«МНОЖЕСТВА И РАССТОЯНИЯ СООТВЕТСТВИЯ» В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ: ГИПЕРПЛОСКОСТИ

Николай Кириченко, Владимир Донченко

***Аннотация**: Рассматриваются общие проблемы кластеризации. Предложена концепция «множеств» и «расстояний соответствия» в построении кластеров, рассмотрены модели кластеризации, в которых «множествами соответствия» являются гиперплоскости, а «расстояниями соответствия» – различные варианты расстояний в связи с соответствующими гиперплоскостями. Развит аппарат псевдообращения по Муру – Пенроузу: приведены рекуррентные формулы возмущения для ортогональных проекторов и R-операторов, связанных с псевдообращением. Рекуррентные формулы возмущения использованы для построения алгебраического варианта Jack Knife'a. Приведена сборка важных для приложений результатов, касающихся псевдообращения.*

***Ключевые слова:** кластеризация, кластеризация по гиперплоскостям, псевдообращение по Муру - Пенроузу, сингулярное представление (SVD), ортогональные проекторы, псевдообращение для возмущённых матриц, преобразование Хока.*

***ACM Classification Keywords**: G.3 Probability and statistics, G.1.6. Numerical analysis: Optimization; G.2.m. Discrete mathematics: miscellaneous.*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Вступление

Статья посвящена алгебраическим аспектам задачи кластеризации (см., например, [Kohonen, 2001]) как задачи группирования информации. В дальнейшем будет обсуждаться вопрос о разбиении имеющихся элементов на два класса с тем, что процедуру такого разбиения можно запускать рекурентно.

Важным, как представляется авторам, во всех методах кластеризации является представление о «множествах соответствия» и «расстояниях соответствия». Типичным представителем первых являются прототипы-представители (prototypes) классов в методе к-средних. Что касается «расстояний соответствия», то, это меры соответствия «множествам соответствия», в соответствии с которыми элемент относят к тому или иному классу: как правило, – по минимальному значению «расстояния». Как правило, такими расстояниями соответствия являются, евклидовы расстояния в соответствующих пространствах признаков.

Заметим также, что процедуры кластеризации построены на применении стандартной рекуррентной процедуры: последовательного объединения (merging), разбиения (splitting) или уточняющих друг друга разбиений.

Разделяют также процедуры кластеризации с учителем (обучение с учителем – supervised learning) и – без учителя (unsupervised learning). В первом случае имеющиеся элементы уже разделены на классы, во втором – следует выделить классы на основе анализа внутренней структуры совокупности $x(1),\ldots,x(n)$ векторов из пространства признаков R^m

В задачах кластеризации следует также выделять этап обучения: построения соответствующих классов-кластеров (этап обучения), и этап использования построенного разбиения: отнесения каждого нового вектора признаков к одному из построенных классов.

В предлагаемой вниманию читателю работе речь идёт об использовании гиперплоскостей в качестве «множеств соответствия», о «расстояниях соответствия», построенным в связи с гиперплоскостями, а

также об обеспечении рекуррентности применения процедуры обучения без учителя; о согласованности обучения с учителем и без него; об аппарате псевдообращения по Муру – Пенроузу ([Moore,1920], [Penrose,1955]); о важном продвижении и расширении возможностей аппарата псевдообращения: о теории возмущения псевдообратных матриц ([Кириченко, 1997]), а также – о её совершенствовании и применении в задачах кластеризации (см. также [Кириченко, Донченко, 2007 a),b)]). Заметим, что важные примеры применения теории псевдообращения к исследованию классических прикладных задач, отличных от задач кластеризации, можно найти в работах [Кириченко, Лепеха, 2002], [Кириченко, Донченко, 2005] .

Заметим также, что важными вехами в развитии аппарата псевдообращения, в частности, в обеспечении эффективности построения соответствующих рекуррентных процедур и вычисления расстояний соответствия в них, – являются: прямые [Албert, 1977] и обратные [Кириченко, 1997] формулы Гревиля; формулы псевдообращения для замены строки или столбца матрицы [Кириченко, Лепеха, 2002,], [Кириченко, Донченко, 2005]; также формулы возмущения для Z - и R-операторов [Кириченко, Донченко., 2007 b)]. Отметем также, что задача кластеризации по гиперплоскостям, порождённым пространствами значений подходящих афиннных операторов, как вариант применения преобразования Хока, рассматривалась в работе [Donchenko, 2003].

