АЛГЕБРАИЧЕСКИЙ JACK KNIFE: КЛАСТЕРИЗАЦИЯ ПО ГИПЕРПЛОСКОСТЯМ

Николай Кириченко, Владимир Донченко

Abstract: The Clustering Problem, when the clustering is carried out along appropriate hyper planes, was investigated. Different variants of the "correspondence distances" have been proposed and investigated including the algebraic Jack Knife one. Efficacy, constructivism and explicit representation for the objects under investigation are provided with the Generalized Inverse Technique including the Generalized Inverse Disturbances. The approach represented is the one to be the variant of the Hough Transform in its mathematical. The Collection of important for the applications results regarding Generalized Inverse is also represented in the paper.

Keywords: clusterization, Hough Transform, Moore – Penrose Generalized Inverse, Single Valued Decomposition (SVD), orthogonal projectors, Generalized Inverse for disturbed matrixes, clustering, clustering along hyper planes.

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

Вступление

Статья посвящена применению методов псевдо обращения (см. например, [Алберт, 1977]) и, в особенности их развития в виде теории возмущения псевдо обращения [Кириченко, Лепеха, 1997, 2002], в решении задач кластеризации. Задача кластеризации в такой постановке может рассматриваться как вариант преобразования Хока[Donchenko, 2003] При таком подходе определяющим является рассмотрение расстояний не до ограниченных множеств, которые отвечают элементам, отнесённым к тому или иному классу, а до подходящих гиперплоскостей, которые ими порождаются. Отметим, что в статистических подходах (см., например [Vapnik,1998]) гиперплоскости связываются с корреляционными матрицами подходящих распределений вероятностей в пространстве признаков. Вычислительные алгоритмы для расстояния от гиперплоскостей использовались, например, в работе [Haykin,1999]. В предлагаемой работе применение аппарата псевдо обращения позволяет выписать явные формулы «расстояний соответствия» в том числе и в варианте алгебраического Jack Knife'а[Эфрон, 1988].

Важными в прменеии рехники псевдообращения являются эквивалентные варианты определения псевдо обращения, прямые [Алберт, 1977] и обратные [Кириченко, Лепеха, 1997, 2002], формулы Гревиля, формулы псевдообращения для замены строки или столбца матрицы [Кириченко, Лепеха, 1997, 2002], [Кириченко, Донченко, 2005], а также формулы возмущения для Z - и R-операторов[Кириченко, Донченко., 2007]. В первой части предлагаемой работы приводится подборка некоторых таких результатов.

Вторая часть посвящена постановке и решению задачи разбиения наличной совокупности векторов на два кластера. Предложены различные варианты определения и вычисления «расстояний соответствия», которые имеют естественную геометрическую интерпретацию и эффективно описываются в терминах псевдо обращения. Явный вид формул, определяющих расстояния соответствия позволяет построить эффективные алгоритмы кластеризации.

Постановка задачи

В задаче о кластеризации по гиперплоскостям каждый из векторов признаков $x(1),\ldots,x(n)$ из пространства признаков R^m может принадлежать одной из двух – для простоты рассмотрений – гиперплоскостей $\Gamma(k) = x_k + L_k \subseteq R^m, k = 1,2$ (x – смещение, L подпространство гиперплоскости). Требуется выделить указанные плоскости и отнести каждый из векторов к той, к которой он принадлежит. Вариантом указанной задачи является такой, в которой дополнительно фиксируется размерность s: s<m – для каждой из гиперплоскостей $\Gamma(k) = x_k + L_k, k = 1,2$.

Эти гиперплоскости подлежат определению на основе внутренней структуры имеющегося набора векторов $x(1),\ldots,x(n)$ с соответствующим их разбиением на две части:

