ОБОБЩЁННЫЕ ВАРИАНТЫ ПРЕОБРАЗОВАНИЯ ХОКА: СТАТИСТИЧЕСКИЙ И АЛГЕБРАИЧЕСКИЙ АСПЕКТЫ

Владимир Донченко

Abstract: The problems within mathematical theory of the Hough Transform are represented in the report including those to be the Clustering Problem in the Euclidean spaces along appropriate hyper planes and to be the arguments in the support of the its using in the "multi parametric estimation" variant. As regarding clustering problem explicit expressions for the grouping distances are represented. The Moore-Penrose Generalized Inverse is the base of the corresponding considerations.

Keywords: Hough Transform, clusterization, 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

Преобразование Хока (ПГ (и) , [Hough,1962]), появившееся и интенсивно развивающееся как инженерный метод обработки изображений (см., например, [Xu , Oja, 1992]), утверждается [Donchenko, 2003] как равноправный метод описания неопределенности, наряду с детерминированным, статистическим, минимаксным (с гарантированной точностью), нечётким.

Что касается последнего из упомянутых методов: нечёткого, – то его место и возможности определяются, – и в значительной мере расширяются – статистической интерпретацией [Donchenko, 2006] и модификацией классического определения введением объекта нечеткости [там же]. Собственно статистическая интерпретация представляет собой статистическую модель классической нечёткости.

В работе [Duda R.O, Hart P.,1972] задача выделения прямых в рамках ПХ(и) была сформулирована как задача кластеризации. Она была сформулирована в виде задачи распределения элементов выборки, по подходящим гиперплоскостям заданной размерности. Определение необходимого количества гиперплоскостей и их описание являлось одним из элементов задачи.

Заметим, что, хотя само появление ПХ (и) было ответом на потребности практики в выделении нескольких прямых на контурном изображении, вопрос, касающийся обоснования возможности его использования в «мультипараметрическом» варианте: для «оценивания» нескольких параметров, представленных в выборке, оставался открытым. Отсутствие категорического ответа на вопрос о возможности использования ПХ (и) для «мультипараметрического оценивания» определило предложение [Risse,1989] использовать технику ПХ (и) рекуррентно. В соответствии с его подходом на каждом шагу такого применения ПХ (и) выделяется одна прямая: та, которая «наиболее выраженным образом» представлена на изображении; из выборки удаляются все точки, выделенной прямой, после чего процедура выделения «наиболее выраженной прямой» повторяется в применении к редуцированной выборке.

В предлагаемом докладе представлено обоснование возможности применения ПХ (и) в его статистической формализации в виде ПГ (с) (статистической модели преобразования Хока) для использования в режиме «мультипараметрического оценивания", по крайней мере,– для достаточно естественных распределений ошибок наблюдения, к которым относятся симметричные, выпуклые по Андерсону [Anderson,1955] распределения. К последним относятся, в частности, гауссовские распределения с нулевым мат ожиданием и симметричные с треугольной плотностью. Кроме того, предложен вариант решения задачи кластеризации по гиперплоскостям из уже упомянутой выше работы [Duda R.O, Hart P.,1972] в ослабленном варианте: когда не задаётся размерность гиперплоскостей, по которым необходимо провести кластеризацию. Полный вариант решения упомянутой задачи можно найти в работе [Кириченко, Донченко., 2007], а также в соответствующем докладе упомянутых авторов в материалах текущей серии конференций в Варне. Применение техники псевдообращения по Муру – Пенроузу ([Moore,1920], [Penrose,1955]) (см., например [Алберт, 1977]) позволяет явным образом описывать формируемые гиперплоскости, а также – явный вид расстояния до них. Заметим, что в аналогичных задачах классификации для статистических совокупностей на основе корреляционных матриц (см., например, [Vapnik,1998]) соответствующе расстояния вычисляются приближённо в ходе выполнения подходящей рекуррентной вычислительной процедуры.

В работе приводится определение псевдообращения через SVD-представление матрицы, а также определение основных операторов, рассматриваемых в связи и определяемые через псевдообращение.

