УДК 519.854.33

В.П. Шало, В.І. Ляшко

СПОСІБ ФОРМУЛЮВАННЯ ЗАДАЧ КЛАСИФІКАЦІЇ ЯК ЗАДАЧ РОЗПОДІЛУ ГРАФА

Пропонується спосіб формулювання задачі класифікації як задачі розподілу графа. Він базується на аналізі спеціально побудованого марковського ланцюга. Наводяться результати розв'язання тестової задачі методом глобального рівноважного пошуку.

Метод глобального рівноважного пошуку (ГРП) [1,2] для задач цілочислового лінійного програмування з булевими змінними успішно розв'язує також задачі оптимального розподілу графа. Тому корисною здається спроба представити задачу класифікації як задачу оптимального розподілу графа.

Опишемо задачу класифікації, яка буде вивчатися у цій статті. Нехай у метричному просторі задана деяка множина точок $T = \{x_i, i = 1,...,n\}$. Потрібно розділити множину T на задану кількість підмножин T_k,

$$T_k \cap T_l = \varnothing, k \neq l, k,l = 1,...,t$$

таким чином, щоб якось враховувалися відстані $\rho(x_i,x_j)$ між точками $x_i,x_j \in T$, i,j=1,...,n. Підмножини T_k далі будемо називати таксонами. Класифікацію точок з множини T (тобто розподілення їх по таксонах) будемо вважати природною, якщо шанси точок $x_i,x_j \in T$ потрапити у один таксон тим вищі, чим менша відстань $\rho(x_i,x_j)$ між ними.

Ясно, що існує багато способів природно класифікувати точки. Наша мета – визначити ймовірне блукання по точках множини T, спостерігаючи за допомогою якого можна було б розподілити точки по таксонах, і визначити природну класифікацію.

Ймовірнісне блукання потрібно побудувати таким чином, щоб ймовірність переходу до точки залежала від відстані до неї і, внаслідок цього, її величина була б пов'язана з шансами цих точок бути в одному таксоні. Марковський ланцюг $\xi^\tau, \tau = 0,1,...$, який визначається нижче, відповідає цим вимогам.

Для точок $x_i \in T$, i=1,...,n перехідні ймовірності задамо такими формулами:

$$p_{ij} = P\{\xi^{\tau+1} = x_j | \xi^\tau = x_i\} = \frac{\exp(-\mu\rho(x_i,x_j))}{\sum\limits_{k=1,k\neq i}^{n} \exp(-\mu\rho(x_i,x_j))},$$

де m – довільна додатна величина.

Нехай π_i – стаціонарна ймовірність знаходження марковського ланцюга ξ^τ у точці x_i, i=1,...,n. Легко впевнитись у тому, що

$$\pi_i = \frac{\sum\limits_{j=1,j\neq i}^{n} \exp(-\mu\rho(x_i,x_j))}{\sum\limits_{k=1}^{n}\sum\limits_{l=1,l\neq k}^{n} \exp(-\mu\rho(x_k,x_l))}, \quad i=1,...,n.$$

Нехай множина T розбита на t таксонів T_k k=1,...,t. Слідкуючи за марковським ланцюгом ξ^τ, можна обчислити середній час θ_k чийого перебування у T_k – у таксоні. Ця величина є мірою відокремленості, компактності цього таксону і тому може слугувати критерієм при класифікації. Виведемо формулу для обчислення величини θ_k.

Визначимо марковський ланцюг η_k^τ наступним чином. Множина його станів складається з двох елементів – 0 та 1 і

$$\eta_k^\tau = \begin{cases} 1, \text{якщо } \xi^\tau \in T_k, \\ 0, \text{якщо } \xi^\tau \notin T_k. \end{cases}$$

Нехай

$$q_{11} = P\{\eta_k^{\tau+1} = 1 | \eta_k^\tau = 1\}, \; q_{10} = P\{\eta_k^{\tau+1} = 0 | \eta_k^\tau = 1\},$$

і φ_1^k – стаціонарна ймовірність знаходження марковського ланцюга η_k^τ у стані 1. Ясно, що

$$\varphi_1^k = \sum_{x_i \in T_k} \pi_i,$$

а середній час θ_k обчислюється за формулою

$$\theta_k = \frac{1}{q_{10}^k}.$$

Можна впевнитись у тому, що справедливі такі вирази:

$$q_{11}^k = \frac{\sum_{x_i \in T_k} \pi_i \sum_{x_j \in T_k} p_{ij}}{\sum_{x_i \in T_k} \pi_i}, \quad q_{10}^k = \frac{\sum_{x_i \in T_k} \pi_i \sum_{x_j \notin T_k} p_{ij}}{\sum_{x_i \in T_k} \pi_i}.$$