$$x(i_1),\ldots,x(x(i_{n_1}))\in \Gamma(1),$$

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

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

Вариантом такой задачи является и задача Duda&Hart'a 1972 года. Задача группировки по гиперплоскостям является одним из вариантов оценивания нескольких зависимостей заданного параметрического семейства, представленных в выборке [Donchenko, 2003]. В такой постановке речь идёт семействе аффинных преобразований между евклидовыми пространствами произвольных размерностей, когда определяющим является структура возможного пространства значений, а не представление оператора в том и ли ином базисе. Стандартным образом, предлагаемый к рассмотрению алгоритм кластеризации носит характер рекуррентной процедуры в ходе, которой после первично произвольного разбиения на две части происходит «рафинирование» исходных частей, в ходе которого осуществляется освобождение от «не отвечающих совокупности» элементов. Аппарат псевдообращения по Муру-Пенроузу позволяет получить явные формулы для расстояний от подходящих гиперплоскостей, которые отвечают тому или иному разбиению. Отметим, что указанные расстояния эффективно и явным видом описываются как в варианте, когда проверяемый элемент не исключается из совокупности, формирующей гиперплоскость, так и при его исключении. В последнем варианте процедура кластеризации аналогична статистической процедуре, которая называется Jack Knife(складной нож) [Эфрон, 1988] . Говоря об эффективности и адекватности применения аппарата псевдообращения для рассматриваемой задачи, отметим также, что формулы вычисления расстояний носят рекуррентный по исключаемым элементам вид. Достижение указанной рекуррентности обеспечивается теорией возмущения псевдообратных матриц [Кириченко, Лепеха, 1997].

Вспомогательные определения и утверждения

Пседообращение A^+ по Муру-Пенроузу для $m \times n$ матрицы A может определяться одним из нескольких еквивалентых способов, среди которых отметим определение через сингулярное представление матриц (SVD-разложение), когда псевдообращение определяется соотношением:

$A^+ = \sum_{i=1}^{r} x_i y_i^T \lambda_i^{-1}$, которое определяется элементами SVD-представления исходной матрицы: $A = \sum_{i=1}^{r} y_i x_i^T \lambda_i$, в котором: $\lambda_1^2 \geq ... \lambda_r^2 > 0$ – общий набор ненулевых собственных чисел; $y_i, i = \overline{1,r}$ и $x_i, i = \overline{1,r}$ – ортонормированные наборы собственных векторов матриц $P(A), P(A^T)$, $A^T A$: соответственно, а $r = rank\ A = rank\ A^T$.

Псевдообращение позволяет в явном виде выписать две пары ортогональных проекторов (ОП) $P(A), P(A^T)$: $P(A) = A^+ A, P(A^T) = A^{T+} A^T = AA^+$ и $Z(A) = E_n - P(A), Z(A^T) = E_m - P(A^T)$ (P- и Z-проекторы соответственно). Первая пара представляет собой ОП на подпространства L_{A^T}, L_A являющиеся множествами значений A^T, A соответственно, вторая – на ортогональные к ним подпространства соответственно $L_{A^T}^{\perp}$, $L_A^{\perp}$. Очевидны образом, $L_{A^T}^{\perp} = KerA$. Заметим также, что каждое из подпространств L_{A^T}, L_A является линейной оболочкой соответственно векторов-столбцов и векторов-строк матрицы A.

Важными в связи с рекуррентными формулами псевдообращения: формулами позволяющими записывать соответствующий оператор при добавлении или вычёркивании строки или столбца матрицы, – являются также - операторы, определяемые соотношениями: $R(A) = A^+ A^{T+}, R(A^T) = A^{+T} A^+$. К таким формулам псевдообращения относятся прямые [Алберт, 1977] и обратные формулы Гревиля(Greville) [Кириченко, Лепеха, 1997]. Такой же характер имеют формулы возмущения псевдообратных матриц [Кириченко, Лепеха, 1997], дающие представление $A^+(a,b) = (A + ab)^T$ через A, A^+, a, b, [там же], а также [Кириченко, Донченко, 2005]. В последней работе также приведены рекуррентные формулы для P-, Z - и R-операторов, описывающие их вид при замене строки или столбца матрицы, для которой они рассматриваются. Формулы, описывающие вид соответствующих операторов при возмущении матрицы, представлены следующей леммой[Кириченко, Донченко., 2007].

Лемма 1.

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:

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

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

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

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

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

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:

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

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

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

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

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

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

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

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

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

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

справедливы следующие соотношения:

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

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

Лемма 2.