В первой части предлагаемой работы приводится подборка результатов, важных в технике применения аппарата псевдообращения.

Во второй части рассматривается собственно задача кластеризации: рассматриваются подходящие гиперплоскости в качестве «множеств соответствия», строятся подходящие «расстояния соответствия» в связи с введёнными в рассмотрение гиперплоскостями, рассматриваются проблема обеспечения рекуррентности в вычислении «расстояний соответствия» как внутри рекуррентного шага, так и между разными шагами.

Отметим, что аппарат псевдообращения позволяет выписывать явные формулы, как для «расстояний соответствия», так и явно описывать «множества соответствия» в терминах смещения и явного описания ортогональных проекторов соответствующих линейных подпространств (ср. с вычислительными процедурами [Vapnik,1998] для статистических вариантов кластеризации на основе ковариационных). Вычислительные алгоритмы для расстояния от гиперплоскостей использовались, к примеру, также в работе [Haykin,1999].

Постановка задачи

Собственно, использование гиперплоскостей как аппарата решения задач группирования информации в статистической постановке, восходит к методу главных компонент: [Pearson , 1901] (другие названия метод Хётеллинга (*Hotelling*), метод Карунена-Лоэва (*Karhunen-Loeve*)) и имеет в основе идею такого ортогонального преобразования имеющегося набора случайных величин, которое бы приводило матрицу ковариаций к главным осям. Ещё раз обратим внимание читателя на специфически статистический вариант постановки и применения метода кластеризации в виде метода главных компонент, связанный с анализом естественного матричного объекта, каковым является матрица ковариаций, и применению классического результата Сильвестра [Sylvester, 1889]. Псевдообращение позволяет анализировать матрицы произвольной размерности, а не обязательно квадратные; позволяет эффективно строить ортогональные проекторы, отвечающие «естественным подпространствам» линейного оператора: подпространству значений и ядру оператора; описывать гиперплоскости, отвечающие всем решениям системы линейных алгебраических уравнений (СЛАУ), а также описывать необходимые и достаточные условия существования таких решений: описывать «наилучшие» приближенные решения (псевдорешения) СЛАУ; явно описывать невязку соответствующего приближения.

В последующем будем рассматривать задачу кластеризации в обучении без учителя для дихотомического варианта постановки задачи: для разбиения имеющейся совокупности *x(1),…,x(n)* векторов:из пространства признаков R^m на две части. В качестве множеств соответствия для каждого из классов-кластеров будут рассматриваться две гиперплоскости $\Gamma(k) \subseteq R^m, k = 1,2$: $\Gamma(k) = x_k + L_k \subseteq R^m, k = 1,2$, *x* – будем называть смещение гиперплоскости, . *L* –подпространством гиперплоскости. Таким образом, решение задачи кластеризации в такой постановке включает в себя

- построение «множеств соответствия» в виде гиперплоскостей: описание их смещений и соответствующих подпространств;
- описание «расстояний соответствия»;
- разбиение векторов *x(1),…,x(n)* обучающей выборки на две части в соответствии с минимумом «расстояния соответствия» на две части:

  $x(i_1),\ldots,x(x(i_{n_1})) \in \Gamma(1)$, $x(j_1),\ldots,x(x(j_{n_2})) \in \Gamma(2)$:

  $\{i_1,\ldots,i_{n_1}\} \cup \{j_1,\ldots,j_{n_2}\} = \{1,2,\ldots,n\}$, $n_1 + n_2 = n$;

- построение решающего правила, в соответствии с которым следует относить объект, не представленный в выборке, к одному из двух классов.

Естественным для рекуррентной процедуры построения классов-кластеров является получение и использование результатов, обеспечивающих рекуррентность.

Заметим, что вариантом указанной задачи кластеризации является такой, в котором дополнительно фиксируется общая размерность s: s<m гиперплоскостей $\Gamma(k) = x_k + L_k, k = 1,2$.

Напомним, что «гиперплоскости соответствия» подлежат определению на основе внутренней структуры имеющегося набора векторов *x(1),…,x(n)*.

