких однорідних кластерів, щоб об'єкти всередині них були схожі один на одного, а об'єкти різних кластерів – несхожі.

У кластерному аналізі виділяють два види групувань: типологічну та структурну. Для виділення однорідних об'єктів користуються такими підходами: ймовірнісно-статистичний, структурний та варіативний.

Виділення кластерів проводиться з використанням множини певних ознак (якісних і кількісних), які визначають міру близькості між об'єктами і вимірюються за такими шкалами: кількісні, якісні, довільні. В залежності від класу задач таких ознак може бути досить багато. Для кількісних шкал існують такі метрики: лінійна відстань, за допомогою якої найкращим способом виділяються "плоскі" кластери, розміщені майже на гіперплощинах, евклідова відстань є найпопулярнішою метрикою в кластерному аналізі, узагальнена степенева відстань, відстань Махаланобіса тощо. Якісні ознаки, за якими класифікуються об'єкти, для представлення у числовій формі формулюються у вигляді мір подібності. Числові значення ознак між заданими об'єктами задаються симетричними матрицями. Для кожної ознаки можуть будуватися різні матриці. Міри подібності можуть бути уведені як між елементами, так і між утвореними кластерами.

Для моделювання цільової функції необхідне чітке визначення поняття кластера. У літературі виділяють такі поняття кластера як евристичні. Якщо виділення кластерів ґрунтується на наперед заданих властивостях, то цей підхід називається процедурою прямої класифікації. Якщо уводяться інші означення кластера і оптимізація проводиться по багатьох означеннях, то процедура називається комбінованою

прямою класифікацією. Інше визначення кластера називають оптимізаційним, при якому загальне уявлення про якість кластеризації формулюється у вигляді деякого функціоналу, оптимальне значення якого відповідає найкращому розв'язку задачі. Але при цьому моделювання цільової функції перетворюється у самостійну проблему, тому що при вибраній функції виникає інша задача: обґрунтування процедури пошуку глобального чи локального оптимуму. Для кластеризації досить складно змоделювати цільову функцію таку, щоб глобальний розв'язок задовольняв дійсну мету дослідження. Оскільки реальна структура вхідних даних невідома, то необхідно уводити кілька цільових функцій. Якщо по заданих функціях одержуються дуже відмінні кластери, то структура даних не зовсім чітка. Якщо кластери по усіх вибраних цільових функціях – однакові, то ймовірно знайдено реальний розв'язок. При евристичній кластеризації цільова функція не формалізується і при розв'язанні задачі вона ураховується на інтуїтивному рівні. При оптимізаційному підході кожному визначенню кластера відповідає певна цільова функція. Третій напрям визначення кластера – апроксимаційний. Він полягає в тому, що відношення, закладені у вхідних даних, необхідно найкращим способом апроксимувати.

Отже, алгоритми, що розробляються для розв'язання задачі кластеризації, можна віднести до таких напрямків: а) алгоритми, що ґрунтуються на розпізнаванні структури вхідних даних; б) ітераційні підходи та алгоритми. Оскільки деякі автори вважають, що в підходах, які грунтуються на розпізнаванні структури вхідних даних, такого поняття як цільова функція не існує, тому сформулюємо цю задачу як задачу комбінаторної оптимізації.

2. Про аргумент цільової функції в задачі кластеризації

В подальшому поняття об'єкт назвемо елементом заданої множини. Задача кластеризації полягає у виділенні однорідних елементів у кластери, а аргументом цільової функції в ній є розбиття n-елементної множини на підмножини (розглядаємо розбиття на неперетинні класи).

Уточнимо деякі поняття. Розбиттям n-елементної множини $A = \{a_1, ..., a_n\}$ на η підмножин (блоків) назвемо множину підмножин $\rho^k = (\rho_1^k, ..., \rho_{\eta^k}^k)$ таку, що $\rho_1^k \bigcup ... \bigcup \rho_{\eta^k}^k = A$, $\rho_s^k \neq \varnothing$, $\rho_p^k \bigcap \rho_s^k = \varnothing$, $p \neq s$, $p, s \in \{1, ..., \eta^k\}$, $\eta^k \in \{1, ..., n\}$ – кількість η^k в ρ^k. Підмножина $\rho_s^k = (a_1, ..., a_{\xi_s^k})$, $a_r \in A$, $r \in \{1, ..., n\}$, може мати від 1 до n елементів ($\xi_s^k \in \{1, ..., n\}$). Верхній індекс k в ρ^k позначає порядковий номер розбиття у множині всіх можливих розбиттів Θ, $k \in \{1, ..., q\}$, q – кількість ρ^k у Θ.

