一种新的聚类有效性函数

彭 勇[1,2],吴友情[3]
PENG Yong[1,2],WU You-qing[3]

1.中国科学院 研究生院,北京 100039
2.中国科学院 沈阳计算技术研究所,沈阳 110171
3.安徽大学 计算机科学与技术学院,合肥 230039
1.Graduate University of Chinese Academy of Sciences,Beijing 100039,China
2.Shenyang Institute of Computing Technology,Chinese Academy of Sciences,Shenyang 110171,China
3.School of Computer Science and Technology,Anhui University,Hefei 230039,China
E-mail:pengyong@mail.ustc.edu.cn

PENG Yong,WU You-qing.New cluster validity function for determining cluster number.Computer Engineering and Applications,2010,46(6):124-126.

Abstract: Cluster validity index is used to evaluate the validity of clustering.The clustering result will tend to be more logical on the condition that the initial clustering number is accurately ascertained.According to the basic theory of fuzzy indetermination and the properties of clustering,a new cluster validity function is proposed to identify the optimal cluster number based on the newly introduced index $D_t(U;c)$ that can measure the clustering compactness.Both the geometry structure of dataset and the membership degree are taken into account in the validity function,which based on the properties of clustering compactness and separation.The experimental results indicate that the new validity function can find out the only cluster number if the dataset has the obvious cluster trend and it is also non-sensitive to the weighting coefficient m.

Key words: fuzzy clustering;clustering validity;fuzzy c-means;clustering compactness;clustering separation

摘 要:聚类有效性函数是用于评价聚类结果优劣的指标,准确地给出初始聚类类别数将使得聚类结果趋于合理化。根据模糊不确定性理论及聚类问题的基本特性,引入了新的紧密度度量指标 $D_t(U;c)$,在此基础上提出了一个旨在寻求最优聚类类别数的有效性函数。该函数基于数据集的紧密度与分离度特征,综合考虑了数据成员的隶属度及数据集的几何结构。实验结果表明该有效性函数能够发现最优的聚类类别数,对于分类结构较为明确的数据集表现出良好的性能,并且对于权重系数具有良好的鲁棒性。

关键词:模糊聚类;聚类有效性;模糊 c 均值;聚类紧密度;聚类分离度
DOI:10.3778/j.issn.1002-8331.2010.06.035 **文章编号:**1002-8331(2010)06-0124-03 **文献标识码:**A **中图分类号:**TP391

1 前言

作为多元统计分析的一种,聚类分析已经广泛地应用于模式识别、数据挖掘等领域。一般地,在明确给定的数据集具有聚类趋势的前提下,可以根据相应的算法进行聚类。但是聚类的结果是否合理,则需要进行有效性分析。通常情况下,聚类有效性分析可以转化为聚类类别数 c 和模糊权重系数 m 的自动确定[1]。

聚类有效性函数的定义方法一般分为 3 类[2]:基于数据集模糊划分的方法;基于数据几何结构的方法和基于数据统计信息的方法。在模糊聚类领域,基于目标函数的模糊 c-均值(Fuzzy C-Means,FCM)类型算法理论最为完善,所以适合于模糊 c-均值聚类算法的聚类有效性函数应用也相当广泛[1,3]。Bezdek 于 1974 年提出了划分系数[4]和划分熵[5],虽然这两个函数具有明确的数学意义和良好的数学性质,但是由于它们是基于样本数据集隶属度定义的,缺少与数据集几何结构特征的联系,所以存在局限性。针对此缺点,许多学者同时考虑隶属度与数据集几何结构信息,提出了一些有效性函数,代表性的有 Xie-Beni 指标[6],kwon 提出的 $V_k(U,V;c)$[7]以及 Fukuyama 和 Sugeno 提出的 $FS(U,V;c)$[8],其理论基础就是类内紧密、类间分离。而基于统计信息的方法就是利用统计的方法,按照数据的分布情况进行判断,如 PFS[9]以及 Sun、Wang 和 Jiang[10]提出的改进指标。

在研究模糊不确定性理论和聚类问题基础上,提出了基于紧密度与分离度的聚类有效性函数,并通过相应的数据集进行了实验。

2 模糊聚类算法

聚类的主要方法有谱系聚类法,等价关系聚类法,图论聚类法以及基于目标函数的聚类法等。下面简单介绍一种基于目标函数的聚类算法:模糊 c-均值聚类算法,后文提出的聚类有

作者简介:彭勇(1985-),男,硕士,CCF 学生会员,主研领域为聚类分析、数据挖掘;吴友情(1984-),女,硕士,主研领域为聚类分析、信息隐藏技术。
收稿日期:2008-09-02 **修回日期:**2008-11-10

效性函数也将基于此算法。

2.1　模糊 c-均值聚类算法

FCM 聚类算法是一种交替优化的聚类算法，通过迭代动态的寻找最佳划分矩阵。简单地说,FCM 算法可以看成在约束条件(1)下的一个非线性规划问题,对于事先给定的聚类类别数 $c(c\in[1,n])$,求取数据集 X 的模糊 c-划分 $\tilde{F}=\{\tilde{F}_1,\tilde{F}_2,\cdots,\tilde{F}_c\}$,使得目标函数 $J_m(\boldsymbol{U},\boldsymbol{V})$ 最小,即 $J_m(\boldsymbol{U},\boldsymbol{V})=\min\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}(d_{ij}^2)$。

$$\text{s.t.}\begin{cases}\sum_{i=1}^{c}u_{ij}=1,1\leqslant j\leqslant n\\ u_{ij}\in[0,1],1\leqslant j\leqslant n,1\leqslant i\leqslant c\\ 0<\sum_{j=1}^{n}<u_{ij}<n,1\leqslant i\leqslant c\end{cases}\tag{1}$$