Вспомогательные определения и утверждения

Псевдообращение и сингулярное (SVD –) представление. Псевдообращение – обозначается A^+ – по Муру - Пенроузу ([[Moore,1920], Penrose,1955], см. также [Алберт, 1977]) для $m \times n$ матрицы *A* может определяться одним из нескольких эквивалентных способов, среди которых отметим определение через сингулярное представление матриц (SVD-разложение), когда псевдообращение определяется соотношением:

$$A^+ = \sum_{i=1}^{r} x_i y_i^T \lambda_i^{-1}, \quad (1)$$

которое определяется элементами SVD-представления исходной матрицы:

$$A = \sum_{i=1}^{r} y_i x_i^T \lambda_i, \quad (2)$$

в котором: $\lambda_1^2 \geq \ldots \lambda_r^2 > 0$ – общий набор ненулевых собственных чисел матриц $AA^T, A^T A$, $y_i, i = \overline{1,r}$ и $x_i, i = \overline{1,r}$, соответственно, – ортонормированные наборы собственных векторов этих матриц, а *r = rank A= rank* A^T.

Чаще всего сингулярное разложение матрицы A представляется в виде, определяемом следующей леммой.

Лемма 1. Для любой $m \times n$ матрицы *A* ранга *r* существуют $Y - m \times r$ и $X - r \times n$ с ортонормированными столбцами и строками соответственно, а также диагональная матрица $\Lambda = diag(\lambda_1,\ldots,\lambda_r), \lambda_1 \geq \ldots \geq \lambda_r > 0$ такие, что

$A = Y\Lambda X$. (3)

Представление (2) является эквивалентным вариантом представления (3), если через $x_i, i = \overline{1,r}$ обозначить столбцы (ортонормированные) матрицы X , а через $y_i^T, i = \overline{1,r}$ – строки (ортонормированные) матрицы Y . В таких обозначениях справедлива следующая лемма.

Лемма 2.Произведение $Y\Lambda X$ матриц в (3) может быть представлено через «столбцовое» для Y и «строчное» для X представление может быть представлено е в виде

$$X\Lambda Y = \sum_{i=1}^{r} y_i x_i^T \lambda_i .$$

Собственно, указанное несложное утверждение вытекает из того, что произведение BC двух матриц со «столбцовым» предоставлением для первой и «строчным» представлением для второй::

$$B = (b(1),\ b(2),\ \ldots,\ b(p)), \qquad C = (c(1),\ c(2),\ \ldots,\ c(p))^T \tag{4}$$

допускает представление в виде

$$BC = (b(1),\ b(2),\ \ldots,\ b(p))(c(1),\ c(2),\ \ldots,\ c(p))^T = \sum_{i=1}^{p} b(i)c_{(i)}^T \tag{5}$$

В дальнейшем представление (4) будет использоваться и для других матриц. При этом обозначение обозначения $b(i)$, $c_{(i)}^T$, $i = \overline{1,p}$, будут использоваться для обозначения соответственно строк и столбцов необходимых матриц.

Основные ортогональные проекторы: *P*-проекторы, *Z*-проекторы. Псевдообращение позволяет в явном виде выписать пару ортогональных проекторов (ОП), – обозначим их $P(A), P(A^T)$, и будем называть P - проекторами, – на подпространства $L(A^T), L(A)$ значений операторов A^T, A соответственно: $P(A) = A^+A, P(A^T) = A^{T+}A^T = AA^+$. Ортогональные проекторы, которые будем обозначать $Z(A), Z(A^T)$ и называть Z - проекторами – определим соотношениями: $Z(A) = E_n - P(A), Z(A^T) = E_m - P(A^T)$ соответственно. Очевидным образом, Z - проекторы являются ортогональными проекторами на подпространства $L_{A^T}^{\perp}, L_A^{\perp}$, ортогональные к подпространствам $L(A^T), L(A)$ соответственно. Заметим, что $L_{A^T}^{\perp} = KerA$, $L_A^{\perp} = KerA^T$. Соответственно, $Z(A), Z(A^T)$ являются ортогональными проекторами на подпространства нулей $KerA, KerA^T$ операторов A, A^T соответственно.

Замечание 1. Обратим внимание также на то, что каждое из подпространств $L(A), L(A^T)$ является линейной оболочкой соответственно векторов-столбцов и векторов-строк матрицы A.

R-операторы. Важными в связи с определением расстояний соответствия и рекуррентными формулами псевдообращения: формулами позволяющими записывать соответствующий оператор при добавлении или вычёркивании строки или столбца матрицы, – являются также операторы, которые будем называть R - операторы. Их будем определять соотношениями:

$R(A) = A^+A^{T+}, R(A^T) = A^{+T}A^+$.