Пусть подпространства гиперплоскостей являются множествами значений матриц $A(k), k = 1,2$ соответственно. Тогда расстояния соответствия $\rho(x, \Gamma(k)), k \quad 1,2$ произвольного вектора $x \in R^m$ до каждой из двух гиперплоскостей $\Gamma(k), k = 1,2$ определяются соотношением:

$$\rho(x, \Gamma_k) = (x \quad x_k))^T Z(U_s^T(k))(x \quad x_k), k = 1,2, \qquad (1)$$

Где $U_s(k) = \begin{cases} A(k) = \sum_{i=1}^{r} y_i(k) x_i^T(k) \lambda_i(k) & \text{разм.s не задана} \\ \sum_{i=1}^{s} y_i(k) x_i^T(k) \lambda_i(k) & \text{разм.s задана} \end{cases}, r = 1.2.$

Кластеризация по гиперплоскостям

Сам алгоритм кластеризации по гиперпл-оскостям состоит в выполнении следующих шагов.

1. Первичное разбиение на две совокупности произвольным образом.
2. Построение смещений $x_k, k = 1,2$. Соответствующие смещения могут быть реализованы или как средние по векторам каждой из отобранных совокупностей или в виде произвольных представителей каждой из них.
3. Построение матриц каждой из совокупностей $A(k), k = 1,2$ из векторов столбцов каждой из групп, центрированных соответствующими векторами смещений, построенных на шаге 2.
4. Определение гиперплоскостей $\Gamma(k), k = 1,2$ как таких, которые задаются смещениями, вычисленными на шаге 2, и подпространствами $L_k = L_{A(k)}, k = 1,2$
5. Вычисление расстояний элементов каждой из совокупностей до каждой из двух построенных гиперплоскостей с использованием результатов леммы 2.
6. Перераспределение векторов между совокупностями и повторением рекуррентных шагов.

Перераспределение элементов можно реализовывать разными способами с введением подходящих параметров алгоритма

Кластеризация по гиперплоскостям – модификация расстояний

Можно показать, что расстояния до гиперплоскостей леммы 2, являются вариантом квадратичной формы с весовыми коэффициентами, $\lambda_i^2, i = 1, r$ Ниже приводится вариант расстояний до гиперплоскостей, реализующий идею взвешивания следующим образом. Модифицированные расстояния ρ_R в этом случае определяются соотношением:

$$\rho_R(x, \Gamma(k)) = \frac{1}{trR(A^T(k)A(k))}(x \quad x_k)^T R(A^T(k))(x \quad x_k), k = 1,2. \qquad (2)$$

Кластеризация по гиперплоскостям алгебраический Jack Knife

При проверке элементов совокупностей на соответствие вычислением расстояний по формулам (1) или (2) тестируемые элементы принимают участие в формировании гиперплоскостей, представляющих кластеры. Поэтому целесообразной является такая процедура проверки соответствия, при которой

тестируемый элемент кластера, исключается из числа объектов, которые определяют исследуемый кластер. 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 $A(k), k = 1,2$. Псевдообращение даёт конструктивную явную формулу проверки соответствующего условия.

Исключение тестируемых элементов из кластера изменяет как сдвиг (центр кластера), так и линейное подпространство кластера. Формулы (1),(2) 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.

Теория возмущения псевдообратных матриц [Кириченко, Лепеха, 1997]. Даёт возможность эффективной организации процедуры «отсеивания», в которой критерий замены строится на основе леммы 1 и имеет вид, определяемый следующей теоремой.

Теорема. В условиях падения ранга

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

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» совокупностей, 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 $A^+(k), k = 1,2$.

Литература

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

[Haykin,1999] Neural networks. A comprehensive Foundation. – New Jersey: Prentice Hall 07458. – 1999.– 842 p.

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

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

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

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

[Кириченко, Донченко, 2005] Кириченко Н.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.

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

[Кириченко, Лепеха, 1997]. Кириченко Н.Ф., Лепеха Н.П. Псевдо обратные и проекционные матрицы в применении к исследованию задач управления, наблюдения и идентификации.//Киб. и СА.- №2, 1997– С.98-122.

[Кириченко, Лепеха., 2002] Кириченко Н.Ф., Лепеха Н.П. Псевдо обратные и проекционные матрицы в применении к исследованию задач управления, наблюдения и идентификации.// Киб. и СА.- №2, 1997– С.98-122.

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

Информация об авторах

Николоай Ф. Кириченко – Институт кибернетики им.В.М.Глушкова НАН Украины, ведущий научный сотрудник, профессор.

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