Про аргумент цільової функції в задачах кластеризації та класифікації

Тимофієва Н. К.

МННЦІТіС НАН та МОН України, пр. Глушкова 42, Київ, Україна

TymNad@gmail.com

Abstract. *Analyzed the properties of a partition n-element set into subsets, which is the argument of the objective function in the problem of clustering. This combinatorial configurations formed by a recurrent combinatorial arithmetic operator and transposition in simultaneously with all the given elements of a basic finite set. Result of solution of this problem by modelling objective function is also determined simultaneously. And so it appertain to static combinatorial optimization problems. For classification problem argument of the objective function is a partial partition of infinite set into subsets with repetitions, which formed sample elements of the base set. Result of solution of this problem is defined stages using partial objective functions. In connection with this mentioned problem appertain to the dynamic combinatorial optimization problems.*

Keywords

Classification, clustering, partitioning an n-element set into subsets, combinatorial configuration, objective function, argument of objective function.

1 Вступ

Розглядається аргумент цільової функції в задачі класифікації та кластеризації. При розв'язанні задачі кластеризації кількість кластерів і їхні характеристики невідомі, зате задано множину елементів, які необхідно розподілити по кластерах, відповідно відома їхня кількість. Тому аргументом цільової функції в ній є розбиття n-елементної множини на підмножини, які утворюються із n елементів скінченної базової множини. У класифікації характеристика класів, як правило, відома, але невідома кількість елементів, які підлягають класифікації. До того ж один і той же об'єкт може належати різним класам. Тому в цій задачі аргументом цільової функції є розбиття нескінченної базової множини на підмножини з повтореннями. Оговорені задачі відносяться як до статичних так і до динамічних задач комбінаторної оптимізації.

2 Огляд робіт

Існує багато робіт, присвячених задачам кластеризації та класифікації, наприклад [1, 2]. Вони полягають в упорядкуванні заданих об'єктів у порівнянно однорідні групи, тобто за розробленими правилами проводиться розбиття елементів заданої множини на підмножини. В літературі описано властивості розбиттів n-елементної множини на підмножини та наведено способи їхнього генерування. Але ці комбінаторні конфігурації як аргумент цільової функції в задачах комбінаторної оптимізації достатньою мірою не вивчалися. Нижче розглянемо їхні властивості для задач класифікації та кластеризації.

3 Аргумент цільової функції в задачі кластеризації

Уточнимо такі поняття як критерій і цільова функція.

Критерій – ознаки або властивості, які характеризують певний об'єкт або зв'язки між об'єктами і є вхідними даними.

Цільова функція – вираз, який формулюється на основі заданих критеріїв з урахуванням особливостей задачі, за яким обчислюється і оцінюється результат розв'язку задачі.

Як правило, цільову функцію ототожнюють з критеріями, а за аргумент цільової функції приймають вхідні дані. Але для одних і тих же критеріїв цільову функцію можна змоделювати по-різному, тобто оцінка проводиться за різними виразами і одержується різний результат. Її аргументом є комбінаторні конфігурації різних типів (перестановки, сполучення, розбиття n-елементної множини на підмножини та ін.).

За способом обчислення цільової функції виділимо задачі комбінаторної оптимізації, у яких для певного варіанту розв'язання значення цільової функції обчислюється одночасно. Такі задачі назвемо статичними. Задачі, в яких в процесі їхнього розв'язання генерується поточна інформація, за якою оцінюється результат, а пошук оптимального розв'язку проводиться поетапно з обчисленням часткових сум цільової функції, назвемо динамічними.