Важную роль в реализации аппарата псевдообращения в прикладных задачах играют прямые (см., например, [Алберт, 1977]) и обратные [Кириченко 1997] формулы Гревиля(Greville), а также формулы возмущения псевдообращения [Кириченко, 1997]. И в том и в другом случае речь идёт о формулах, связывающих псевдообращение преобразованной матрицы с псевдообращением исходной. В первом случае (прямых или обратных формулах Гревиля) речь идёт о преобразовании матрицы введением или вычёркиванием дополнительного строки или столбца. Во втором – о преобразовании исходной матрицы

аддитивной добавкой ab^T . Таким образом, в формулах псевдообращения для возмущённых матриц речь идёт о выражении псевдообращения возмущенной матрицы $(A+ab^T)^+$ через A, A^+, a, b .

Прямые и обратные формулы Гревиля приведены ниже. Формулы возмущения псевдообращения можно найти в уже цитированной работе [Кириченко 1997]. Ниже приведены полученные на их основе формулы возмущения для *Z*- и *R*-операторов.

Заметим, что комбинация прямых и обратных формул Гревиля позволяет получить формулы псевдообращения при замене строки или столбца исходной матрицы. Соответствующие представления можно найти в работах [Кириченко, Лепеха, 2002], [Кириченко, Донченко, 2005]. Там же приведены формулы, определяющие вид, *Z* - и *R* - операторов при замене строки или столбца матрицы, для которой они рассматриваются.

Прямые формулы Гревиля (Greville).

Напомним, что прямые формулы Гревиля – это формулы, определяющие вид псевдообращении матрицы при её дополнении строкой или столбцом. Они определяются соотношениями, в которых используется блочное представление псевдообращения расширенной матрицы: через $P-m\times n$ - матрицу и, q - $n\times 1$ - вектор для расширения матрицы строкой

$$\begin{pmatrix} A \\ a^T \end{pmatrix}^+ = (P \vdots q) \tag{6}$$

и через $Q-n\times m$ - матрицу и, q - $m\times 1$ - вектор

$$(A \vdots a)^+ = \begin{pmatrix} Q \\ q^T \end{pmatrix} \tag{7}$$

при дополнении матрицы столбцом.

Замечание 2. Обратим внимание читателя, что вектор a в (6) и (7) имеет разные размерности: размерность $n\times 1$ в первом и – $m\times 1$ во втором.

Теорема 1.(прямые формулы Greville– дополнение строкой). В представлении (6)

$$\begin{cases} P=(E-qa^T)A^+ \\ q=\begin{cases} \dfrac{Z(A)a}{a^TZ(A)a}, a^TZ(A)a>0(\text{нез.}) \\ \dfrac{R(A)a}{1+a^TR(A)a}, a^TZ(A)a=0(\text{зав.}) \end{cases} \end{cases}, \tag{8}$$

Теорема 2(прямые формулы Greville– дополнение столбцом) В представлении (7)

$$\begin{cases} Q==A^+(E-aq^T) \\ q=\begin{cases} \dfrac{Z(A^T)a}{a^TZ(A^T)a}, a^TZ(A^T)a>0(\text{нез.}) \\ \dfrac{R(A^T)a}{1+a^TR(A^T)a}, a^TZ(A^T)a=0(\text{зав.}) \end{cases} \end{cases}, \tag{9}$$

Замечание 3. Вид вектора q в прямых формулах Гревиля определяется линейной зависимостью вводимого вектора a^T или a от, соответственно, строк или столбцов матрицы A. Линейная независимость обеспечивается нулевым значением квадратичной формы (с соответствующей матрицей - *Z*–оператором) на векторе a .

Обратные формулы Гревиля. Как и в прямых формулах Гревиля, вид выражений, связывающих псевдообращения исходной и преобразованной матрицы, выписывается в рамках блочного представлением (6) или (7), и так же – определяется линейной зависимостью или независимостью вычёркиваемой строки или столбца: сохранением или падением ранга преобразованной матрицы.

Изменяется только вид соответствующего условия. Теперь условием независимости является условие $a^T q = 1$.

