УДК 004.923:65.011.56
© Я. Меденець[1], О. Тимченко[1], 2012

АНАЛІЗ КРИТЕРІЇВ ЯКОСТІ КЛАСТЕРИЗАЦІЇ

Здійснено огляд функціоналів якості кластеризації. Запропоновано методи міжкластерних відстаней. Описано властивості, що характеризують якість кластеризації.

The review of the quality clusterization's functionals had been made. The methods of the distances between clusters are offered. The characteristics of the quality clusterization had been described.

1. ВСТУП

У багатьох прикладних задачах вимірювати міру подібності об'єктів набагато простіше, ніж формувати опис признаків. Завдання класифікації об'єктів на основі їх схожості один з одним, коли приналежність об'єктів будь–яким класам не задається, є завданням кластеризації.

Нехай існує навчальна вибірка $X^l = \{x_1,...,x_l\} \subset X$ і функція відстані між об'єктами $\rho(x,x')$. Потрібно розбити вибірку на підмножини, що не перетинаються – кластери, так, щоб кожен кластер складався з об'єктів, близьких по метриці ρ, а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту $x_i \in X^l$ приписується мітка (номер) кластера y_i.

Вирішення задачі кластеризації принципово неоднозначне, чому є ряд причин. По–перше, не існує однозначно найкращого критерію якості кластеризації. По–друге, число кластерів, як правило, невідоме заздалегідь і встановлюється відповідно до деякого суб'єктивного критерію. По–третє, результат кластеризації істотно залежить від метрики ρ, вибір якої, як правило, також суб'єктивний і визначається експертом.

2. МЕТА РОБОТИ

Провести аналіз найбільш важливих функціоналів якості кластеризації та запропонувати методи знаходження оптимальних міжкластер-

[1] Українська академія друкарства

них відстаней і описати параметри, які характеризують якість кластеризації.

3. ФОРМАЛЬНІ КРИТЕРІЇ ЯКОСТІ КЛАСТЕРИЗАЦІЇ

Завдання кластеризації можна звести до завдання дискретної оптимізації: необхідно так надати номери кластерів y_i об'єктам x_i, щоб значення обраного функціонала якості набуло найкращим.

Оскільки існує багато видів функціоналів якості кластеризації, тому кожен метод кластеризації можна розглядати як точний або наближений алгоритм пошуку оптимуму деякого функціонала.

Найважливішими функціоналами являються середня внутрішньокласова та середня міжкласова відстані.

Середня внутрішньокласова відстань повинна бути якнайменшою:

$$F_0 = \frac{\sum_{i<j}\left[y_i = y_j\right]\rho\left(x_i, x_j\right)}{\sum_{i<j}\left[y_i = y_j\right]} \to \min \cdot$$

Середня міжкласова відстань навпаки повинна бути найбільшою:

$$F_1 = \frac{\sum_{i<j}\left[y_i \neq y_j\right]\rho\left(x_i, x_j\right)}{\sum_{i<j}\left[y_i \neq y_j\right]} \to \max \cdot$$

Якщо алгоритм кластеризації обчислює центри кластерів $\mu_y, y \in Y$, то можна визначити функціонали більш ефективно. Сума середніх внутрішньокластерних відстаней має бути якомога меншою:

$$\Phi_0 = \sum_{y \in Y} \frac{1}{|K_y|} \sum_{y_i = y} \rho^2(x_i, \mu_y) \to \min,$$

де $K_y = \left\{x_i \in X^l \mid y_i = y\right\}$ – кластер з номером y.

В наведеній формулі беруться квадрати відстаней, а не самі відстані, оскільки, якщо ρ – евклідова метрика, то внутрішня сума в Φ_0 набуває фізичного сенсу моменту інерції кластера K_y відносно його центру мас, тобто можна розглядати кластер як матеріальне тіло, що складається з $|K_y|$ точок однакової маси.

Сума міжкласових відстаней повинна бути як можна більшою:

$$\Phi_1 = \sum_{y \in Y} \rho^2(\mu_y, \mu) \to \max,$$

де μ – центр мас усієї вибірки.

На практиці обчислюють відношення пари функціоналів, щоб врахувати як міжкласові, так і внутрішньокласові відстані:

$$\frac{F_0}{F_1} \to \min \text{ або } \frac{\Phi_0}{\Phi_1} \to \min .$$

Формальні критерії якості кластеризації найкраще простежуються на прикладі алгоритмів ієрархічної кластеризації. Дівізимні або низхідні алгоритми розбивають вибірку на дрібніші кластери. Більш поширенішими є агломеративні або висхідні алгоритми, в яких об'єкти об'єднуються у більші кластери. Реалізація даної ідеї представлена у Алгоритмі 1.

