划分系数和总变差相结合的聚类有效性函数

范九伦,吴成茂
(西安邮电学院信息与控制系,陕西西安 710061)

摘 要: 划分系数是聚类有效性检测中常用方法之一.针对划分系数存在的严重不足,本文从一个新的角度对划分系数进行修改.结合数据的模糊划分得到的总变差,提出了二个新的聚类有效性标准.实验结果表明,本文提出的方法具有良好的分类性能.

关键词: 模糊划分; 模糊-均值聚类; 聚类有效性; 划分系数
中图分类号: O235 **文献标识码**: A **文章编号**: 0372-2112 (2001) 11-1561-03

Clustering Validity Function Based on Partition Coefficient Combined with Total Variation

FAN Jiu-lun, WU Cheng-mao
(Department of Informotion and Control, Xi'an Institute of Post and Telecommunications, Xi'an, Shanxi 710061, China)

Abstract: Partition coefficient is a common method for clustering validity test. Based on the serious limitation of partition coefficient, modified definitions are given through a new point of view. Two new clustering validity functions are proposed using partition coefficient combined with total variation in fuzzy-partition. Experimental results show that the new methods have good classification performance.

Key words: fuzzy partition; fuzzyc-mean clustering; clustering validity; partition coefficient

1 引言

模糊 c-均值聚类算法[1]是非监督模式识别中的一个基本方法,在应用模糊 c-均值聚类算法时,一个事先需要确定的参数是数据集的分类数 c.确定数据集的分类数问题属于"聚类有效性问题",目前已提出了各种各样的有效性标准[2]."划分系数"是 Dunn[3]引入的,用于度量数据的模糊划分的模糊性程度.将划分系数应用于聚类有效性的判决归功于 Bezdek,特别是 Gunderson[4]应用划分系数对星域数据进行了成功的分类,使得划分系数成为模糊聚类中的第一个实用的聚类有效性标准.划分系数具有良好的数学性质,但正如 Bezdek 在其专著[1]中指出的,划分系数具有随类数增加而单调递减的趋势,缺乏与数据集的几何结构的直接联系. Trawvaest[5]对划分系数进行了更为深入的分析,指出划分系数作为有效性检测标准,存在不尽人意之处.在文[6]中,通过模糊集理论中的可能性分布对划分系数进行了讨论.本文,将从一个新的角度对划分系数进行修改.实验结果表明,这种修改是令人满意的.

2 划分系数

模糊聚类问题可表述成下面的数学规划问题

问题 FCL $$\min J_m(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^m d_{ij}^2 \tag{1}$$

使得 $$\sum_{j=1}^{c}u_{ij}=1,1\le i\le n \tag{1a}$$

$$u_{ij}\ge 0,1\le i\le n,1\le j\le c \tag{1b}$$

$$\sum_{i=1}^{n}u_{ij}>0,1\le j\le c \tag{1c}$$

这里 $X=\{x_1,x_2,\cdots,x_n\}\subset R^s$ 是数据集,n 是数据集中样本的个数,c 是聚类中心数 $(1<c<n)$,m 是权重因子 $(m>1)$,$d_{ij}=\|x_i-V_j\|$ 是样本点 x_i 和聚类中心 V_j 的欧氏距离,$V_j\subset R^s(1\le j\le c)$. u_{ij} 是第 i 个样本属于第 j 个聚类中心的隶属度,$U=[u_{ij}]$ 是一个 $n\times c$ 矩阵,$V=[V_1,V_2,\cdots,V_c]$ 是一个 $s\times c$ 矩阵. Bezedk[1]给出求解上述数学规划问题的循环迭代法:模糊 c-均值聚类算法.在应用模糊 c-均值聚类算法时,必须事先给定数据的分类数.为了确定数据集的分类数,Bezedk[1]提出如下标准.

定义 1 对于给定的分类数 c 和隶属度矩阵 U,划分系数定义为: $F(U;c)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^2$.

