

КОНТЕЙНЕРНАЯ КЛАСТЕРИЗАЦИЯ СИГНАЛОВ И ФОРМИРОВАНИЕ ДЕРЕВА КЛАСТЕРОВ

Н.Ф. Кириченко, А.С. Корлюк

Институт кибернетики им. В.М. Глушкова НАН Украины

e-mail: arheys@gmail.com

Введение

В работах [1], [2] были предложены средства классификации ультразвуковых сигналов базирующиеся на использовании свойств многомерных сфер в L_p пространствах, которые получили название контейнеров. В данной работе идея построения контейнеров в пространстве признаков развивается в направлении формирования контейнеров в виде многомерных эллипсоидов определенных на основе взвешенных проекционных матриц. Известные свойства этих матриц позволили по новому взглянуть на решение проблемы кластеризации и классификации сигналов в пространстве информационных признаков.

Контейнерная кластеризация сигналов и формирование дерева кластеров

Пусть $x(1), \dots, x(m)$ являются точками (векторами $x(j) \in R^m$) в пространстве признаков, соответствующие различным объектам или процессам, которые необходимо кластеризовать. Например $x(1)$ - соответствует изображению буквы А, $x(2)$ - Б, $x(n)$ - Я. Необходимо по некоторому принципу подобия объединить часть этих векторов в одну совокупность (кластер), а оставшуюся во второй кластер. А затем последовательно осуществлять эти же разбиения для кластеров, организовав, таким образом, дихотомное дерево кластеризации исходных данных. Наиболее простое средство такой кластеризации представляет собой известный алгоритм К-средних [3] использующий идею объединения точек по минимальному Евклидовому расстоянию от точек до центров кластеров, т.е. этот алгоритм имеет вид

1. Определим первоначальные центроиды кластеров $x(1), \dots, x(k)$.
2. Найти ближайшие по Евклидовому расстоянию точки к каждому центроиду.
3. Пересчитать центроиды.
4. Если все точки отнесены к кластерам и центроиды не изменились на выход, иначе на шаг 2.

Здесь предлагается развитие этой идеи К-средних но с учетом вычисления расстояния по специальной мере, которую формируют сами точки или другие совокупности, возникающей в процессе выполнения алгоритма. Эта мера учитывает плотность и факторную структуру соответствующего подмножества точек, что далее положительно проявляется на качестве, как кластеризации, так и распознавании сигналов. Итак, опишем этот подход формирования кластеров. Пусть изначально рассматриваемые точки $x(j), j = \overline{1, n}$ в пространстве признаков уже разбиты на 2 кластера (например, с помощью алгоритма 2-средних) $D_1 = \{x : x(i_k), k = \overline{1, n_1}\}$ и $D_2 = \{x : x(i_k), k = \overline{1, n_2}\}$. Для этих подмножеств D_1 и D_2 определим соответственно матрицы $\tilde{X}(1), \tilde{X}(2)$.

$$\tilde{X}(1) = (\tilde{x}(i_1) : \dots : \tilde{x}(i_{n_1})), \quad \tilde{X}(2) = (\tilde{x}(j_1) : \dots : \tilde{x}(j_{n_2})),$$

$$\tilde{x}(i_k) = x(i_k) - x(1), \tilde{x}(j_s) = x(j_s) - x(2), k = \overline{1, n_1}, s = \overline{1, n_2}$$

$$\mathfrak{X}(1) = \frac{1}{n_1} \sum_{k=1}^{n_1} x(i_k), \mathfrak{X}(2) = \frac{1}{n_2} \sum_{s=1}^{n_2} x(j_s)$$

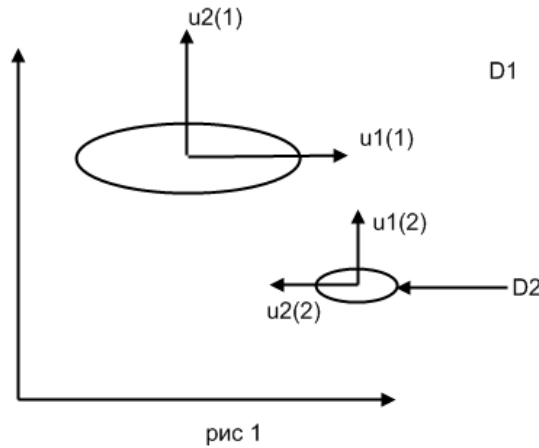
Будем предполагать, что n_1 и n_2 имеют настолько достаточно большие значения, что векторы $\tilde{x}(i_k), k = \overline{1, n_1}$ и соответственно $\tilde{x}(j_s), s = \overline{1, n_2}$ обладают свойством