Теорема *3.* (обратные формулы Гревиля – вычёркивание строки) В обозначениях (6) имеет место соотношение

$$A^+ = \begin{cases} \left(I_n - \dfrac{qq^T}{\| q \|^2} \right) P, a^T q = 1, (\text{нез.}), & \text{ранг падает} \\ \left(I_n - \dfrac{qa^T}{1 - a^T q} \right) P, a^T q < 1 (\text{зав.}), & \text{ранг сохраняется} \end{cases} . \qquad (10)$$

Теорема *4.* (обратные формулы Гревиля – вычёркивание столбца) В обозначениях (7) имеет место соотношение

$$A^+ = \begin{cases} Q\left(I_m - \dfrac{qq^T}{\| q \|^2} \right), a^T q = 1, (\text{нез.}), & \text{ранг падает} \\ Q\left(I_m - \dfrac{aq^T}{1 - a^T q} \right), a^T q < 1 (\text{зав.}), & \text{ранг сохраняется} \end{cases} .$$

Теорема 5 (формулы возмущения для *Z*- и *R*- операторов). При возмущении матрицы A матрицей $a \times b^T$ *Z*- и *R*- операторы для возмущённой матрицы определяется следующими соотношениями, вид которых определяется линейной зависимостью или независимостью векторов - составляющих возмущения от соответствующих составляющих матрицы A, а также от того, сохраняется или падает ранг возмущённой матрицы:

1) Для векторов a и b^T линейно не зависимых от, соответственно, столбцов и строк матрицы A, т.е. при выполнении условий $a^T Z(A^T)a > 0, b^T Z(A)b > 0,$ справедливы следующие соотношения

$$Z(A + ab^T) = Z(A) + \frac{Z(A)bb^T Z(A)}{b^T Z(A)b};$$

$$Z((A + ab^T)^T) = Z(A^T + ba^T) = Z(A^T) + \frac{Z(A^T)aa^T Z(A^T)}{a^T Z(A^T)a};$$

$$R(A + ab^T) = R(A) - R(A)\frac{bb^T Z(A)}{b^T Z(A)b} \quad \frac{Z(A)bb^T}{b^T Z(A)b} R(A) \quad cA^+ ab^T Z(A) \quad cZ(A)ba^T A^{+^T} +$$

$$+ \frac{A^+ aa^T A^{+^T}}{a^T Z(A^T)a} + \frac{b^T R(A)ba^T Z(A^T)a + \left(1 + b^T A^+ a\right)^2}{a^T Z(A^T)a\left[b^T Z(A)b\right]^2} Z(A)bb^T Z(A),$$

где $c = \dfrac{1 + b^T A^+ a}{a^T Z(A^T)ab^T Z(A)b}$.

2) Для вектора a линейно зависимого от столбцов матрицы *A*, а вектора b^T – линейно не зависимого от строк матрицы таким образом, что, – для упрощения представления результата, – $b \perp L_{A^T}$, т.е. при выполнении условий $a^T Z(A^T)a = 0, b^T Z(A)b = \| b \|^2$, справедливы соотношения:

$$Z(A + ab^T) = Z(A) + \frac{k_{A,a,b} k_{A,a,b}^T}{\| k_{A,a,b} \|^2} \quad \frac{bb^T}{\| b \|^2},$$

где: $k_{A,a,b} = A^{+}a \quad \frac{b}{||b||^2}$,

$Z((A+ab^T)^T) = Z(A^T + ba^T)$,

$R(A+ab^T) = \left(I_n - \frac{kk^T}{||k||^2}\right)R(A)\left(I_n - \frac{kk^T}{||k||^2}\right)$.

3) Для векторов a и b^T одновременно линейно зависимых от соответственно столбцов и строк матрицы A, при условии падения ранга возмущённой матрицы: $rank(A+ab^T) = rankA-1$, т.е. при выполнении условий: $a^T Z(A^T)a = 0, b^T Z(A)b = 0, b^T A^{+}a = -1$, справедливы следующие соотношения:

$$Z(A+ab^T) = Z(A) + \frac{A^{+}aa^T(A^{+})^T}{a^T R(A^T)a},$$

$$Z((A+ab^T)^T) = Z(A^T + ba^T) = Z(A) + \frac{(A^{+})^T bb^T A^{+}}{b^T R(A)b},$$

$$R(A+ab^T) = A^{+}(a,b)A^{+T}(a,b),$$

где: $A^{+}(a,b) = A^{+} \quad \frac{A^{+}aa^T R(A^T)}{a^T R(A^T)a} \quad \frac{R(A)bb^T A^{+}}{b^T R(A)b} + cA^{+}ab^T A^{+}$, $c = \frac{b^T R(A)A^{+}a}{a^T R(A^T)a\, b^T R(A)b}$.