Алгоритм 1. Алгомеративна кластеризація Ланса–Уільямса

ініціалізувати множину кластерів C_1:

$$t{:=}1;\ C_t = \{\{x_1\},...,\{x_l\}\};$$

для усіх t=2,…,l, t – номер ітерації:

знайти в C_{t-1} два найближчих кластери:

$$(U,V) := \arg\min_{U \neq V} R(U,V);$$

$$R_t := R(U,V);$$

вилучити кластери U, V, додати об'єднуючий кластер $W = U \cup V$:

$$C_t := C_{t-1} \cup \{W\} / \{U,V\};$$

для усіх $S \in C_t$ обчислити відстань $R(W,S)$ по формулі Лапласа–Уільямса.

Спочатку кожен об'єкт вважається окремим кластером. Для одноелементних кластерів природним чином визначається функція відстані:

$$R(\{x\},\{x'\}) = \rho(x,x').$$

Після цього запускається процес об'єднання. На кожній ітерації замість пари самих близьких кластерів U і V утворюється новий кластер $W = U \cup V$.

Відстань від нового кластера W до будь–якого іншого кластера S обчислюється відстанню $R(U,V)$, $R(U,S)$, $R(V,S)$, котрі повинні бути відомими:

$$R(U \cup V, S) = \alpha_U R(U,S) + \alpha_V R(V,S) + +\beta R(U,V) + \gamma |R(U,S) - R(V,S)|,$$

де α_U, α_V, β, γ – числові параметри.

Дана універсальна формула узагальнює практично усі способи визначення відстані між кластерами. Вона була запропонована Лансом та Уільямсом ще у 1967р.

На практиці використовуються наступні способи обчислення відстаней $R(W, S)$ між кластерами W і S.

Для кожного з них доведена відповідність формулі Ланса–Уільямса при певних поєднаннях параметрів:

Відстань до найближчого сусіда:

$$R^{\delta}(W,S) = \min_{\omega \in W, s \in S} \rho(\omega, s);$$

$$\alpha_U = \alpha_V = \frac{1}{2};\ \beta = 0;\ \gamma = -\frac{1}{2}.$$

Відстань до найдальшого сусіда:

$$R^{\partial}(W,S) = \max_{\omega \in W, s \in S} \rho(\omega, s);$$

$$\alpha_U = \alpha_V = \frac{1}{2};\ \beta = 0;\ \gamma = \frac{1}{2}.$$

Середня відстань:

$$R^{c}(W,S) = \frac{1}{|W||S|} \sum_{\omega \in W} \sum_{s \in S} \rho(\omega, s);$$

$$\alpha_U = \frac{|U|}{|W|};\ \alpha_V = \frac{|V|}{|W|};\ \beta = \gamma = 0.$$

Відстань між центрами:

$$R^{ц}(W,S) = \rho^2\left(\sum_{\omega \in W} \frac{\omega}{|W|}, \sum_{s \in S} \frac{s}{|S|}\right);$$

$$\alpha_U = \frac{|U|}{|W|};\ \alpha_V = \frac{|V|}{|W|};\ \beta = -\alpha_U \alpha_V;\ \gamma = 0.$$

Відстань Уорда:

$$R^{У}(W,S) = \frac{|S||W|}{|S| + |W|} \rho^2\left(\sum_{\omega \in W} \frac{\omega}{|W|}, \sum_{s \in S} \frac{s}{|S|}\right);$$

$$\alpha_U = \frac{|S| + |U|}{|S| + |W|};\ \alpha_V = \frac{|S| + |V|}{|S| + |W|};\ \beta = \frac{-|S|}{|S| + |W|};\ \gamma = 0.$$

Можливих варіантів досить багато, проте задачу, який з них вважати найкращим, слід вирішувати з врахуванням додаткових властивостей, що характеризують якість кластеризації.

4. ВЛАСТИВІСТЬ МОНОТОННОСТІ

Позначимо через R_t відстань між найближчими кластерами, обраними на t–кроці для об'єднання. Кажуть, функція відстані R володіє властивістю монотонності, якщо при кожному злитті відстань між об'єднуваними кластерами лише збільшується: $R_2 \le R_3 \le ... \le R_l$.

Властивість монотонності дозволяє представити процес кластеризації у вигляді спеціального графіка – дендрограми. По вертикальній осі відкладаються об'єкти, по горизонтальній – відстані R_t. Можна довести, що якщо кластеризація володіє властивістю монотонності, то дендрограму можна побудувати так, щоб вона не мала перехресть. При цьому будь–який кластер з множини C_t представляється суцільною послідовністю точок на вертикальній осі. Якщо ж процес кластеризації йде не монотонно, то замість дендрограми отримуємо переплутаний клубок ліній.