$$\text{rank } \tilde{X}(1) = m, \text{rank } \tilde{X}(2) = m. \quad (1)$$

Тогда согласно свойствам взвешенных проекционных матриц имеют место соотношения

$$\begin{aligned} x(i_k) &\in \{x : \frac{1}{d_1^2} (x - \mathfrak{X}(1))^T R(\tilde{X}^T(1))(x - \mathfrak{X}(1)) \leq 1\}, k = \overline{1, n_1} \\ x(j_s) &\in \{x : \frac{1}{d_2^2} (x - \mathfrak{X}(2))^T R(\tilde{X}^T(2))(x - \mathfrak{X}(2)) \leq 1\}, s = \overline{1, n_2} \end{aligned} \quad (2)$$

Геометрическая интерпретация для $m = 2$ этих соотношений приведена на рис 1.



$$\text{Здесь } d_1 = \max_{k=1, n_1} (x(i_k) - x(1))^T R(\tilde{X}^T(1))(x(i_k) - x(1)),$$

$$d_2 = \max_{s=1, n_2} (x(j_s) - x(2))^T R(\tilde{X}^T(2))(x(j_s) - x(2)),$$

$$R(\tilde{X}^T(1)) = \tilde{X}^{+T}(1)\tilde{X}^+(1), R(\tilde{X}^T(2)) = \tilde{X}^{+T}(2)\tilde{X}^+(2)$$

Тогда в качестве меры расстояния от некоторой точки $x \in R^m$ до множества $\Omega_j, j = 1, 2$ будем рассматривать вместо евклидова расстояния величину

$$\rho(x, \Omega_j) = \frac{1}{d_j} ((x - x(j))^T R(\tilde{X}^T(j))(x - x(j)))^{\frac{1}{2}} \quad (3)$$

Что позволяет реорганизовать принадлежность точек $x(j), j = \overline{1, n}$ множествам D_1 и D_2 т.е. если некоторая точка $x(j)$ удовлетворяет условию

$$\rho(x(j), \Omega_1) \leq \rho(x(j), \Omega_2)$$

То её далее будем относить в множество точек D_1 , и, если

$$\rho(x(j), \Omega_1) > \rho(x(j), \Omega_2)$$

То $x(j)$ отсылаем в множество D_2 . Таким образом алгоритм формирования кластеров на основании использования множеств-контейнеров имеет следующее представление

1. Сформируем кластеры D_1 и D_2 .
2. Пересчитываем центроиды согласно (3).
3. Если условия $\rho(x(j), \Omega_1) \leq \rho(x(j), \Omega_2)$, $\rho(x(j), \Omega_1) > \rho(x(j), \Omega_2)$ выполнены переходим на шаг 2 пока есть не пересчитанные точки, иначе на выход.

Матрицы $R(\tilde{X}^T(1))$, $R(\tilde{X}^T(2))$ приобретают рекуррентную форму трансформации при перемещении некоторой точки $x(j)$ из множества D_1 в D_2 или из D_2 в D_1 за счет изъятия или пополнения из матриц $\tilde{X}(1)$, $\tilde{X}(2)$ соответствующего вектора и изменения векторов $\mathbf{x}(1)$ и $\mathbf{x}(2)$.

Заключение

В данной работе описаны контейнерные средства кластеризации и классификации сигналов, которые могут быть использованы для решения задач кластеризации и классификации массива текстовых печатных документов, ультразвуковых сигналов, графических образов и т.д.

Литература

1. Кириченко Н.Ф. Куц Роман (США), Лепеха Н.П. Множества принадлежности в задачах классификации сигналов. // Проблемы управления и информатики. - 2001, №5. - С.71-85.
2. Кириченко М.Ф. Куц Роман (США), Лепеха М.П. Множини приналежності в задачах класифікації сигналів // П'ята Всеукраїнська міжнародна конференція УкрОБРАЗ'2000. Праці Інститут Кібернетики НАНУ. – Київ, 2000. - С.83-87.
3. Котов А., Красильников Н. Кластеризация данных. [Электрон. ресурс] // 2006.- <http://logic.pdmi.ras.ru/~yura/internet/02ia-seminar.ppt>