Два розбиття ρ^k і ρ^i назвемо ізоморфними, якщо $\eta^k = \eta^i$, і для будь-якої підмножини $\rho_p^k \subset \rho^k$ знайдеться підмножина $\rho_s^i \subset \rho^i$, для якої $\xi_p^k = \xi_s^i$. Підмножину ізоморфних розбиттів позначимо $\Theta_{\eta^k} \subset \Theta$.

За кількістю підмножин і кількістю в них елементів розбиття ρ^k розділяються на чотири типи [6]. До першого типу відносяться ρ^k, кількість елементів у всіх підмножинах яких – різна. Кількість елементів у підмножинах ρ_s^k розбиття другого типу – однакова. В розбиття третього типу входять дві і більше підмножини, які містять один елемент. Хоча б одна підмножина повинна містити більше ніж один елемент. В розбиття четвертого типу входять дві і більше підмножини, кількість елементів у яких однакова. З них одна підмножина повинна мати порівнянно з іншими найбільше елементів.

В задачі кластеризації закономірність зміни значень цільової функції залежить від упорядкування аргументу, від структури вхідних даних та від типів розбиття $\rho^k \in \Theta$.

2. Математична постановка задачі кластеризації

Для моделювання цільової функції в задачі кластеризації необхідно а) урахувати множину ознак заданих елементів; б) для визначення подібності елементів увести міру подібності; в) визначити спосіб оцінки кластера. Розглянемо ці три чинники детальніше.

Позначимо множину ознак елементів $a_r \in A$, $r \in \{1, ..., n\}$, упорядкованою множиною $V^{(t)} = (v_{a_1}^{(t)}, v_{a_2}^{(t)}, ..., v_{a_n}^{(t)})$. Елементи $v_{a_r}^{(t)} \in V^{(t)}$ визначають часткові критерії якості, по яких оптимізується цільова функція, $t \in \{1, ..., z\}$, де z – кількість часткових критеріїв. Ці критерії задаються мірами подібності між елементами a_r множини A. Запишемо $u^{(t)}(a_l, a_r)$ елементарну міру подібності між $a_l, a_r \in A$, яка задає t-й критерій. Оскільки міри подібності можуть бути уведені як між елементами, так і між кластерами, то уведемо міру подібності $\tilde{u}^{(t)}(\rho_s^k, \rho_p^k)$ між кластерами $\rho_s^k, \rho_p^k \in \rho^k$. Числове значення мір подібності $u^{(t)}(a_l, a_r)$, яке назвемо вагами між $a_l, a_r \in A$, задамо симетричною матрицею $C^{(t)} = \left\| c_{lr}^{(t)} \right\|_{n \times n}$, де $c_{lr}^{(t)} \sim u^{(t)}(a_l, a_r)$.

Як було оговорено вище, при оптимізаційному способі визначення кластера варто уводити кілька цільових функцій. Використаємо такі способи оцінки кластера: 1) оптимізацію проводимо так, щоб сумарне значення ваг між елементами одного кластера було найбільшим; 2) оптимізацію проводимо так, щоб середнє значення ваг між елементами одного кластера було найбільшим.

Змоделюємо цільову функцію за першим способом оцінки кластера, використавши метод моделювання структури вхідних даних функціями натурального аргументу. Для k-го розбиття при обчисленні цільової функції ураховуються ваги між елементами $a_l, a_r \in A$, які знаходяться в одній підмножині. Тому уведемо симетричну (0,1)-матрицю $Q(\rho^k) = \left\| g_{lr}(\rho^k) \right\|_{n \times n}$. Якщо елементи a_l, a_r знаходяться в одній підмножині, то $g_{lr}(\rho^k) = 1$, в іншому випадку $g_{lr}(\rho^k) = 0$.

Послідовність наддіагональних елементів матриці $C^{(t)}$ за t-ю ознакою подамо числовою функцією $\varphi^{(t)}(j)|_1^m$, а матриці $Q(\rho^1)$ – комбінаторною $\beta(f(j), \rho^1)|_1^m$, яка змінюється в залежності від розбиття ρ^k, де $m = \frac{n(n-1)}{2}$, $\rho^1, \rho^k \in \Theta$. Ця функція змінюється в залежності від типу розбиттів і не залежить від ознак заданих елементів. Цільова функція для цього випадку набуде вигляду

$$F_1^{(t)}(\rho^k) = \sum_{j=1}^{m} \beta_j (f(j), \rho^k) \varphi^{(t)}(j). \qquad (1)$$

Із виразу (1) видно, що для фіксованого аргументу послідовність величин добутку значень числової і комбінаторної функцій є комбінації елементів заданої матриці. Цю послідовність назвемо варіантом розв'язку задачі.