4) Для векторов a и b^T одновременно линейно зависимых от, соответственно, столбцов и строк матрицы A, но при условии неизменности ранга возмущённой матрицы по сравнению с рангом A, т.е. при выполнении условий

$a^T Z(A^T)a = 0, b^T Z(A)b = 0, b^T A^{+}a \neq -1$,

справедливы следующие соотношения:

$$Z(A+ab^T) = Z(A), \qquad Z((A+ab^T)^T) = Z(A^T + ba^T) = Z(A^T),$$

$$R(A+ab^T) = R(A) \quad \frac{A^{+}ab^T R(A)}{1+b^T A^{+}a} \quad \frac{R(A)ba^T A^{+T}}{1+b^T A^{+}a} + \frac{b^T R(A)b}{1+b^T A^{+}a} A^{+}aa^T A^{+T}.$$

Основные элементы кластеризации по гиперплоскостям – множества соответствия

Как уже упоминалось, построение «множеств соответствия» в виде гиперплоскостей предполагает конструктивное описание их смещений и соответствующих линейных подпространств.

Смещение гиперплоскостей.

Смещения предлагается определять как средние векторов, принадлежащих к каждой из частей разбиения. Можно также выбрать в качестве смещения один из элементов разбиения.

Подпространства гиперплоскостей. При наличии смещений подпространства гиперплоскостей определяются как подпространства, натянутые на центрированные смещением (преобразованные вычитанием определённого вектора) векторы каждой из частей разбиения. В дальнейшем будет предполагаться, что центрирование каждой части разбиения производится соответствующими средними $\bar{x}_k, k = 1,2$.

Конструктивное описание подпространств, натянутых на каждую из центрированных совокупностей векторов, обеспечивается построением для каждой из гиперплоскости подходящей матрицы $A_k, k = 1,2$ так, чтобы подпространство - множество значений $L_k, k = 1,2$ каждой из них совпадало с подпространством соответствующей гиперплоскости, т.е. с линейной оболочкой каждой из центрированных групп векторов. В соответствии с замечанием 1 в качестве таких матриц можно выбрать матрицы, столбцами которых являются центрированные векторы каждой из частей разбиения

соответственно. В этом случае ортогональными проекторами $P_{L_k}, k = 1,2$ для каждого из подпространств гиперплоскостей будут P - проекторы для транспонированных к соответствующим матрицам:

$P_{L_k} = P(A_k^T), k = 1,2$.

Таким образом, гиперплоскости $\Gamma_k, k = 1,2$ определяются парами $(\bar{x}_k, A_k), k = 1,2$:

$$\Gamma_k = \Gamma(\bar{x}_k, A_k), k = 1,2 \quad (11)$$

Основные элементы кластеризации по гиперплоскостям – расстояния соответствия

В качестве «расстояний соответствия» векторов до каждого «множеств соответствия» $\Gamma_k = \Gamma(\bar{x}_k, A_k), k = 1,2$ предлагается рассматривать евклидово расстояние векторов до гиперплоскостей $\Gamma_k = \Gamma(\bar{x}_k, A_k), k = 1,2$, каковыми эти «множества соответствия» являются. Средства псевдообращения позволяют конструктивно описать соответствующие расстояния. Такое конструктивное описание возможно и в том случае, когда задаётся размерность s: $s \le min(rank A_k, k = 1,2)$ – подпространств гиперплоскостей. Формулы, определяющие соответствующие расстояния, являются предметом следующей леммы.

Лемма 1. Для $\Gamma_k = \Gamma(\bar{x}_k, A_k), k = 1,2$ расстояния соответствия $\rho(x, \Gamma_k), k = 1,2$ произвольного вектора $x \in R^m$ до каждой из двух гиперплоскостей $\Gamma_k, k = 1,2$ определяются соотношением:

$$\rho(x, \Gamma_k) = (x - \bar{x}_k))^T Z(U_s^T(k))(x - \bar{x}_k), k = 1,2, \quad (12)$$

Где $U_s(k) = \begin{cases} A_k = \sum_{i=1}^{r} y_i(k) x_i^T(k) \lambda_i(\mathrm{k}) & \text{ðàçì .s í å çàäàíà} \\ \sum_{i=1}^{s} y_i(k) x_i^T(k) \lambda_i(k) & \text{ðàçì .s çàäàíà} \end{cases}, r = 1.2$.