划分系数具有如下良好性质

定理 1 对于 $1<c<n$,有

(1) $\frac{1}{c}\le F(U;c)\le 1$; (2) $F(U;c)=1$ 当且仅当 U 是硬划分;

收稿日期:2000-12-04;修回日期:2001-03-16
基金项目:国家自然科学基金(No. 69972041)

(3) $F(U;c)=1/c$ 当且仅当 $U=[1/c]$.

从上述性质可见,划分系数是一个判定分类模糊性程度的标准. 分类越分明时, $F(U;c)$ 的值越大;分类越模糊时, $F(U;c)$ 的值越接近于 $1/c$. 记 Ω_c 为所有分类矩阵的"最优"有限集,如果存在 (U^*,c^*) 满足 $F(U^*;c^*)=\max_c\{\max_{\Omega_c}F(U;c)\}$,则 (U^*,c^*) 是最佳的有效性聚类, c^* 是最佳的分类数.

3 新标准的提出

对于模糊 c-划分矩阵 $U=[u_{ij}]$,文[8]给出如下定义.

定义 2 对于给定的模糊 c-划分, $u_{ij}\|x_i-V_j\|$ 叫样本 x_i 到第 j 个类的模糊偏差.

定义 3 对于给定的模糊 c-划分, $\sigma_j=\sum_{i=1}^{n}u_{ij}^2\|x_i-V_j\|^2$ 叫第 j 个类的变差.

定义 4 对于给定的模糊 c-划分, $J(U;c)=\sum_{j=1}^{c}\sigma_j=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^2\|x_i-V_j\|^2$ 叫 X 的总变差.

上述定义是独立于模糊聚类算法的. 当 $m=2$ 时,模糊 c-均值聚类算法中的目标函数 $J_2(U;V)$ 就是 X 的总变差 $J(U;c)$. 单独用 $J(U;c)$ 或 $J_m(U;V)$ 来定义聚类有效性函数是不可行的,这一点在文[1]中已经指出. 应用 $J(U;c)$ 来定义聚类有效性标准, Fukugama 和 Sugeno[7]作了初步尝试,但正如文[2]所指出的,这种定义有太多的弊病. 1991 年, Xie 和 Beni[8]把 $J(U;c)$ 看成数据集的"紧致性"的一种表示,应用"紧致性"和"分离性"提出了一个有效性标准. 如果将划分系数 $F(U;c)$ 和 $J(U;c)$ 对比一下,可以看出, $F(U;c)$ 和 $J(U;c)$ 的差异在于后者应用到了数据的几何结构信息,前者仅用到了数据的模糊隶属度信息. 为此,本文的设想是将 $F(U;c)$ 和 $J(U;c)$ 结合在一起来考虑定义有效性标准. 换句话说,希望将数据的模糊划分信息和数据的几何结构信息溶合在一起来定义有效性函数.

一个应当注意的事实是,应用模糊 c-均值聚类算法是有条件的. 一般地,模糊 c-均值聚类算法适用于每类样本数相差不大且每类离差相差不大的团状数据. 有关这一方面的理论分析在文[9]中有所涉及. 文[9]指出,当每类离差相同时,基于模糊最大分类相关原则的模糊聚类算法就是常见的模糊 c-均值聚类算法. 因此,在定义有效性标准时,应将模糊 c-均值聚类算法适用条件的信息考虑进去. 但迄今为止,所有的有效性标准,并未充分考虑这方面的约束. 为此,本文定义中加入了这方面的信息. 实验结果表明,这样做是有效的.

对于给定的数据集 $X=\{x_1,x_2,\cdots,x_n\}$,记 X 的中心 $V_0=\sum_{i=1}^{n}x_i/n$,用 J_0 表示所有样本到 V_0 的距离之和,即 $J_0=\sum_{i=1}^{n}\|x_i-V_0\|^2$. $F(U;c)$ 的最大值代表最佳分类信息, $J(U;c)$ 的最小值尽管不能完全代表最佳分类信息,但也对最佳分类有所体现. 由于 $F(U;c)$ 和 $J(U;c)$ 相差很大,因此给 $J(U;c)$ 除以 J_0 来抑制 $J(U;c)$.