Розглянемо задачу кластеризації, в якій аргументом цільової функції є розбиття n-елементної базової множини A на η підмножин [3, 4]. Назвемо множину підмножин $\rho = (\rho_1, ..., \rho_\eta)$ таку, що $\rho_1 \cup ... \cup \rho_\eta = A$, $\rho_p \cap \rho_l = \varnothing$, $p \neq l$, $\rho_p \neq \varnothing$, $p, l \in \{1, ..., \eta\}$. Непуста підмножина $\rho_p = \{a_1, ..., a_{\xi_p}\}$, $a_s \in A$, $s \in \{1, ..., n\}$, може мати від 1 до n елементів ($\xi_p \in \{1, ..., n\}$). Кількість підмножин ρ_p у розбитті ρ може бути від 1 до n ($\eta \in \{1, ..., n\}$). Їхню множину позначимо Θ.

Розбиття ρ у множині Θ утворюється двома рекурентними комбінаторними операторами: або арифметичним або транспозицією.

Дійсно, утворення ρ у множині Θ можна проводити таким чином, що елемент $a_s \in \rho$ з однієї підмножини забирається, а до другої добавляється. Якщо підмножина не містить жодного елемента, то вона забирається. По необхідності утворюється нова підмножина. Іншими словами, кількість підмножин ρ_j множини ρ і кількість у кожній з них елементів визначається певним розбиттям числа. Розбиття числа n утворюється арифметичним рекурентним комбінаторним оператором. З цього випливає, що утворення розбиттів у Θ проводиться оговореним оператором.

Нескладно замітити, що при генеруванні множини Θ у деяких ρ елементи, які знаходяться в різних підмножинах, змінюють порядок їхнього слідування, тобто для утворення розбиттів необхідно, крім арифметичного рекурентного комбінаторного оператора використовувати і транспозицію.

Два розбиття ρ^k і ρ^i назвемо ізоморфними, якщо кількість їхніх підмножин однакова, і для будь-якої підмножини $\rho_p^k \subset \rho^k$ можна знайти у множині ρ^i підмножину ρ_l^i, яка не відрізняється від ρ_p^k кількістю елементів, а відрізняється самими елементами; $k, i \in \{1, ..., q\}$ – порядкові номера ρ^k і ρ^i у множині Θ, q – їхня кількість у Θ.

Підмножину $\Theta_\eta \subset \Theta$ назвемо підмножиною ізоморфних комбінаторних конфігурацій, якщо її елементи – ізоморфні комбінаторні конфігурації.

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.

Позначимо множину ознак елементів $a_s \in A$ 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, за якими оптимізується цільова функція, $t \in \{1, ..., z\}$, де z – кількість часткових критеріїв. Ці критерії задаються мірами подібності між елементами a_s множини A. Запишемо $u^{(t)}(a_s, a_r)$ елементарну міру подібності між $a_s, a_r \in A$, $s, r \in \{1, ..., n\}$, яка задає t-й критерій. Оскільки міри подібності можуть бути уведені як між елементами, так і між кластерами, то уведемо міру подібності $\tilde{u}^{(t)}(\rho_p^k, \rho_l^k)$ між кластерами $\rho_p^k, \rho_l^k \in \rho^k$. Числове значення мір подібності $u^{(t)}(a_s, a_r)$, яке назвемо вагами між $a_s, a_r \in A$, задамо симетричною матрицею $C^{(t)} = \left\| c_{sr}^{(t)} \right\|_{n \times n}$, де $c_{sr}^{(t)} \sim u^{(t)}(a_s, a_r)$.

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 $a_s, a_r \in A$, які знаходяться в одній підмножині. Тому уведемо симетричну (0,1)-матрицю $Q(\rho^k) = \left\| g_{sr}(\rho^k) \right\|_{n \times n}$. Якщо елементи a_s, a_r знаходяться в одній підмножині, то $g_{sr}(\rho^k) = 1$, в іншому випадку $g_{sr}(\rho^k) = 0$.

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

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

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

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

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 $\Phi^{(t)}(\rho^k)=\sum_{p=1}^{\eta^k}\sum_{l=1}^{\eta^k}\tilde{u}^{(t)}(\rho_p^k,\rho_l^k)$ – інтегральну міру подібності, яка визначає t-й критерій якості між утвореними кластерами для k-го варіанту розв'язку задачі.

