关于聚类有效性函数 $FP(u,c)$ 的研究

于 剑,程乾生
(北京大学数学科学学院,北京 100871)

摘 要: 本文对已有的划分系数 $P(u,c)$ 作了进一步研究,指出 $P(u,c)$ 可以作为聚类有效性函数使用,其性能与划分系数 $V_{pc}(u,c)$ 相仿. 据此对原有的聚类有效性函数 $FP(u,c)$ 作了一定的理论分析,并就本文使用的数据进行了计算机模拟. 理论分析与计算机模拟得出了同样结论: $FP(u,c)$ 作为 FCM 算法的聚类有效性函数是不合适的.

关键词: 模糊聚类; 划分系数; 聚类有效性
中图分类号: TP391.41 **文献标识码**: A **文章编号**: 0372-2112 (2001) 07-0899-03

On Cluster Validity Function FP(u, c)

YU Jian, CHENG Qian-sheng
(*School of Mathematical Science, Peking University, Beijing* 100871, *China*)

Abstract: Based on the property of $P(u,c)$, we can conclude that the performance of $P(u,c)$ and $V_{pc}(u,c)$ as cluster validity function is alike. At the same time, $FP(u,c)$ is not suitable for cluster validity function according to theoretical analysis and experiments.

Key words: fuzzy clustering; partition coefficient; cluster validity function

1 引言

聚类算法是一种无监督的学习算法,事先对给定数据的结构一无所知,无论用什么算法聚类,其聚类结果的合理性都有待评价. 例如,FCM 算法只能保证收敛到目标函数 J_m 的局部极值,这样不同的聚类数、不同的初值就可能得到不同的聚类结果. 如何评价此时得到的不同聚类结果? 这被称为聚类有效问题. 自 1974 年,Bezdek[2] 提出了上述问题并给出了第一个聚类有效性函数 $V_{pc}(u,c)$ (同时,Bezdek 也称其为划分系数(Partition Coefficient)) 以来,文献中已经提出了很多检验聚类有效性的函数,如文[3]中提出的 $V_{FS}(U,V,X,c)$,文[4]中提出的 $V_{XB}(U,V,X,c)$ 等. 文[1]中提出了新的划分系数 $P(U,c)$ 和聚类有效性函数 $FP(U,c)$,并指出: (1) 直接以 $P(U,c)$ 作为聚类有效性函数不是可取的. (2) 聚类有效性函数 $FP(U,c)$ 具有良好判决功能和鲁棒性. 本文对 $P(U,c)$ 和 $FP(U,c)$ 理论分析与实验结果表明,上述结论似乎缺乏成立的理由.

本文采用 1974 年 Bezdek 在文[2]中推广的 FCM 算法,即:设 $X=\{X_1,X_2,\cdots,X_n\}\subset R^p$,期望使目标函数 J_m 极小以将 X 分为 c 个模糊子集(类),即: $\min\{J_m(U,V;X)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m\|x_k-v_i\|_A^2\}$ 其中: $U=\{u_{ik}\}_{c\times n}$ 为划分矩阵, $\forall i\in\{1,2,\cdots,c\}$, $\sum_{k=1}^{n}u_{ik}=1$, $u_{ik}\geq 0$, $0<\sum_{i=1}^{c}u_{ik}<n$; $V=\{v_1,v_2\cdots,v_c\}$ 为类中心, $\forall i\in\{1,2,\cdots,c\}$, $v_i\in R^P$, $1<m<+\infty$. 其具体算法叙述见文献[2].

Bezdek 在文[2]中定义了如下划分系数(partition coefficient):

$$V_{pc}(U,c)=(1/n)*\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^2$$

文献[2]指出,可以以 $V_{pc}(U,c)$ 作为聚类有效性函数,即如果存在 (U^*,c^*) 满足下式, $V_{pc}(U^*,c^*)=\max_{2\leq c\leq n-1}\{\max_{\Omega_c}V_{pc}(U,c)\}$,则以 (U^*,c^*) 为"最优"的聚类结果. 其中, $\Omega_c=\{$聚类数为 c 时,FCM 算法得到的所有最佳划分矩阵 $U\}$.