ПГ (с) в варианте «мультипараметрического» оценивания

Возможность использования ПГ (и) в «мультипараметрическом» варианте оценивания следует из базовых результатов математической теории ПГ (и) – из ПГ (с) - теории (см., например, [Donchenko, 2003]). Напомним, что в рамках ПГ (с), предметом анализа является выборка (последовательность наблюдений) $s_i = s_{x_i} = (x_i, y_i), i = \overline{1,N}$, в которой компоненты каждой из пар связаны соотношением: $y_i = g_{\theta_i}(x_i) + \varepsilon_i, i = \overline{1,N}$, где $g_\theta, \theta \in \Theta$ – параметрическое семейство отображений между евклидовыми пространствами, $\varepsilon_i, i = \overline{1,N}$ – независимые ошибки наблюдений. Преобразованием Хока выборки называется последовательность $L_i = L_{s_i} = \{\ \theta \in \Theta : y_i = g_\theta(x_i)\}, i = \overline{1,N}$, каждый элемент которой называется преобразованием Хока соответствующего наблюдения. Основой анализа ПГ (с) является аккумуляторная функция (АФ) $A_N(\pi)$ или нормированная АФ (НАФ) $A_N^{(f)}(\pi)$ как функция множества, заданная на подходящей совокупности $\pi \in \Pi$ (совокупности множеств-зондов) подмножеств множества параметров Θ соотношениями, соответственно

$$A_N(\pi) = \sum_{i=1}^{N} \delta(\pi \cap L_i)\ ,\ A_N^{(f)}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \delta(\pi \cap L_i)\ .$$

Основой вывода о корректности использования ПГ (и) в «мультипараметрическом» варианте являются теоремы о предельном поведении нормированных значений НАФ при неограниченном увеличении объёма выборки и неограниченном уменьшении геометрических размеров множеств-зондов. В качестве множеств-зондов Π в рассматриваемой ситуации рассматривается либо совокупность всех замкнутых шаров $S_\rho(\theta) = \theta + \rho S$ радиуса $\rho > 0$ с центром в $\theta \in \Theta$, либо множества $\pi = V_\rho(\theta) = \theta + \rho V$ с замкнутым симметричным выпуклым множеством V единичного радиуса, содержащим шар ненулевого радиуса с центром в нуле. Такая нормировка может рассматриваться как для общего варианта размерности

евклидовых пространств, так и, в частности, для скалярных значений функций параметрического семейства. Ниже приведены две теоремы именно для этого случая.

Теорема 1. Пусть для каждого $x \in X$ функции параметрического семейства со скалярными значениями являются непрерывно дифференцированными относительно θ в области определения Θ; пусть также плотности распределений ошибок наблюдений $h_{x,\theta^{(0)}}(z)$, $x \in X, \theta^{(0)} \in \Theta$ являются непрерывными по z и определяются только комбинацией аргумент-параметр, представленной в наблюдении; пусть в наблюдениях представлено конечное множество параметров $\Theta^{(0)} = \{\theta_1^{(0)},...,\theta_K^{(0)}\}$ та конечное множество значений аргументов $X^{(0)} = \{x_1^{(0)},...,x_K^{(0)}\}$ с предельными ненулевыми частотами $f_{k,q}^{(0)}$, $k = \overline{1,K}, q = \overline{1,Q}$ для соответствующих комбинаций аргумент параметр в выборке. Тогда

$$\lim_{\rho \to 0} \rho^{-1} A_\infty(S_\rho(\theta)) = 2\sum_{k=1}^{K}\sum_{q=1}^{Q} f_{kq}^{(0)} h_{kq}(g_\theta(x_q^{(0)}) - g_{\theta_k^{(0)}}(x_q^{(0)}))\,\|\,grad_\theta g_\theta(x_q^{(0)})\,\|, \quad (1)$$

где $h_{kq} = h_{x_q^{(0)},\theta_k^{(0)}}$, $k = \overline{1,K}, q = \overline{1,Q}$,

а A_∞ - предельное значение НАФ при неограниченном увеличении объёма выборки.

Если множество S заменить на V, то соответствующий результат приобретает вид.

Теорема 2. В условиях и обозначениях теоремы 1

$$\lim_{\rho \to 0} \rho^{-1} A_\infty(S_\rho(\theta)) = 2\sum_{k=1}^{K}\sum_{q=1}^{Q} \varphi_q f_{kq}^{(0)} h_{kq}(g_\theta(x_q^{(0)}) - g_{\theta_k^{(0)}}(x_q^{(0)}))\,\|\,grad_\theta g_\theta(x_q^{(0)})\,\|, \quad (2)$$

где $\varphi_q : \varphi_q \in (0,1), q = 1,Q$.