Кластеризации по гиперплоскостям – основные шаги алгоритма

Алгоритм кластеризации по гиперплоскостям состоит в последовательном, рекуррентном уточнении «множеств соответствия», каковыми являются гиперплоскости. На каждом рекуррентном шаге происходит уточнение набора элементов, порождающих «множества соответствия», построение пар $(\bar{x}_k, A_k), k = 1,2$ отвечающих уточнённому разбиению, после чего происходит новое «уточнение разбиения» отбором в каждую часть разбиения векторов исходного набора по минимуму расстояний до вновь построенных гиперплоскостей. В общем, алгоритм состоит в выполнении следующих шагов.

1. На первом шаге производится разбиение на две совокупности произвольным образом.
2. На втором шаге для каждой из частей разбиения вычисляются:
   - смещения $\bar{x}_k, k = 1,2$, как средние по векторам каждой из частей разбиения;
   - матрицы $A_k, k = 1,2$, как матрицы, построенные из центрированных соответствующими средними векторов каждой из групп как из столбцов.

3. На третьем шаге происходит «уточнение» разбиения: вычисляются «расстояния соответствия» каждого из векторов *x(1),…,x(n)* до каждого из двух построенных «множеств соответствия»: до каждой из двух гиперплоскостей, – и происходит отнесение каждого из векторов *x(1),…,x(n)* к той части разбиения, к которой он оказался ближе по «расстоянию соответствия» (12). В результате происходит формирование нового, «уточнённого» разбиения векторов *x(1),…,x(n)* на две части.

4. На четвёртом шагу происходит возвращение ко второму шагу алгоритма.

Кластеризация по гиперплоскостям – модификация расстояний

Расстояния до гиперплоскостей в лемме 1 определяются значениями квадратичных форм с матрицами $\sum_{i=1}^{s} y_i(k)x_i^T(k), k=1,2$,

Их можно рассматривать как взвешенное среднее матриц $y_i(k)x_i^T(k), i=\overline{1,r}, k=1,2$ с весами $\omega_i = \begin{cases} 1, i \le s \\ 0, i = \overline{s+1,r} \end{cases}$, соответственно, в нормированном варианте $\omega_i = \begin{cases} 1/s, i \le s \\ 0, i = \overline{s+1,r} \end{cases}$.

Рассмотрение взвешенных варианта сумм из $y_i(k)x_i^T(k), i=\overline{1,r}, k=1,2$ с нормированными весами $\lambda_i^2(k)$, $i=\overline{1,r}$, $k=1,2$ даёт следующий вариант расстояний ρ_R до «множеств соответствия». Они для этого случая определяются соотношением:

$$\rho_R(x,\Gamma_k) = \frac{1}{trR(A_k^T(k)A_k)}(x - \bar{x}_k)^T R(A_k^T)(x - \bar{x}_k), k=1,2. \quad (13)$$

Использование в качестве «расстояний соответствия» расстояний, определяемых соотношением (13) приводит к очевидному изменению алгоритма кластеризации: в нём «уточнение» разбиения третьего шага происходит на основе «расстояний соответствия», определяемых соотношениями (13) вместо – (12).

Рекуррентные формулы для алгебраического Jack Knife'а

При проверке элементов совокупностей на соответствие вычислением расстояний по формулам (12) или (13) тестируемые элементы принимают участие в формировании гиперплоскостей, представляющих кластеры. Резонной является также построение такой процедура проверки соответствия, при которой тестируемый элемент кластера, исключается из числа объектов, которые его определяют. В статистике такая процедура исключения носит название "Jack Knife"(складной нож) [Эфрон, 1988]. Поэтому процедуру тестирования на принадлежность кластеру с исключением тестируемых элементов из описания кластера будем называть алгебраическим Jack Knife'ом.

Заметим, что естественным является вариант кластеризации, когда исключение элемента приводит к падению ранга матрицы $A(k), k=1,2$ ((п.3 теоремы 5)). Псевдообращение даёт конструктивную явную формулу проверки соответствующего условия.

Исключение тестируемых элементов из кластера изменяет как сдвиг (центр кластера), так и линейное подпространство кластера. Формулы (12),(13) при таком исключении, очевидным образом, переписываются в виде, для изменённых смещений (будем считать их средними) и изменённых матриц: $x_k^{(0)}, A^{(0)}(k), k=1,2$ соответственно.

Лемма 1 даёт возможность эффективной организации процедуры «отсеивания», в которой критерий замены строится на основе леммы 1 и имеет вид, определяемый следующей теоремой.