文献[1]中,给出了一个新的划分系数 $P(U,c)$,定义了如下聚类有效性函数 $FP(U,c)$:

$$FP(U;c)=V_{pc}(U;c)=P(U;c);$$

其中: $P(U;c)=(1/c)*\sum_{i=1}^{c}(\sum_{k=1}^{n}u_{ik}^2/\sum_{k=1}^{n}u_{ik})$

2 对 $V_{pc}(U,c)$、$P(U,c)$ 和 $FP(U,c)$ 的理论分析

N. R. Pal 与 J. C. Bezdek 在文[5]中,指出对于 FCM 算法来说:权重系数 m 的较为合理的取值范围为 1.5 到 2.5 之间. 一般取 $m=2$ 较为常见. 在本文中,始终取 $m=2$.

下面,为了研究 $FP(U,c)$ 的性质,先给出函数 $P(U,c)$ 与 $Vpc(U,c)$ 具有的如下性质.

收稿日期:2000-01-15;修回日期:2001-01-09
基金项目:国家自然科学基金(No. 69872003)

定理 1 当 $1 < c < n$ 时,

(1) $\frac{1}{c} \le P(u, c) \le 1$

(2) $P(u, c) = 1 \Leftrightarrow u$ 是硬划分;

(3) $P(u, c) = \frac{1}{c} \Leftrightarrow u[a_1, a_2, \cdots, a_c]^T \times \overbrace{[1, 1, \cdots, 1]}^{n}$,

$\forall 1 \le i \le c; a_i > 0, \sum_{i=1}^{c} a_i = 1$

证明 见本文附录.

定理 2 当 $1 < c < n$ 时,

(1) $\frac{1}{c} \le V_{pc}(u, c) \le 1$

(2) $V_{pc}(u, c) = 1 \Leftrightarrow u$ 是硬划分;

(3) $V_{pc}(u, c) = \frac{1}{c} \Leftrightarrow u = [\frac{1}{c}]$

证明 见文献[2].

显然由上述定理 1、2 可知, $P(u, c)$ 具有与 $V_{pc}(u, c)$ 类似的性质,仿照 $V_{pc}(u, c)$ 定义最优聚类结果的方式,可以给出 $P(u, c)$ 计算最优聚类结果的方法. 因此在理论上,文献[1]所断言的如下结论未必成立:直接以 $P(u, c)$ 作为聚类有效性函数不是可取的.

文献[1]又指出:一个好的分类是 $V_{pc}(u, c)$ 与 $P(u, c)$ 的差别最小. 因此设计了聚类有效性函数 $FP(u, c)$,并认为如果有 (u^*, c^*) 满足 $FP(u^*, c^*) = \min_c\{\min_{\Omega_c} FP(u, c)\}$,则 (u^*, c^*) 为"最优"的聚类结果. 但是,如果注意到 $\forall u$, $FP(u, c)$ 并不恒成立,则其设计 $FP(u, c)$ 的初衷与 (u^*, c^*) 的计算规则是相矛盾的. 因为有可能存在 (u^*, c^*) 满足 $FP(u^*, c^*) = \min_c\{\min_{\Omega_c} FP(u, c)\}$,而此时 $V_{pc}(u^*, c^*)$ 和 $P(u^*, c^*)$ 的差别较大. 如果再注意到定理 1、2 的结论,可知在划分最分明与划分最模糊的情形下, $FP(u, c)$ 皆为零. 这样可得到如下一个推断: $FP(u, c)$ 作为聚类有效性函数其性能是较差的.

下文将用实际数据实验来验证上述论断.

3 实验方法、数据与实验结果

3.1 实验方法

(1) 数据 m 有 k 个样本,样本 $m_i = [m_{i1}, m_{i2}, \cdots, m_{il}]$,其最大聚类数为 c_{max},用 FCM 算法求出最佳划分矩阵 d 次.