Змоделюємо цільову функцію за другим способом оцінки кластера. Для цього визначимо кількість одиниць у комбінаторній функції для s-ї підмножини $J_s^k = \frac{\xi_s^k!}{(\xi_s^k - 2)!2!}$, $\xi_s^k > 1$. Запишемо середнє значення ваг для t-го критерію

$$F_2^{(t)}(\rho^k) = \sum_{s=1}^{\eta^k} \left(\sum_{j=1}^{m} \beta_j (f(j), \rho_s^k) \varphi^{(t)}(j) \right) \Big/ J_s^k. \quad (2)$$

Вирази (1)–(2) є інтергальними мірами подібності, які визначають постійні часткові критерії якості, якщо подібність установлюється між заданими елементами. Якщо в процесі розв'язання задачі виникає ситуація невизначеності, то уводяться змінні критерії, які ураховують подібність між кластерами. Запишемо $\Phi^{(t)}(\rho^k) = \sum_{s=1}^{\eta^k} \sum_{p=1}^{\eta^k} \tilde{u}^{(t)}(\rho_s^k, \rho_p^k)$ – інтегральну міру подібності, яка визначає t-й критерій якості між утвореними кластерами для k-го варіанту розв'язку задачі. Уведення мір подібностей між кластерами і знаходження для них інтергальної міри подібності присвячено багато робіт, наприклад [7].

Запишемо векторний критерій $F(\rho^k) = (F_1^{(1)}(\rho^k), ..., F_1^{(z)}(\rho^k), F_2^{(z+1)}(\rho^k), ...$ $..., F_2^{(2z)}(\rho^k), \Phi^{(2z+1)}(\rho^k), ..., \Phi^{(3z)}(\rho^k))$, по якому оптимізуємо задану цільову функцію. Урахування постійних часткових критеріїв ефективне, якщо виділені кластери не перетинаються. Вони можуть бути враховані в процесі розв'язання задачі лише один раз. Уведення змінних часткових критеріїв проводиться у випадку, якщо кластери перетинаються, що створює ситуацію невизначеності. Ці критерії використовуються як один раз так і багато разів в ітераційному режимі таким чином: на k-му етапі часткового розв'язку певної задачі обчислюються міри подібності між кластерами, які ураховуються для знаходження нового варіанту розв'язку задачі. Якщо векторний критерій – зважений, то для часткових критеріїв будується матриця відстаней $C = \left\| c_{lr} \right\|_{n \times n}$ між елементами $a_l, a_r \in A$, в якій враховано усі їхні ознаки, а цільова функція обчислюється за виразами (1)–(2).

Отже, задача кластеризації полягає в знаходженні такого ρ^k, для якого векторна цільова функція була б максимальною або часткові критерії $F(\rho^k)$ по відношенню один до одного не погіршували результат.

3. Підкласи розв'язних задач із класу задачі кластеризації

В літературі для деяких класів задач комбінаторної оптимізації (задача комівояжера, задача розміщення, задача про призначення) описані підкласи, що мають певну структуру вхідних даних, для яких відомий спосіб аналітичного знаходження глобального розв'язку. Ці підкласи задач називають розв'язними.

Але в кластеризації, крім кількості операцій, затрачених на знаходження глобального розв'язку, необхідно ураховувати і міри подібності, які в цій задачі відіграють основну роль і від вибору яких в значній мірі залежить сам розв'язок. В цьому випадку говоримо про складність задачі з урахуванням міри подібності. Вона оцінюється по шкалі "так" або "ні", де "так" означає, що вибраний спосіб оцінки подібності дозволяє знаходити глобальний розв'язок, а "ні" – вибраний спосіб оцінки подібності не дає жодного розв'язку.

Отже, складність розв'язання задачі кластеризації оцінюється як за кількістю затрачених на знаходження глобального розв'язку операцій так і за способом моделювання цільової функції (визначенням подібності елементів), а підкласи розв'язних задач із класу задачі кластеризації виділяються за певною ознакою подібності, за обчислювальною складністю з урахуванням структури вхідних даних і за структурою аргумента.