Несложно убедиться, что в (1), (2) множитель - плотность распределения для $\theta \in \Theta$ в ПГ (с) отвечает преобразованию Хока истинного (без ошибок) наблюдения, «размытого» плотностью ошибки наблюдения. Если плотность ошибки наблюдения имеет максимум в нуле, как это имеет место для симметричных, выпуклых по Андерсону [Anderson,1955] распределений, то максимумы предельной функции будут отвечать максимумам для «безошибочной» НАФ. И, таким образом, ПХ можно использовать в «мультипараметрическом» режиме, поскольку в ПХ оценивание параметров, представленных в выборке осуществляется на основе максимумов АФ или НАФ.

Кластеризация по гиперплоскостям: 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 гиперплососте й $\Gamma(k) = x_k + L_k, k = 1,2$.

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 о 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 (см. ниже), так и при его исключении [Кириченко, Донченко., 2007]. 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 насматриваемой задачи, отметим также, что формулы вычисления расстояний носят рекуррентный по исключаемым элементам вид. Достижение указанной рекуррентности обеспечивается теорией возмущения пседообратных матриц [Кириченко, Лепеха, 1997].

Кластеризация по гиперплоскостям: вспомогательные определения и утверждения

Пседообращение A^+ по Муру-Пенроузу 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

$A^+=\sum_{i=1}^{r}x_i y_i^T\lambda_i^{-1}$, которое определяется элементами SVD-представления исходной матрицы:

$A=\sum_{i=1}^{r}y_i x_i^T\lambda_i$, в котором: $\lambda_1^2\geq\ldots\lambda_r^2>0$ – 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^TA: 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 $L^{\perp}_{A^T}$, $L^{\perp}_{A}$. Очевидны образом, $L^{\perp}_{A^T} = KerA$. Заметим также, что каждое из подпространств L_{A^T}, L_A является линейной оболочкой соответственно векторов-столбцов и векторов-строк матрицы A.

Не ограничивая общности, будем считать. Что $x(1),\ldots,x(n)$ из пространства признаков R^m могут 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). Подпространства могут задаваться, как множества значений подходящих матриц $A(k), k = 1,2$. В этом случае будут использоваться обозначения как $L_k = L_{A(k)}, k = 1,2$.

Лемма. Пусть подпространства гиперплоскостей являются множествами значений матриц $A(k), k = 1,2$ соответственно. Тогда 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

$$\rho(x, \Gamma_k) = (x \quad - \quad x_{\mathrm{k}}))^T Z(A^T(k))(x - x_{\mathrm{k}}), k = 1,2$$

Кластеризация по гиперплоскостям: основной алгоритм

Сам алгоритм кластеризации по гиперплоскостям 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. Опребделение 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
6. гиперплоскостей с использованием результатов леммы 2.
7. Перераспределение векторов между совокупностями и повторением рекуррентных шагов.

Перераспределение элементов можно реализовывать разными способами с введением подходящих параметров алгоритма

Литература

[Anderson, 1955] Anderson T.W. The integral of a symmetric random functions over symmetric convex sets // Proc. Amer. Math. Soc. – № 6.–1955.–P.170-175.

[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.

[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.

[Donchenko, 2006] Donchenko V.S. Fuzzy sets: abstraction axiom, statistical interpretation, observation of fuzzy sets. // International Journal on Information Theory and Applications.–Vol.13, №3– 2006(Bulgaria). – 233-238.

[Duda R.O, Hart P., 1972] Duda R.O, Hart P. Using the Hough Transform to detect Lines and Curves in pictures.//Communications of ACM.–v.15.–1972,–p.11-15.

[Hough, 1962] Hough P.V.C. Method and Means for Recognizing Complex Patterns. - U.S. Patent 3069354, 1962.

[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.

[Risse, 1989] Risse T. Hough Transformation for line recognition: complexity for evidence accumulation and cluster detection.// Computer Vision, Graphics and Image Processing. – Vol.46 –1989.–p.327-345.

[Xu, Oja, 1992] Xu L., Oja E. Further Developments on RHT: Basic Mechanisms, Algorithms, and Computational Complexities.// ICPR (92): Proceedings, International Conference on Pattern Recognition. – Vol.1. –1992.– P.125-132.

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

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

Информация об авторе

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