(2) 固定一大于 1 而不大于 c_{max} 的聚类数 c. 用 FCM 算法求出数据 m 的 d 次最佳划分系数 u,分别计算 $V_{pc}(u, c)$、$P(u, c)$ 和 $FP(u, c)$,选出其中最大的 $V_{pc}(u, c)$、$P(u, c)$ 和最小的 $FP(u, c)$,分别记入向量 vpc, puc 和 fpuc 之中,作为各自的第 c 个分量.

(3) 令聚类数 c 取遍从 2 到 c_{max} 的所有整数,重复上述第二步.

3.2 实验数据、结果

数据 1:共有 200 样本,分 3 类,其中以 $[0, 0]$ 为类中心的有 100 个样本,以 $[4, 4]$ 为类中心的有 50 个样本,以 $[4, -4]$ 为类中心的有 50 个样本,每类样本皆由类中心加服从 $N(0, 1)$ 分布的白噪声生成. 数据 1 的样本分布图见图 1.

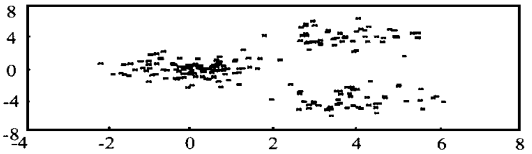

图 1 数据 1 的样本分布图

对数据 1 最大聚类数 $c_{max} = 9$, $d = 10$ 时进行上述实验. 实验结果如表 1.

表 1

C	2	3	4	5	6	7	8	9
fpuc	0.0100	0.0057	-0.0039	0.0230	0.0167	0.0106	0.0138	0.0118
vpc	0.7886	0.8369	0.7271	0.6745	0.6273	0.5959	0.5718	0.5511
puc	0.7733	0.8312	0.7309	0.6515	0.6094	0.5788	0.5507	0.5329

数据 2:iris 数据见文献[6].

(1) 对 iris 数据在最大聚类数 $c_{max} = 10$, $d = 1$ 时进行上述实验. 进行多次实验,选择与文献[1]中表 2 差别较小的实验结果列成如下表 2.

表 2

	2	3	4	5	6	7	8	9	10
fpu	0.0053	0.0031	0.0229	0.0449	0.0572	0.0068	0.0713	0.0221	0.0230
vpc	0.8922	0.7834	0.7068	0.6657	0.6071	0.5544	0.5568	0.5113	0.4934
puc	0.8870	0.7803	0.6839	0.6208	0.5499	0.5476	0.4855	0.4892	0.4704

(2) 对 iris 数据在最大聚类数 $c_{max} = 10$, $d = 10$ 时进行上述实验. 实验结果如表 3.

表 3

C	2	3	4	5	6	7	8	9	10
fpuc	0.0053	0.0031	0.0089	-0.0037	0.0026	0.0034	-0.0016	0.0005	0.0018
vpc	0.8922	0.7834	0.7068	0.6657	0.6071	0.5925	0.5578	0.5093	0.4934
puc	0.8870	0.7803	0.6839	0.6283	0.5927	0.5480	0.5182	0.4964	0.4742

数据 3:文献[1]中的数据 I

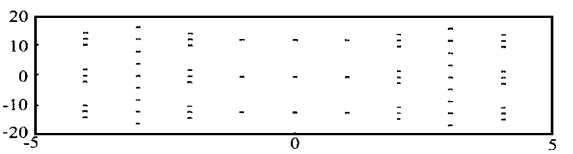

图 2 数据 I 的分布图

为了说明本文的实验方法与文献[1]中使用的方法相同,对数据 I 在最大聚类数 $c_{max} = 8$, $d = 5$ 时进行上述实验. 实验结果如表 4.

表 4

c	2	3	4	5	6	7	8
fpuc	0	-0.000335	0.018272	0.011096	-0.000135	0.010173	0.011151
vpc	0.8045	0.7932	0.7235	0.6977	0.6875	0.6405	0.6123
puc	0.8045	0.7935	0.7053	0.6866	0.6876	0.6296	0.5896

4 实验分析和结论

就数据一来说,根据表 1 和图 1, $V_{pc}(u, c)$、$P(u, c)$ 都选出了正确聚类数为 3, $FP(u, c)$ 却错误地选择了聚类数为 4. 多次重复本次实验,都得到与表 1 类似的数据. 因此, $FP(u, c)$ 作为聚类有效性函数就数据 1 来说是不合适的,而 $V_{pc}(u,$

$c)$、$P(u,c)$ 的表现可以说是一样良好.