对于未标识的数据集,如果事先对数据结构一无所知,从理论上说若分成 C 个类,假设每类的样本数相差不大常常是可行的. 用 $\sum_{i=1}^{n}u_{ij}$ 表示隶属于第 j 个类的样本数,给出的第一个有效性标准为:

$$F_1(U;c)=\frac{\min_{j=1}^{c}(\sum_{i=1}^{n}u_{ij})}{\max_{j=1}^{c}(\sum_{i=1}^{n}u_{ij})}\left[F(U;c)+1-\frac{J(U;c)}{J_0}\right]$$

其中 $\frac{\min_{j=1}^{c}(\sum_{i=1}^{n}u_{ij})}{\max_{j=1}^{c}(\sum_{i=1}^{n}u_{ij})}$ 可看作是对 $F(U;c)+1-\frac{J(U;c)}{J_0}$ 进行了"加权"处理. 作为 $F_1(U;c)$ 的一种对偶表示,给出第二个有效性标准:

$$F_2(U;c)=\frac{\min_{j=1}^{c}(\sum_{i=1}^{n}u_{ij})}{\max_{j=1}^{c}(\sum_{i=1}^{n}u_{ij})}\left[1-F(U;c)+\frac{J(U;c)}{J_0}\right]$$

和 $F(U;c)$ 一样,满足 $F_1(U^*;c^*)=\max_c\{\max_{\Omega_c}F_1(U;c)\}$ 的 $(U^*;c^*)$ 或 $F_2(U^*;c^*)=\min_c\{\min_{\Omega_c}F_2(U;c)\}$ 的 $(U^*;c^*)$ 是最佳的有效性聚类, c^* 是最佳的分类数.

4 实验结果

在文[2]中, Pal 和 Bezdek 讨论了权重因子 m 对模糊 c-均值聚类有效性的影响,指出 m 的取值范围可选为 $[1.5,2.5]$. 限于篇幅本节只给出对 1 个人造数据和著名的 IRIS 数据的测试结果. 为了方便,我们限制最大分类数 $c_{max}=10$ 且选择权重因子的三个典型值 $m=1.5$, $m=2.0$ 和 $m=2.5$.

立方体数据:由均值分别为 $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ 和 $(1,1,1)$,各维方差均为 0.5 的正态分布生成的空间数据. 每类 100 个样本,共计 400 个样本,图 1 给出该数据的分布图. 从表 1 可见, $F_1(U;c)$ 和 $F_2(U;c)$ 均得到正确的分类数.

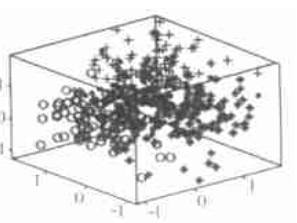

图 1 立方体数据的分布图

表 1(a) 立方体数据的 $F_1(U;C)$ 值

C	$F_1(U;C)$		
	$m=1.5$	$m=2.0$	$m=2.5$
2	1.152993	1.076358	1.023000
3	1.135512	1.121638	1.053268
4	1.291692	1.158688	1.056196
5	1.179919	1.084703	1.006726
6	1.024113	1.083308	0.999051
7	0.986384	1.032254	0.984579
8	0.983898	1.046326	0.974065
9	0.968812	0.976083	0.951899
10	0.933325	0.923435	0.951798

表 1(b) 立方体数据的 $F_2(U;C)$ 值

C	$F_2(U;C)$		
	$m=1.5$	$m=2.0$	$m=2.5$
2	0.828840	0.920776	0.976609
3	0.764117	0.859187	0.943372
4	0.598439	0.806755	0.934392
5	0.669587	0.874620	0.983519
6	0.773608	0.882019	0.993092
7	0.795057	0.928738	1.008534
8	0.806947	0.920952	1.020768
9	0.801473	0.987338	1.044430
10	0.831453	1.044356	1.045220

IRIS 数据:该数据由 4 维空间的 150 个样本组成. 每一个样本的四个分量分别表示 IRIS 的 petal length, petal width, sepal length, sepal width. IRIS 数据共有三个种类 setosa, versicolor, virginica,每一个种类有 50 个样本. 第一个种类与其它二类完全分离,第二个种类与第三个种类之间有交叉. 该数据常被用来

检验聚类算法和聚类有效性函数的性能.从表 2 可见, $F_1(U;c)$ 和 $F_2(U;c)$ 均得到正确的分类数.