Оскільки Θ – скінченна множина, то закономірність зміни значень цільової функції в задачі кластеризації залежить від упорядкування ρ^k в Θ [5]. Упорядкуємо ρ^k в Θ_{η^k} так, що значення цільової функції на цьому упорядкуванні в Θ_{η^k} змінюється як монотонна функція (неспадна або незростаюча). Підмножини Θ_{η^k} упорядкуємо так, що $\eta^i \le \eta^k$. Для цього упорядкування сформулюємо теореми.

Теорема 1. Якщо оптимізація в задачі кластеризації проводиться за виразом (1), то цільова функція для заданого упорядкування ρ^k в Θ – дискретна кусково-монотонна функція (відповідно неспадна або незростаюча).

Теорема 2. Якщо оптимізація в задачі кластеризації проводиться за виразом (2), то максимальні значення цільової функції у підмножинах Θ_{η^k} на заданому вище упорядкуванні ρ^k в Θ змінюються як неспадна кусково-монотонна, а мінімальні – як незростаюча кусково-монотонна функція. Початок обох функцій – розбиття, яке містить один кластер, для якого значення цільової функції – середнє значення кількості зв'язків між усіма заданими елементами.

Доведення теорем 1–2 проводиться з використанням властивостей комбінаторних функцій, які уводяться для оцінки результату на k-му кроці і значення яких не залежить від структури вхідних даних.

Як правило, одержаний за цільовими функціями (1)–(2) глобальний розв'язок у цій задачі не завжди збігається з метою дослідження. Оскільки реальна структура вхідних даних у ній – невідома, то для точного розв'язання задачі оцінку кластера варто проводити за кількома цільовими функціями.

З вищевикладеного випливає, що для задачі кластеризації розбиття n-елементної множини A на η підмножин і оцінка результату за змодельованою цільовою функцією проводиться одночасно, тобто вона є статичною.

4 Аргумент цільової функції в задачі класифікації

Розглянемо задачу класифікації. Для неї виділимо такі підзадачі:

а) задано скінченну базову множину A. Класи можуть бути як задано так і не задано. Необхідно розподілити елементи базової множини по класах так, щоб останні не перетиналися. Ця задача зводиться до задачі кластеризації;

б) задано скінченну базову множину A. Класи можуть бути як задано так і не задано. Елементи множини A розподіляються так, що один елемент може належати різним класам. В даному разі аргументом цільової функції є розбиття n-елементної множини A на η підмножин з повтореннями;

в) задано нескінченну базову множину, частина елементів якої відома, а частина визначається в процесі розв'язання задачі, тобто інформація поступає в процесі розв'язання задачі і змінюється в часі. Аргументом цільової функції в ній є часткове розбиття нескінченної множини A на η підмножин з повтореннями. В цьому разі уводиться часткова цільова функція і часткове розбиття.

Оскільки для перших двох задач розбиття утворюється із елементів скінченної множини, яке характерне для задачі кластеризації, розглянемо аргумент цільової функції для третьої задачі. Уведемо базову нескінченну множину $\tilde{A}$, в якій елементи $\tilde{a}_s$ для $s=\overline{1,n}$ задано, а для $s>n$ визначаються в процесі розв'язання задачі. З відомих елементів $\tilde{a}_r \in \tilde{A}$, $r=\overline{1,\tilde{q}}$, утворюємо часткове розбиття множини $\tilde{A}$ на η підмножин (блоків) $\tilde{\rho}=(\tilde{\rho}_1,...,\tilde{\rho}_\eta)$, $\tilde{q}>n$ – кількість відомих елементів. Тоді множина підмножин $\tilde{\rho}=(\tilde{\rho}_1,...,\tilde{\rho}_\eta)$ має такі характеристики: $\tilde{\rho}_1 \cup ... \cup \tilde{\rho}_\eta = \tilde{A}$, $\tilde{\rho}_p \cap \tilde{\rho}_l = \varnothing$ або $\tilde{\rho}_p \cap \tilde{\rho}_l \neq \varnothing$, $p \neq l$, $\tilde{\rho}_p \neq \varnothing$, $p,l \in \{1,...,\eta\}$. Непуста підмножина $\tilde{\rho}_p=\{\tilde{a}_1,...,\tilde{a}_{\xi_p}\}$ може мати від 1 до q' елементів ($\xi_p \in \{1,...,q'\}$), $\eta \in \{1,...,\tilde{q}\}$, $q'>\tilde{q}$, $\tilde{a}_r=\tilde{a}_s$ або $\tilde{a}_r \neq \tilde{a}_s$, $\tilde{a}_r,\tilde{a}_s \in \rho_p$, $r,s \in \{1,...,\xi_p\}$,. Їхню множину позначимо $\tilde{\Theta}$.