对 iris 数据,根据表 2、表 3,$V_{pc}(u,c)$、$P(u,c)$ 选出的最佳聚类数均为 2,次佳聚类数为 3. 据文献[5]等,iris 数据实际为三类,但有两类存在交叉,就此看来,$V_{pc}(u,c)$、$P(u,c)$ 的表现应该说差强人意. 据表 2,对 iris 数据 $FP(u,c)$ 选择的最佳聚类数为 3,次佳聚类数为 2. 这与文献[1]中的结论一致. 而据表 3,对 iris 数据 $FP(u,c)$ 选择的最佳聚类数为 5,次佳聚类数为 10. 如果考虑到表 2、表 3 分别代表的实验,FCM 算法的特性和 $FP(u,c)$ 选择最佳聚类数的准则,可以认为表 3 表示的实验结果比表 2 表示的实验结果更有代表性,因此,根据表 2 表示的实验结果选择的最佳聚类数有很大的偶然性,以此为依据评价 $FP(u,c)$ 作为聚类有效性函数性能并不合适. 而多次重复本次实验,结果与表 3 类似. 因此,就 iris 数据来说,$FP(u,c)$ 作为聚类有效性函数的表现并不象文献[1]中所指出的那样好,实际上据本文的实验结果,对 iris 数据 $FP(u,c)$ 作为聚类有效性函数并不合适.

本文对文献[1]中的数据 I 进行多次本文设计的实验,得到的结果与表 4 几乎没有差别,注意到表 4 与文献[1]中表 1 非常相似,这说明本文中的实验方法与文献[1]中的实验方法并无二致. 同时对数据 I 来说,文献[1]已经指出了其可分为 2 类,3 类或 6 类,从数据 I 的分布图可明显看出此点. 因此,聚类有效性函数选出的最佳聚类数为 2,3 或 6 都算达到了目的. 从表 4 来看,$V_{pc}(u,c)$、$P(u,c)$ 和 $FP(u,c)$ 对数据 I 的表现难分优劣.

总上所述,实验结果较充分地证实了上文的理论推断:$V_{pc}(u,c)$、$P(u,c)$ 作为聚类有效性函数的性能是非常接近的,而 $FP(u,c)$ 作为聚类有效性函数是不合适的.

参考文献:

[1] 范九伦,裴继红,谢维信. 基于可能性分布的聚类有效性 [J]. 电子学报,1998,26(4):113 - 115.

[2] J C Bezdek. Cluster validity with fuzzy sets [J]. Journal of Cybernetics,1974:3(3):58 - 72.

[3] Y Fukuyanma, M Sugeno. A new method of choosing the number of clusters for the fuzzy c-means method [A]. in Proc. 5th Fuzzy Syst. Symp [C],1989:247 - 250(in Japanese).

[4] X L Xie, G Beni. A validity method for fuzzy clustering [J]. IEEE Trans Pat Anal. Mach. Intell,1991,13(8):841 - 847.

[5] N R Pal, J C Bezdek. On cluster validity for the fuzzy c - mean model [J]. IEEE Trans fuzzy systems,1995,3(3):370 - 379.

[6] E Aanderson. The irises of the Gaspe penisnsula [J]. Bull. Aner. IRIS. Soc. 1935,59(3):381 - 406.

附录

定理 1 的证明

(1) 不等式右边显然成立. 只须证明不等式左边成立.

由于 $\forall 1 \le i \le n, a_i \ge 0, \dfrac{a_1^2 + a_2^2 + \cdots + a_n^2}{n} \ge \left(\dfrac{a_1 + a_2 + \cdots + a_n}{n}\right)^2$

故

$$P(u;c) = (1/c) \times \sum_{i=1}^{c}\left(\sum_{k=1}^{n} u_{ik}^2 / \sum_{k=1}^{n} u_{ik}\right)$$

$$\ge (1/c) \times \sum_{i=1}^{c}\left(n \times \left(\frac{\sum_{k=1}^{n} u_{ik}}{n}\right)^2 \times \left(\sum_{k=1}^{n} u_{ik}\right)^{-1}\right)$$

$$\ge (1/c) \times \sum_{i=1}^{c}\left(\sum_{k=1}^{n} u_{ik}/n\right) = \frac{1}{c}$$,证毕.