Виділимо підкласи розв'язних задач за певною ознакою подібності. Позначимо $u^+(a_l, a_r) = 1$ міру подібності, за допомогою якої отримуємо глобальний розв'язок при умові, що $u(a_l, a_d) = 0$ і $a_l, a_r \in \rho_s^k$, а $a_d \in \rho_p^k$. Якщо $u^-(a_l, a_r) = 0$, то за вибраною мірою подібності не знаходимо жодного розв'язку, якщо $u(a_l, a_r) \in \{1,...,\delta\}$, то вибрана міра подібності дозволяє знайти допустимий розв'язок, де δ – найменша величина міри подібності, для якої існує допустимий розв'язок. Таким чином, якщо для певної задачі $u^+(x, y) = 1$, то задача є розв'язною за ознакою подібності.

Знаходження підкласів розв'язних задач за обчислювальною складністю проводимо за допомогою методу моделювання структури вхідних даних функціями натурального аргументу. Виділимо підкласи розв'язних задач з регулярною структурою вхідних даних (функції натурального аргументу змінюються як лінійні, монотонні, унімодальні опуклі, вгнуті тощо). Для установлення зміни значень цільової функції для певного упорядкування аргументу за способом утворення варіантів розв'язку задачі їхню множину розділимо на підмножини. У першій підмножині знаходяться послідовності, значення яких вибрані з матриці, починаючи з елемента за адресою 1, у другій підмножині – починаючи з адреси 2 і т.д. Кількість таких підмножин для різних типів розбиттів – різна. Відповідно упорядковується і множина розбиттів. Якщо числова функція – дискретна лінійна неспадна або незростаюча, то найбільше значення цільової функції за першим означенням кластера для першого типу розбиття знаходиться у першій підмножині. Якщо вона змінюється як вгнута унімодальна функція, то найбільше значення цільової функції знаходиться в останній підмножині Якщо $\varphi(j)|_1^m$ змінюється як опукла унімодальна функція, то найбільше значення цільової функції знаходиться у першій підмножині.

Виділення підкласів розв'язних задач за структурою аргумента цільової функції розглянемо на прикладі задачі розпаралелювання обчислень. Для розв'язання цієї проблеми математичну модель задачі певного класу розробляють так, щоб вона природно розділялася на незалежні підзадачі, які можна розв'язувати незалежно одна від однієї. Якщо це можливо, то задача розпаралелювання є розв'язною задачею.

В задачі кластеризації цільова функція задана на скінченній множині комбінаторного характеру W, тому закономірність зміни значень цільової функції в них залежить від упорядкування комбінаторних конфігурацій у W. Оскільки множина W складається з підмножин ізоморфних розбиттів W_η, а за способом утворення варіантів розв'язку задачі множина W_η розділяється на незалежні підмножини, то генерування аргументу у виділених підмножинах проводиться незалежними процедурами. Відповідно множина значень цільової фунції також розділяється на незалежні підмножини, в кожній з яких оптимальне значення цільової функції знаходиться в паралельному режимі відомим або спеціально розробленим методом. Отже розпаралелювання обчислень в задачі кластеризації за структурою аргумента також зводиться до розв'язної задачі.

Висновок

Використання теорії комбінаторної оптимізації дає можливість виявити характерні властивості задачі кластеризації, задати цільову функцію в явному вигляді, виявити підкласи розв'язних задач за різними ознаками. Ускладнюючи структуру вхідних даних, можна для них виявляти закономірність зміни значень цільової функції, виділяти структури, для яких цільова функція змінюється однаково і розробляти для них однакові правила розв'язання задачі кластеризації.

Література

1. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. – 176 с.

2. Жамбю М. Иерархический кластер-анализ и соответствия.–М: Финансы и статистика / Пер. с англ. – 1988.– 342 с.

3. Винцюк Т.К. Анализ, распознавание и интерпретация речевых сигналов.– К.: Наукова думка, 1987. – 262 с.

4. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. К.: наук. Думка, 2004. – 545 с.

5. Дорофеюк А.А., Гучук В.В., Десова А.А., Дорофеюк Ю.А., Покровская И.В. Классификационный анализ характеристик пульсового сигнала в задачах диагностики сердечно-сосудистых заболеваний // Таврический вестник информатики и математики. – 2008. – №1. – С.152–158

6. Тимофеева Н.К. О некоторых свойствах разбиений множества на подмножества// УСиМ. – 2002. – N 5.– С. 6–23.

7. Кириченко Н.Ф., Донченко В.С. Псевдообращение в задачах кластеризации // Кибернетика и системный анализ. – 2007. – №4. – С. 73–92.