Нескладно довести, що розбиття $\tilde{\rho}$ у множині $\tilde{\Theta}$ з елементів нескінченної множини утворюється трьома рекурентними комбінаторними операторами: вибиранням, арифметичним або транспозицією.

Як правило, при моделюванні задачі класифікації аргументом цільової функції вважають вхідні дані. Але в цій задачі оцінка результату проводиться за частковими цільовими функціями, аргументом якої є часткове розбиття нескінченної множини на підмножини з повтореннями $\tilde{\rho}^k$, тобто $\tilde{F}(\tilde{\rho}^{k*}) = \underset{\tilde{\rho}^k \in \tilde{\Theta}}{extr}\, \tilde{F}(\tilde{\rho}^k)$.

В класифікації характеристика кластерів відома, об'єкти, які необхідно визначити, до якого вони класу відносяться, аналізуються не одночасно, а групами чи окремими елементами. Оскільки результат визначається не одночасно, а за частковою цільовою функцією, то задача класифікації відноситься до динамічних задач комбінаторної оптимізації.

5 Висновки

Отже, аргументом цільової функції в задачі кластеризації є розбиття n-елементної множини A на η підмножин, яке в множині Θ утворюється двома рекурентними комбінаторними операторами: або арифметичним або транспозицією. Тобто, для генерування цих комбінаторних конфігурацій користуємося алгоритмом розбиття натурального числа і генерування перестановок. При розробленні алгоритмів розв'язання задачі кластеризації необхідно враховувати, що цільові функції (1)–(2) для розглянутого упорядкування розбиттів змінюються як кусково-монотонні функції незалежно від вхідних даних, тобто в цій задачі за способом моделювання цільової функції та за структурою аргументу виникає ситуація невизначеності. Щоб одержати коректний результат варто оцінку розв'язання задачі для різних підмножин ізоморфних розбиттів проводити за додатковими критеріями. Із аналізу задачі кластеризації випливає, що вона відноситься до статичних задач.

В задачі класифікації характеристика кластерів відома, елементи, які необхідно визначити, до якого класу вони відносяться, змінюються в процесі розв'язання задачі. Для них проводиться часткове обчислення цільової функції. Із аналізу задачі класифікації випливає, що вона відноситься до динамічних задач.

Посилання

[1] Классификация и кластер/Под редакцией Дж.Вэн Райзина / Пер. с англ. – М.: Мир, 1980. – 389 с.
[2] Мандель И.Д. Кластерный анализ/ И.Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с.
[3] Тимофеева Н.К. О природе неопределенности и переменных критериях в задачах разбиения/ Н.К. Тимофеева // Проблемы управления и информатики.– 2009, № 5. – С. 88–99.
[4] Тимофеева Н.К. О некоторых свойствах разбиений множества на подмножества/ Н.К. Тимофеева // УСиМ. – 2002. – № 5. – С. 6–23.
[5] Тимофеева Н.К. Зависимость целевой функции задач комбинаторной оптимизации от упорядочения комбинаторных конфигураций / Н.К. Тимофеева // Компьютерная математика: Сб. науч. тр. – 2005. – № 2. – С. 135–146.