(2) 参见文献[2].

(3) 证明:利用拉格朗日乘子法. 设,

$\lambda = [\lambda_2, \lambda_2, \cdots, \lambda_n]$, $u_i = [u_{i1}, u_{i2}, \cdots, u_{in}]$

$S_i = \sum_{k=1}^{n} u_{ik}$, $T_i = \sum_{k=1}^{n} u_{ik}^2$, $\lambda S = \sum_{k=1}^{n} \lambda_k$,则 $u = [u_1^T, u_2^T, \cdots, u_c^T]^T$,

记 $L(\lambda, u) = P(u,c) - \sum_{k=1}^{n} \lambda_k \times \left(\sum_{i=1}^{c} u_{ik} - 1\right)$ 令 L 在 (λ, u) 的偏微分为零以求出最小值:

$$\frac{\partial L}{\partial \lambda_k} = \sum_{i=1}^{c} u_{ik} - 1 = 0;\ \forall 1 \le k \le n \qquad (A)$$

$$\frac{\partial L}{\partial u_{ik}} = \frac{2 \times u_{ik}}{c \times S_i} - \frac{T_i}{c \times S_i^2} - \lambda_k = 0;\ \forall 1 \le k \le n \qquad (B)$$

由式 (B) 可得:$\dfrac{2 \times u_{ik}}{c \times S_i} - \dfrac{T_i}{c \times S_i^2}$ $(B1)$

对式 $(B1)$ 从 1 到 n 对 λ_k 求和得:$\lambda S = \dfrac{2}{c} - \dfrac{n \times T_i}{c \times S_i^2}$ $(B2)$

对式 $(B1)$ 整理可得:$c \times S_i \times \lambda_k = 2 \times u_{ik} - \dfrac{T_i}{S_i}$ $(B3)$

对式 $(B3)$ 从 1 到 c 对下标 i 求和得:

$$n \times c \times \lambda_k = 2 - \sum_{i=1}^{c} \frac{T_i}{S_i} \qquad (B4)$$

对式 $(B4)$ 整理可得:$\lambda_k = \dfrac{2}{n \times c} - \dfrac{1}{n \times c} \times \sum_{i=1}^{c} \dfrac{T_i}{S_i}$ $(B5)$

对式 $(B5)$ 从 1 到 n 对 λ_k 求和得:$\lambda S = \dfrac{2}{c} - \dfrac{1}{c} \times \sum_{i=1}^{c} \dfrac{T_i}{S_i}$ $(B6)$

由式 $(B2)$ 与 $(B6)$ 联立可得:$\dfrac{n \times T_i}{S_i^2} = \sum_{i=1}^{c} \dfrac{T_i}{S_i}$ $(B7)$

由式 $(B5)$ 与 $(B7)$ 联立可得:$\lambda_k = \dfrac{2}{n \times c} - \dfrac{T_i}{c \times S_i^2}$ $(B8)$

由式 $(B1)$ 与 $(B8)$ 联立可得:$u_{ik} = \dfrac{S_i}{n}$, $\forall \le k \le n$ $(B9)$

令 $a_i = \dfrac{S_i}{n}$, $\forall 1 \le i \le c$,则本结论得证.

作者简介:

于　剑　男. 1969 年生于山东淄博. 博士,讲师. 1991 年,1994 年和 2000 年在北京大学分别获得学士,硕士和博士学位. 现在北方交通大学计算机系人工智能研究所工作. 主要研究兴趣包括模式识别、数据挖掘和模糊信息处理等.

程乾生　男. 1940 年生于安徽怀宁. 北京大学数学科学学院信息科学系教授,博士生导师,主要研究领域为信号处理和模式识别. 中国信号处理协会副理事长,中国工业与应用数学学会常务理事兼学术委员会主任.