Вони безпосередньо випливають з такої тотожності:

$$\sum_{x_i \in T_k} \pi_i \sum_{x_j \in T_k} p_{ij} + \sum_{x_i \notin T_k} \pi_i \sum_{x_j \in T_k} p_{ij} =$$
$$= \sum_{x_j \in T_k} \sum_{x_i \in T_k} \pi_i p_{ij} + \sum_{x_j \in T_k} \sum_{x_j \notin T_k} \pi_i p_{ij} =$$
$$\sum_{x_j \in T_k} \sum_{x_j \in T_k} \pi_i p_{ij} = \sum_{x_j \in T_k} \pi_j .$$

На наш погляд, чим більше значення величини θ_k, тим більше таксон T_k відповідає природним класифікаційним поняттям. Тому має сенс формулювати задачу класифікації таким чином:

розподіленням точок по таксонах T_k, k=1,...,t, мінімізувати функцію

$$\max_{k=1,\ldots,t} \frac{\sum_{x_i \in T_k} \pi_i \sum_{x_j \notin T_k} p_{ij}}{\sum_{x_i \in T_k} \pi_i}, \qquad (1)$$

або

$$\sum_{k=1}^{t} \sum_{x_i \in T_k} \pi_i \sum_{x_j \notin T_k} p_{ij} .$$

З метою перевірки якості запропонованої цільової функції вигляду (1) був розв'язаний тестовий приклад з дослідження [3]. Множина T складалась з 21 точки, декартові координати яких наведено у таблиці.

Таблиця. Декартові координати множини T

Номер точки	x	y	Номер точки	x	y	Номер точки	x	y
1	4.3	2.4	8	8.4	3.3	15	4.7	4.3
2	5.3	2.6	9	8.0	3.0	16	4.8	4.7
3	3.6	3.0	10	5.5	3.4	17	5.4	4.9
4	5.0	3.0	11	8.0	3.5	18	6.3	4.2
5	4.7	3.2	12	5.0	4.0	19	9.7	5.4
6	7.2	3.2	13	8.5	4.0	20	9.9	5.6
7	8.2	3.2	14	4.8	4.1	21	9.0	5.9

Множину T потрібно було розбити на 2 таксони. Методом ГРП при m=0.1 ця задача розв'язувалась 10 разів (з різними початковими значеннями генератора псевдовипадкових чисел). У всіх 10 випадках було знайдено один і той же найкращий розв'язок:

T_1={1,2,3,4,5,10,12,14,15,16},

T_2={6,7,8,9,11,13,17,18,19,20,21},

цільова функція дорівнює 0.492 (=2.151, =2.031). Ці таксони виділені на рисунку. Середнє число ітерацій, необхідних для знаходження найкращого розв'язку методом ГРП, дорівнює 194.

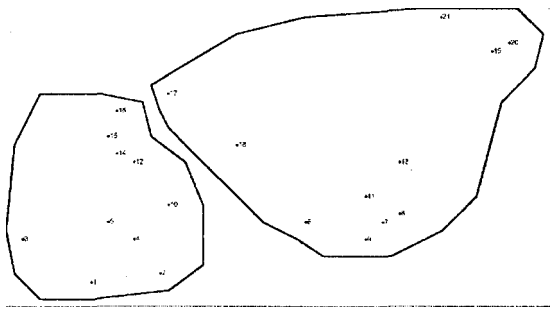

Рисунок. Таксони

1. *Шило В.П.* Метод глобального равновесного поиска //Кибернетика и системный анализ- 1999.- №1 .- С. 74-80.

2. *Шило В.П.* Результаты экспериментального исследования эффективности метода глобального равновесного поиска //Кибернетика и системный анализ,- 1999.- №2 .-С. 94-103.

3. *Котов В.Н., Терентьева Н.Г.* Классифицирование в биологии. Экспресс-метод ФЛАМЕНКО.- Киев: Наук, думка, 1993.- 68 с.

V.P. Shylo, V.I. Lyashko

A WAY OF FORMULATING CLASSIFICATION PROBLEMS AS GRAPH PARTITIONING PROBLEMS

The way of a formulation of a classification problem as graph partitioning problems is suggested. It is based on the analysis of specially constructed Markov chain. The results of the solution of the test task by global equilibrium search method are given.