Дендрограма дозволяє представити кластерну структуру у вигляді плоского графіка незалежно від того, яка розмірність вихідного простору. Існують й інші способи візуалізації багатовимірних даних, такі як багатовимірні шкали або карти Кохонена, але вони вносять в картину штучні завади, вплив яких досить важко оцінити.

Проте, не будь–яке поєднання коефіцієнтів в формулі Ланса–Вільямса призводить до монотонної кластеризації.

Теорема 1. (Мілліган). *Якщо виконуються наступні три умови, то кластеризація – монотонна:*

1) $\alpha_U \ge 0, \alpha_V \ge 0$;

2) $\alpha_U + \alpha_V + \beta \ge 1$;

3) $\min\{\alpha_U, \alpha_V\} + \gamma \ge 0$.

З наведених вище відстаней лише R^u не являється монотонною. Відстань Уорда відрізняється мультиплікативною поправкою, котра і робить її монотонною.

5. ВЛАСТИВІСТЬ РОЗТЯГНЕННЯ І СТИСНЕННЯ

Деякі відстані володіють властивістю розтягнення. У міру того, як кластер зростає, відстані від нього до інших кластерів збільшуються, неначе простір довкола кластера розтягується. Властивість розтягування вважається бажаною, оскільки вона сприяє чіткішому відокремленню кластерів. З іншого боку, при дуже сильному розтягуванні мож-

ливо знайти кластери там, де їх спочатку не було. Розтягуючими є відстані R^{∂} та $R^{У}$.

Деякі відстані, навпаки, володіють властивістю стиснення. Під час зростання кластера відстані від нього до інших кластерів зменшується, і здається, що простір довкола кластера стискається. Природна кластеризація при цьому розмазується. Відстань ближнього сусіда $R^{б}$ є такою, що сильно стискається.

Властивості стиснення і розтягування визначаються через відношення $R_t / \rho(\mu_U, \mu_V)$, де $R_t = R(U,V)$ – відстань між найближчими кластерами, що об'єднуються на *t*–кроці, μ_U, μ_V – центри кластерів. Якщо це відношення на кожному кроці більше одиниці, то відстань *R* є розтягуючою; якщо воно менше одиниці, то відстанню стискання. Є відстані, які не являються відстанями стискання чи розтягування, наприклад, R^{c} і $R^{ц}$. Про них говорять, що вони зберігають метрику простору.

На практиці часто застосовують гнучку відстань, яка є компромісом між методами ближнього та далекого сусіда і середньої відстані. Вона визначається одним параметром β замість чотирьох:

$$\alpha_U = \alpha_V = \frac{1-\beta}{2},\ \gamma = 0,\ \beta < 1.$$

Гнучка відстань являється відстанню стиснення при $\beta > 0$ та відстанню розтягу при $\beta < 0$. Стандартною є рекомендація: $\beta = -0,25$.

6. ВЛАСТИВІСТЬ РЕДУКТИВНОСТІ

Найважчою операцією у Алгоритмі 1 є пошук пари найближчих кластерів на кроці 3. Вона вимагає $O(l^2)$ операцій всередині основного циклу. Відповідно, для побудови усього таксономічного дерева вимагається $O(l^2)$ операцій. Це обмежує використання алгоритму вибірками довжиною в декілька сотень об'єктів.

Ідея прискорення алгоритму полягає в тому, щоб перебирати лише найближчі пари. Задається параметр δ і перебір обмежується скороченою множиною пар $\{(U,V): R(U,V) \le \delta\}$. Коли всі такі пари будуть вичерпані, параметр δ збільшується, і формується нова скорочена множина пар. Так відбувається далі до повного злиття всіх об'єктів в один кластер. Реалізація представлена в Алгоритмі 2.

Алгоритм 2. Швидка алгомеративна кластеризація на основі редуктивності

ініціалізувати множину кластерів C_l:

$t := 1$; $C_t = \{\{x_1\},...,\{x_l\}\}$;

обрати початкове значення параметра β;

обчислити $P(\delta) := \{(U,V) \mid U,V \in C_t, R(U,V) \le \delta\}$; для усіх $t = 2,...,l$, t – номер ітерації: поки $P(\delta) = \varnothing$

збільшити δ;

знайти в $P(\delta)$ пару найближчих кластерів:

$(U,V) := \underset{(U,V)\in P(\delta)}{\arg\min}\ R(U,V)$; $R_t := R(U,V)$;

вилучити кластери *U*, *V*, додати об'єднаний кластер $W = U \cup V$:

$C_t := C_{t-1} \cup \{W\} / \{U,V\}$;

для усіх $S \in C_t$ обчислити відстань *R(W,S)* по формулі Ланса–Уільямса;

якщо $R(W,S) \le \delta$ то $P(\delta) := P(\delta) \cup \{(W,S)\}$;

Неважко показати, що даний алгоритм приводить до тієї ж кластеризації, що й Алгоритм 1, якщо відстань *R* володіє властивістю редуктивності:

Визначення 1 (Брюінш). *Відстань R називається редуктивною, якщо для будь–якого* $\delta > 0$ *і будь–яких* δ*–близьких кластерів U, V об'єднання* δ*–площин U,V містить у собі* δ*–площину кластера* $W = U \cup V$*:*

$$\{S \mid R(U \cup V, S) < \delta, R(U,V) \le \delta\} \subseteq$$
$$\subseteq \{S \mid R(S,U) < \delta \text{ або } R(S,V) < \delta\}.$$

Теорема 2 (Дідо і Моро). *Якщо виконуються наступні три умови, то відстань R – редуктивна:*

1) $\alpha_U \ge 0$, $\alpha_V \ge 0$;

2) $\alpha_U + \alpha_V + \min\{\beta, 0\} \ge 1$;

3) $\min\{\alpha_U, \alpha_V\} + \gamma \ge 0$.

Порівняння умов Теореми 1 і Теореми 2 показує, що будь–яка редуктивна відстань є монотонною, отже, дозволяє відображувати про-

цес кластеризації у вигляді дендрограми. З перерахованих вище відстаней лише R^{ll} не є редуктивною.

У Алгоритмі 2 можливі різні евристичні стратегії вибору параметра δ на кроках 2 і 6.

Загальні міркування такі: якщо δ настільки велике, що $R(\delta)$ містить практично всі пари кластерів, то ми фактично повертаємося до неефективного Алгоритму 1; якщо ж значення δ є малим, то доводиться дуже часто формувати множину $P(\delta)$.

На практиці поступають таким чином: якщо число кластерів у C_t не перевищує поріг n_1, то в якості $P(\delta)$ беруть множину усіх можливих пар *(U,V)* з C_t. У інакшому випадку обирається випадковим чином n_2 відстаней *R(U,V)*, і δ вважається рівним найменшому з них.

У випадку редуктивності параметри алгоритму n_1 і n_2 впливають лише на час виконання алгоритму, але не на результат кластеризації.

7. ВИЗНАЧЕННЯ ЧИСЛА КЛАСТЕРІВ

Визначення числа кластерів найпростіше здійснювати шляхом відсікання фінальної ділянки дендрограми.

На горизонтальній осі знаходиться інтервал максимальної довжини $|R_{t+1} - R_t|$, і в якості результуючої кластеризації отримуємо множину кластерів C_t. Число кластерів рівне $K = l - t + 1$.

При необхідності можна задати обмеження на мінімальне і максимальне число кластерів $K_0 \le K \le K_1$ та обрати *t*, що задовольнятиме обмеження $l - K_1 + 1 \le t \le l - K_0 + 1$.

У багатьох прикладних задачах цікавість представляє таксономічне дерево в цілому, а визначення оптимального числа кластерів не має особливого змісту.

8. ВИСНОВКИ

Проведений аналіз показує, що точної відповіді на питання, який алгоритм кластеризації кращий, не існує. Кожна з відстаней, перерахованих вище, має свої недоліки і підходить не для всіх завдань.

Метод ближнього сусіда володіє ланцюговим ефектом, коли незалежно від форми кластера до нього приєднуються найближчі до межі об’єкти. У окремих випадках це призводить до того, що кластери розростаються подібно деревам. У залежності від завдання ця властивість може бути як корисною, так і такою, що заважає. Метод ближнього сусіда добре підходить для виділення кластерів стрічкової форми.

Метод дальнього сусіда ланцюгового ефекту не має, але на ранньому етапі може об'єднувати досить несхожі групи. Відстань між центрами мас не є монотонною і редуктивною, тому рідко використовується на практиці, хоча інтуїтивно здається «золотою серединою».

Метод Уорду виявився найкращим за результатами експериментального порівняння на наборі модельних завдань. Він частіше за інші методи встановлює інтуїтивно найкращу кластеризацію.

1. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999. 2. Кулаичев А. П. Методы и средства комплексного анализа данных. – М.: ИНФРА–М, 2006. 3. Боюн В.П. Динамическая теория информации: основы и применение. – К.: ИК НАН Украины, 2001. – 326 с. 4. Лагутин М. Б. Наглядная математическая статистика. – М.: П–центр, 2003.