

УДК 004.021

АЛГОРИТМ К-ГИПЕРПЛОСКОСТНОЙ КЛАСТЕРИЗАЦИИ ТЕКСТОВ

А.С. Корлюк

Институт кибернетики им. В.М. Глушкова НАН Украины

e-mail: arheys@gmail.com

Математическая модель

Рассмотрим математическую постановку задачи. Пусть в n -мерном пространстве расположено множество точек, каждая из которых представляет собой вектор признаков. В случае с текстовой информацией компонентами вектора являются слова (термы), а сам вектор представляет текстовый документ в пространстве признаков R^n . Необходимо построить гиперплоскость которая, разделит документы на два кластера, такую гиперплоскость можно представить в виде системы линейных уравнений:

$$Ax = b \quad (1)$$

где A - матрица коэффициентов системы уравнений размерностью $m \times n$, где m - количество документов в наборе, n - число отобранных термов в каждом документе, x - решение системы A , b - вектор свободных членов.

Будем рассматривать прямоугольную матрицу размерности $m \times n$, где $m \neq n$, так как количество документов и термов в них по умолчанию не совпадает. Однако для удобства реализации алгоритма будем считать, что количество термов в документе равно константе из диапазона 30 – 70 термов на документ.

Необходимо разделить точки пространства на два кластера, которые образуют гиперплоскости. Любая гиперплоскость разбивает R^n на две гиперплоскости. Необходимо найти такую гиперплоскость, которая будет оптимально по Евклидовому расстоянию разделять исходное пространство и документы в нем.

Множество решение (псевдорешений) $\Gamma(A, b)$ системы уравнений (1) согласно [1] имеет вид:

$$Ax = b, A \in R^{m \times n}, x \in \Gamma(A, b) = \{x : x = A^+ b + Z(A)\omega, \forall \omega \in R^n\}, \quad (2)$$

где A^+ - псевдообратная матрица, $Z(A) = E - A^+ A$ - ортогональный проектор.

Расстояние от точки $x(j) \in R^n$ до множества решений (псевдорешений) $\Gamma(A, b)$ описывается соотношением:

$$\rho^2(x(j), \Gamma(A, b)) = \|A^+(b - Ax(j))\|^2 = (b - Ax(j))^T R(A)(b - Ax(j)). \quad (3)$$

Последнее вытекает из определения $R(A^T) = A^{+T} A^+$.

Расстояние от множества точек $x(j), j = \overline{1, n}$ до $\Gamma(A, b)$

$$\rho^2(\{x : x(j), j = \overline{1, n}\}, L(A, b)) = \sum_{j=1}^n (b - Ax(j))^T R(A^T)(b - Ax(j)), \quad (4)$$

где $R(A) = A^+ A^{+T}$.

Оптимальный выбор вектора b_{opt} .

Вектор b , при фиксированной A , $x(j)$, $j = 1, \dots, n$

$$b_{opt} = \arg \min_{b \in R^n} \rho^2(\{x : x(j), j = \overline{1, n}\}, \Gamma(A, b)),$$

$$b_{opt} = A\hat{x}, \quad \hat{x} = \frac{1}{n} \sum_{j=1}^n x(j) \quad (5)$$

Значение расстояния между множеством точек $\{x : x(j), j = \overline{1, n}\}$ и плоскостью $\Gamma(A, b)$ при b_{opt}

$$\rho^2(\{x : x(j), j = \overline{1, n}\}, \Gamma(A, b_{opt})) = \sum_{j=1}^n (\hat{x} - x(j))^T A^+ A (\hat{x} - x(j)) = tr \tilde{X}^T A^+ A \tilde{X} = tr A^+ A \tilde{X} \tilde{X}^T \quad (5')$$

$$\tilde{X} = (\tilde{x}(1) \dots \tilde{x}(n)), \tilde{x}(j) = x(j) - \hat{x} \quad (6)$$

где $tr A^+ A \tilde{X} \tilde{X}^T$ - след матрицы $A^+ A \tilde{X} \tilde{X}^T$.

Из (6) следует что выражение (5') можно представить в виде сингулярного разложения

$$tr A^+ A \tilde{X} \tilde{X}^T = tr A^+ A \sum_{i=1}^r u(i) u^T(i) \lambda_i^2, \quad (7)$$

$$\tilde{X} = \sum_{i=1}^r u(i) v^T(i) \lambda_i,$$

$u^T(i) u(i) = \delta_{ij}$, $v^T(i) v(j) = \delta_{ij}$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$, где λ - сингулярные числа матрицы A .

Оптимальный выбор матрицы A_{opt} .

При b_{opt} и заданных $x(j)$, $j = \overline{1, n}$ для получения матрицы A_{opt} необходимо применить к матрице U ортогонализацию – достраивая тем самым ортогональные вектора согласно [2], до получения матрицы размером $m \times n$.

$$A_{opt} = \begin{pmatrix} u^T(m-s+1) \\ \vdots \\ u^T(m) \end{pmatrix} = \arg \min_{A \in R^{s \times m}} tr A^+ A \tilde{X} \tilde{X}^T, \quad (8)$$

$$tr A_{opt}^T A_{opt} \tilde{X} \tilde{X}^T = \sum_{j=m-s+1}^r \lambda_j^2 \quad (9)$$

λ_j^2 - собственные значения.

Полученную в (8) матрицу A_{opt} подставим в (5) и получим вектор b_{opt} для новых значений $x(j)$, $j = 1, \dots, n$.

Алгоритм К-гиперплоскостной кластеризации.

На основании вышеизложенного можно выделить следующие шаги работы алгоритма:

1. Построить начальное произвольное разбиение $\Omega_x = \{x : x(j), j = \overline{1, n}\}$ на два кластера $\Omega_x(1), \Omega_x(2)$

$$\Omega_x = \Omega_x(1) \cup \Omega_x(2), \Omega_x(1) \cap \Omega_x(2) = \emptyset$$

Разбиение реализуется с помощью метода К-средних.

2. Построить s-мерные гиперплоскости

$$A_1 x = b_1, \Gamma(A_1, b_1) \text{ для } \Omega_x(1),$$

$$A_2 x = b_2, \Gamma(A_2, b_2) \text{ для } \Omega_x(2),$$

3. Сформировать новые множества $\Omega_x(1), \Omega_x(2)$ из точек Ω_x , для которых выполняются соответствующие условия

$$\|A_1^+(b_1 - A_1 x(j))\|^2 \leq \|A_2^+(b_2 - A_2 x(j))\|^2$$

$$\|A_2^+(b_2 - A_2 x(j))\|^2 < \|A_1^+(b_1 - A_1 x(j))\|^2$$

Если условия не выполнились то перейти на 5.

4. Перейти на шаг 2.
5. Конец.

Список литературы

1. Кириченко Н.Ф., Кривонос Ю.Г., Лепеха Н.П. Синтез систем нейрофункциональных преобразований в решении задач классификации. Кибернетика и системный анализ №3, 2007г. - С. 47-57.
2. Гаращенко О.Ф., Кириченко Н.Ф. Об одном методе последовательного построения матриц ортогональных преобразований. Проблемы управления и информатики №1, 2005г. - С.75-88.