

最小一乘聚类中心模型及算法

蔡广基¹ 严玉清^{2,3} 李师贤²

(广东湛江师范学院信息学院 湛江 524048)¹ (中山大学计算机科学系 广州 510275)²

(广东外语外贸大学信息学院 广州 510420)³

摘要 本文首次提出最小一乘聚类中心的数学模型,对于不可微目标函数推导了最优解的条件,同时构造了最小一乘聚类中心的逼近点列,给出了相应的算法,并研究了一个不带类别标记的两类分类问题。最后讨论了最小一乘聚类中心能抵御样本中噪声的数学机理,并通过例子说明了此方法的有效性。

关键词 最小一乘方法,聚类中心,不可微函数

Model and Algorithm of Cluster Center Based on the Least Norm Method

CAI Guang-ji¹ YAN Yu-qing^{2,3} LI Shi-xian²

(Information College, Zhanjiang Normal College, Zhanjiang 524048, China)¹

(Department of Computer Science, Zhongshan University, Guangzhou 510275, China)²

(Information College, Guangdong University of Foreign Studies, Guangzhou 510420, China)³

Abstract In this paper, we first put forward the mathematical model of the least norm cluster center and deduced the condition of the optimal solution to the corresponding nondifferential goal function, and gave an approximating point sequence to the optimal solution. we still discussed the geometric explanation that the least norm cluster center can oppose the sample errors. Finally we studied the classification problem of two classes with non-marked sample points.

Keywords Least norm method, Cluster center, Nondifferential function

1 引言

聚类分析也称群分析、点群分析,是一种研究分类的方法。常见的聚类方法有 K 均值聚类法、模糊 C 均值聚类法、系统聚类法、指标聚类等。在具体的聚类方法指导下,对样本空间进行了分类,接下来一件非常重要的事情是计算聚类中心。通常对聚类中心的计算都是基于某种形式的最小二乘方法,即目标函数中含有聚类中心与数据点间的欧氏距离的平方项。最小二乘方法的特点是数学处理简单,可得到聚类中心的解析表达式,但求解的聚类中心易受极端样本点的影响,抗噪声能力差。

本文研究用最小一乘确定聚类中心,聚类中心的目标函数只含有聚类中心与数据点间欧氏距离的一次项。这样所确定的聚类中心不易受到极端样本点的影响,抗噪声能力强,但需要使用迭代的方法来确定聚类中心。

下一节给出最小一乘聚类中心的数学模型,推导不可微目标函数达到最小的条件。第 3 节构造最小一乘聚类中心的逼近点列,给出相应的算法。最后一节对最小一乘聚类中心的数学特点作了讨论,并给出具体应用例子,与其它方法作了比较,说明了该方法的有效性。

2 最小一乘数学模型及最优解的条件

设聚类分析中,某一类含有 m 个样本点:

$$x_j \in R^n, j=1, 2, \dots, m$$

聚类中心 c 满足:

$$\min_{c \in R^n} f(c), f(c) = \sum_{j=1}^m \|x_j - c\| \quad (1)$$

其中 $\|\cdot\|$ 为 n 维空间中的欧氏距离。

称

$$F(c) = \sum_{\substack{j=1 \\ x_j \neq c}}^m \frac{x_j - c}{\|x_j - c\|} \quad (2)$$

为目标函数 $f(c)$ 的广义负梯度。在样本点处, $f(c)$ 不可微,也就不存在通常意义下的梯度,但其广义负梯度是存在的。故(1)式是一个不可微的优化问题。

引理 1 对任意 $c, c' \in R^n$, 有

$$F(c)^T (c' - c) \geq f(c) - f(c') + |\{j; x_j = c\}| \cdot \|c' - c\|$$

其中 $|\cdot|$ 为集合元素的个数。

证:

$$\begin{aligned} F(c)^T (c' - c) &= \sum_{\substack{j=1 \\ x_j \neq c}}^m \frac{(x_j - c)^T}{\|x_j - c\|} [(x_j - c) - (x_j - c')] \\ &\geq \sum_{\substack{j=1 \\ x_j \neq c}}^m \frac{\|x_j - c\|^2 - \|x_j - c\| \cdot \|x_j - c'\|}{\|x_j - c\|} \\ &= \sum_{j=1}^m \|x_j - c\| - \sum_{j=1}^m \|x_j - c'\| + \sum_{x_j=c} \|x_j - c'\| \\ &= f(c) - f(c') + |\{j; x_j = c\}| \cdot \|c' - c\| \end{aligned}$$

下面定理给出(1)式的最优解的条件。

定理 1 如果

$$\|F(c)\| \leq |\{j; x_j = c\}| \quad (3)$$

则 c 是(1)式的最优解。上式右端表示与 c 点重合的样本点的个数。

证:在引理 1 中,令 $c' = c + dc$, 其中 d 为某一实数,则 $c' - c = dc$,

$$F(c)^T dc \geq f(c) - f(c + dc) + |\{j; x_j = c\}| \cdot \|dc\|$$

因而

$$\begin{aligned} f(c+dc) &\geq f(c) - F(c)^T dc + |\{j; x_j = c\}| \cdot \|dc\| \\ &\geq f(c) + (|\{j; x_j = c\}| - \|F(c)\|) \cdot \|dc\| \\ &\geq f(c) \end{aligned}$$