Теорема 6. В условиях падения ранга (п.3 теоремы 5) расстояния до гиперплоскостей, после исключения элемента из числа порождающих элементов, определяется следующим соотношением для одного из значений $k=1$ или $k=2$

$$\rho(x_j(k),\Gamma_j^{(0)}(k)) = \frac{n_k^2}{\left|\left|\left(E_m \quad \frac{q_j(k)q_j^T(k)}{||q_j(k)||^2}\right)\sum_{l \ne j} q_l(k)\right|\right|^2}, j=\overline{1,n_k}, k=1,2,$$

Где $x_j(k)$, $\Gamma_j^{(0)}(k)$ $j=\overline{1,n_k}, k=1,2$ – исключаемые элементы каждой из совокупностей и гиперплоскости, отвечающие «усечённым» совокупностям, а $q_j(k)$, $j=\overline{1,n_k}, k=1,2$ столбцы с номером j, $j=\overline{1,n_k}$ в каждой из матриц $A_k^+, k=1,2$.

Заключение

В работе рассмотрены задачи кластеризации на основе концепции «множеств» и «расстояний соответствия», предложены варианты алгоритмов кластеризации, когда «множествами соответствия» являются гиперплоскости, а «расстояния соответствия» построены на основе вариантов расстояний до них. Применение аппарата псевдообращения позволяет описать все элементы соответствующих построений явными формулами, включая варианты алгебраического Jack Knife'a

Литература

[Алберт, 1977] Алберт А. Регрессия, псевдоинверсия, рекуррентное оценивание. М.: Наука, 1977.–305 с.

[Donchenko, 2003] Donchenko V.S. Hough Transform and Uncertainty//Proceedings International Conference "Knowledge Dialog – Solution". – V. – June 16-23, 2003.–Varna (Bulgaria). – P.391-395.

[Haykin,1999] Neural networks. A comprehensive Foundation. – New Jersey. – 1999.– 842 p.

[Кириченко, 1997] Кириченко Н.Ф. Аналитическое представление псевдообратных матриц//Киб. и СА. - №2, 1997– С.98-122.

[Кириченко, Донченко, 2007, a)] Кириченко Н.Ф., Донченко. В.С. Псевдообращение в задачах кластеризации.// Киб. и СА. - №4, 2007.– С.98-122.

[Кириченко, Донченко, 2007, b)] Кириченко Н., Донченко В. Алгебраический Jack Knife: кластеризация по гиперплоскостям// Proceedings: XIII-th International Conference "Knowledge –Dialog – Solution".–June 18-24, 2007, Varna (Bulgaria). – 2007. – V.1.– P.89-95.

[Кириченко, Лепеха, 2002] Кириченко Н.Ф., Лепеха Н.П. Псевдообратные и проекционные матрицы в применении к исследованию задач управления, наблюдения и идентификации// Киб. и СА.- №4, 2002. – С.107-123.

[Kohonen, 2001] Kohonen T., Self-Organizing Maps.– Third Extended Edition.– New York, 2001.– 501 p.

[Кириченко, Донченко, 2005] Кириченко М.Ф, Донченко В.С. Задача термінального спостереження динамічних системи: множинність розв'язків та оптимізація //Ж. Обч. та пр. мат. – Вип..3 , 2005.– С. 63-78.

[Moore,1920] Moore E.H. On the reciprocal of the general algebraic matrix//Bull. Amer. Math. Soc. – 26, 1920. – P. 394-395.

[Pearson, 1901] Pearson K., On lines and planes of closest fit to systems of points in space//Philosophical Magazine.–1901, N 2.– P. 559—572.

[Penrose,1955] Penrose R. A generalized inverse for matrices// Proc. Cambr. Philosophical Soc.- 51, 1955.– P. 406-413.

[Sylvester, 1889] Sylvester J.J. On the reduction of a bilinear quantic of the nth order to the form of a sum of n products by a double orthogonal substitution, Messenger of Mathematics, – 1889.– N19.– P., 42—46;

[Vapnik, 1998] Vapnik V.N. Statistical Learning Theory.–New York: Wiley. – 1998.

[Эфрон, 1988] Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Фин. и стат. – 1988.– 263 с.

Информация об авторах

Кириченко Николай Ф.– *Профессор, Институт кибернетики им. В.М.Глушкова НАН Украины, ведущий научный сотрудник.*

Донченко Владимир С. – *Профессор, Киевский национальный университет имени Тараса Шевченко, факультет кибернетики, Украина, e-mail: voldon@unicyb.kiev.ua*