表 2(a)　IRIS 数据的 $F_1(U;C)$ 值

C	$F_1(U;C)$		
	$m=1.5$	$m=2.0$	$m=2.5$
2	0.987234	1.091523	1.159510
3	1.238752	1.330863	1.314881
4	0.897578	0.897012	0.935421
5	0.959150	1.033372	1.088798
6	0.585749	0.686708	0.784882
7	0.761896	0.679003	0.698931
8	0.803652	0.759790	0.747958
9	0.642296	0.748209	0.733225
10	0.614769	0.596651	0.625613

表 2(b)　IRIS 数据的 $F_2(U;C)$ 值

C	$F_2(U;C)$		
	$m=1.5$	$m=2.0$	$m=2.5$
2	0.413710	0.457456	0.548304
3	0.265415	0.392102	0.529262
4	0.378866	0.656594	0.821406
5	0.407629	0.651380	0.765746
6	0.652085	1.011438	1.092701
7	0.538791	1.076820	1.263044
8	0.537468	1.008241	1.211516
9	0.669986	1.037999	1.248869
10	0.760171	1.324377	1.473211

5　结论

本文通过将划分系数和总变差结合在一起,并进行加权处理的方式定义了二个有效性标准.实验结果表明,这样做是可行的.文[2]中报道了划分系数和 *X-B* 指标对 IRIS 数据的分类结果.和本结果相比,本方法具有明显的优势.应该指出的是,在定义有效性函数时溶入"样本数"信息是行之有效的.这一点,在其它工作中也有所体现.

参考文献:

[1] J C Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms [M]. New york, 1981:309 - 321.

[2] N R Pal, J C Bezdek. On cluster validity for the fuzzy C-means model [J]. IEEE Trans. Fuzzy System, 1995, 3(3):370 - 379.

[3] J C Dunn. Some recent investigations of a new fuzzy partition algorithm and its application to pattern classification problems [A], J. Cybernet [C], 1974, 4:1 - 15.

[4] R Gunderson. Applications of Fuzzy ISODATA algorithms to startracker printing systems [A]. in Proc. 7th Triennual world IFAC Congr. [C], 1978:1319 - 1323.

[5] E Trawvaest. On the meaning of Dunn's partitions coefficient for fuzzy clusters [J]. Fuzzy sets and systems, 1988, 25:217 - 242.

[6] 范九伦,裴继红,谢维信. 基于可能性分布的有效性函数 [J]. 电子学报, 1998, 4:103 - 105.

[7] Y Fukuyama, M Sugeno. A new method of choosing the numbers of clusters for the Fuzzy C-means method [A]. in Proc. 5th Fuzzy syst. Symp. [C], in Japanese, 1989:247 - 250.

[8] X L Xie, G A Beni. Validity measure for Fuzzy clustering [J]. IEEE Trans. Pattern Anal. Machine Intell. 1991, 3(8):841 - 846.

[9] M S Yang. On a class of fuzzy classification maximum likelihood procedures [J]. Fuzzy Sets and Systems. 1993, 57:365 - 375.

[10] M R Rezaee, et al. A new cluster validity index for the fuzzy C-mean [J]. Pattern recognition Letters, 1998, 19:237 - 246.

作者简介:

范九伦　男. 1964 年生于陕西省西安市. 博士后,教授. 1985 年和 1988 年在陕西师范大学数学系分获学士、硕士学位. 1998 年在西安电子科技大学获得博士学位. 1998 年 ——2000 年在西北工业大学博士后流动站工作. 主要研究兴趣包括模糊集理论、不确定性推理、模糊模式识别和图像处理. 出版专著一部,在国内外杂志已发表四十余篇论文.

吴成茂　男. 1968 年生于四川省仪陇县. 学士,工程师, 1990 年在西安工业学院计算机系获学士学位. 主要从事模糊信息处理、智能管理决策等方面的研究. 发表十余篇论文.