故 c 是目标函数 f 的极小点。易知 f 是 R^n 上的凸函数, 因此 c 也是 f 的最小点。

3 聚类中心的逼近点列

引理 2 若 c 是(1)式的最优解, c' 不是, 则 $F(c')^T(c-c') > 0$

证: 引理 1 中, 调换 c 与 c' 的位置, 得

$$F(c')^T(c-c') \geq f(c') - f(c) + |\{j; x_j = c'\}| > 0$$

构造聚类中心 c 的逼近点列如下:

$$\begin{cases} c_0 \in R^n \\ c_{k+1} = c_k + \mu \cdot F(c_k), k=0, 1, 2, \dots \end{cases} \quad (4)$$

其中 $0 < \mu < 1$ 为常数。

定理 2 设 c 是(1)式的最优解, 而 c_k 不是最优解, 则当

$$0 < \mu < 2 \frac{F(c_k)^T}{\|F(c_k)\|^2} (c - c_k) \quad (5)$$

时, 有

$$\|c_{k+1} - c\| < \|c_k - c\|$$

证: 引理 2 保证(5)式右端为正。

$$\begin{aligned} \|c_{k+1} - c\|^2 &= \|c_k - c + \mu \cdot F(c_k)\|^2 \\ &= \|c_k - c\|^2 + \mu^2 \|F(c_k)\|^2 + 2\mu \frac{F(c_k)^T}{\|F(c_k)\|^2} (c - c_k) \\ &< \|c_k - c\|^2 \end{aligned}$$

定理 2 表明, 当 μ 足够小时, 由(4)式定义的点列一步步逼近聚类中心 c 。由于 f 是 R^n 上的凸函数, 点列 $\{c_k\}$ 收敛。

构造聚类中心 c 的逼近算法如下:

I. 初始化

i) 选取参数 μ , 足够小的正数 ϵ ;

ii) 任取 $c \in R^n$, 如 $c = \frac{1}{m}(x_1 + \dots + x_m)$ 。

II. 计算 $F_i = F(c)$ 和 $t_i = |\{j; x_j = c\}|$

i) $t_i = 0; F_i = 0$;

ii) 对 $j = 1, 2, \dots, m$, 循环

a) $x_j = x_j - c$

b) 若 $\|x_j\| < \epsilon$, 则 $t_i = t_i + 1$;

否则 $F_i = F_i + \frac{x_j}{\|x_j\|}$;

III. 若 $\|F_i\| < t_i + \epsilon$, 则 c 为所求, 结束; 否则 $c_i = c + \mu F_i$, 转 II。

4 最小一乘聚类中心的数学特点

4.1 最小一乘中心能抵御噪声影响

图 1 所示的 4 个样本点中, 右端那一个为受噪声影响的极端样本点。按上一节的算法计算出的最小一乘中心只受到极端样本点的微小影响, 而最小二乘中心(样本点的算术平均值)受到了极端样本点的显著影响。

4.2 广义负梯度的几何意义

由(2)式定义的广义负梯度为

$$F(c) = \sum_{\substack{j=1 \\ x_j \neq c}}^m u_j \quad (6)$$

其中 $u_j = \frac{x_j - c}{\|x_j - c\|}$ 为单位矢量。若矢量 $u_1, u_2, \dots, u_m$ 的始

点在最小一乘聚类中心 c 处, 则它们的终点落在以 c 为中心、半径为 1 的超球面上, 如图 2 所示。

图 2 中, 若样本点 x_j 受噪声影响而移至 x'_j , 则 u_j 不变, 因而由(6)式定义的 $F(c)$ 的值不变, 从而由(3)式确定的聚类中心 c 也不改变。这就是最小一乘聚类中心能抵御极端样本点的影响、抗噪声能力强的数学机理。

5 最小一乘聚类中心用于聚类分析

研究图 3 中不带类别标记的分类问题, 9 个样本点可以区分为两个类区: ω_1 和 ω_2 。图中标出的最小一乘中心和最小二乘中心都是对全部 9 个样本点计算而得到。最小二乘中心处于两个类区的中间区域, 而最小一乘中心则处于样本点较密集的一类区 ω_1 中。

最小一乘中心用于不带类别标记的两类聚类分析, 方法如下: 对全部样本点计算最小一乘聚类中心, 它处于样本点较密集的一个类区, 由此确定该类区的样本点, 余下即为另一类区的样本点。

本文提出的最小一乘聚类中心也可用于不带类别标记的多个类区的聚类分析, 以及用于带类别标记的聚类分析, 这将另文讨论。

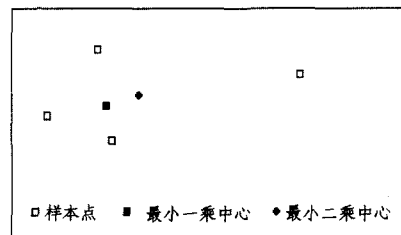


图 1 最小一乘和最小二乘方法比较

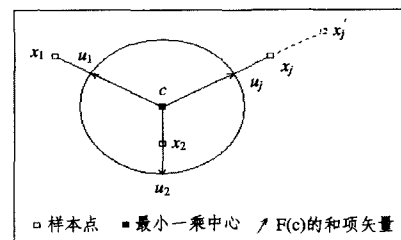


图 2 广义负梯度的几何意义

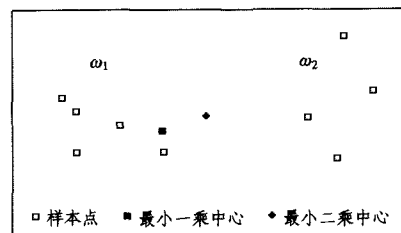


图 3 最小一乘方法用于聚类分析

参考文献

- [1] 边肇祺, 等. 模式识别. 北京: 清华大学出版社, 2001
- [2] 张智星, 等. 神经-模糊和软计算. 西安: 西安交通大学出版社, 2000
- [3] 高新波. 模糊聚类分析及其应用. 西安电子科技大学, 2004