其中 $\boldsymbol{U}=[u_{ij}]_{c\times n}$ 是模糊划分矩阵。聚类中心向量集为 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_c\}$,$v_i=\{v_{i1},v_{i2},\cdots,v_{ik}\}$ 为第 i 个聚类中心,d_{ij} 表示两聚类中心之间的距离。m 为权重系数,m 越大,对应的模糊度越大。对应不同的 m 值,显然有不同的最佳 c 划分。

FCM 算法的过程描述可以参考文献[2]。

2.2　基于 FCM 算法的聚类有效性分析

基于 FCM 聚类算法的有效性分析过程如下:

(1)确定可能的聚类类别数 $[c_{\min},c_{\max}]$,一般地,可以取 $c_{\min}=2$,$c_{\max}=\sqrt{n}$;

(2)对于 $c\in[c_{\min},c_{\max}]$:初始化聚类中心,修正划分矩阵和聚类中心,若收敛则计算有效性指标 Val,否则迭代直至收敛或终止条件满足;

(3)比较不同 c 值对应的有效性指标,选择有效性指标的最优值对应的聚类类别数 c 去初始化聚类原型。

3　新的聚类有效性函数

从模糊不确定性基本原理出发,结合聚类基本特性,引入了新的度量紧密度的指标 $D_i(U;c)$,并且使用数据点到其最小隶属度的聚类中心距离作为分离度指标 s,建立了新的聚类有效性函数。

3.1　紧密度指标

一般地,在模糊理论中,模糊测度 $d(\mu)$ 具有以下性质:

$P1$:$d(\mu)=0$ 当且仅当 $\mu=0$ 或 $\mu=1$;

$P2$:当且仅当 $\mu=0.5$ 时,$d(\mu)$ 取 $d(\mu)_{\max}$;

$P3$:当 $\mu\geqslant0.5$ 时,若 $\mu<\mu^*$ 则 $d(\mu)\geqslant d(\mu^*)$,

　当 $\mu<0.5$ 时,若 $\mu>\mu^*$ 则 $d(\mu)\geqslant d(\mu^*)$;

$P4$:$d(\mu_{\tilde{A}})=d(\mu_{\tilde{A}^c})$。

对于有限集的情况，能够满足上述四条基本性质的 $d(\mu)$ 的数学形式是[2]:

$$d(\mu)=F\left[\sum_{i=1}^{N}c_i f_i(\mu_{\tilde{A}}(x_i))\right]$$

其中 F 是非负递增函数;c_i 是正实数,$i=1,2,\cdots,N$;对于所有的 i,f_i 是实函数且 $f_i(0)=f_i(1)=0$,$f_i(0.5)$ 是 f_i 唯一的最大值,f_i 在 $[0,0.5]$ 区间内单调递增,在 $[0.5,1]$ 内单调减,对 $\forall\mu\in[0,1]$ 有 $f_i(\mu)=f_i(1-\mu)$。

采用 Minkowski 距离来度量模糊测度,得到 $d(\mu)=\left\{\sum_{i=1}^{N}\left|\mu_{\tilde{A}}(x_i)-\mu_{\tilde{A}0.5}(1/c)\right|^q\right\}^{1/q}$,此时 $d(\mu)$ 表示 $\tilde{A}$ 到它的最贴近普通集合 A 的 Minkowski 距离。

结合 FCM 模糊聚类基本特性，对于给定的聚类中心 c 和模糊划分矩阵 $\boldsymbol{U}$,可以改造 $d(\mu)$ 为 $D_i(\boldsymbol{U};c)=\left\{\sum_{j=1}^{N}\left[\mu_{ij}-\mu_{ij(1/c)}\right]^2\right\}^{1/2}$，式中 $[\mu_{ij(1/c)}]$ 表示用 $1/c$ 去截 $[\mu_{ij}]$ 得到的 0-1 阵,将 $D_i(\boldsymbol{U};c)$ 归一化到 $[0,1]$ 区间,即

$$D_i(\boldsymbol{U};c)=\frac{c}{c-1}\left\{\frac{1}{N}\sum_{j=1}^{N}\left[\mu_{ij}-\mu_{ij(1/c)}\right]^2\right\}^{1/2}$$

此模糊不确定性测度具有以下性质:

$P1$:$0\leqslant D_i(\boldsymbol{U};c)\leqslant1$;

$P2$:$D_i(\boldsymbol{U};c)=0$ 当且仅当 $\boldsymbol{U}$ 是硬划分;

$P3$:$D_i(\boldsymbol{U};c)=1$ 当且仅当 $\boldsymbol{U}=[1/c]$。

相应地,整个聚类模糊集的模糊测度可定义为 $D(\boldsymbol{U};c)=\sum_{i=1}^{c}D_i(\boldsymbol{U};c)$。

明显地,划分结果越分明,$D(\boldsymbol{U};c)$ 值越小,当 $D(\boldsymbol{U};c)=0$ 时为硬划分;划分越模糊,$D(\boldsymbol{U};c)$ 值越大,当 $D(\boldsymbol{U};c)=c$ 时,各数据对象均匀属于相应的类,此时 $\boldsymbol{U}=[1/c]$,即聚类的“最模糊状态”。当采用距离度量法时,隶属度的大小取决于样本点到相应聚类中心的距离,若每个数据样本点都有较明确地归属类别(即每各样本数据点相对于某一聚类中心的隶属度明显大于对其他的聚类中心的隶属度),那么这种模糊的不确定性就很小,整个隶属度的分布就远离“最模糊”状态,也即聚类模糊度较小,整个聚类显示较好的紧密度。

3.2　分离度指标

1974 年,Dunn 提出了指标[12],即

$$D=\min_{i=1,2,\cdots,c}\left\{\min_{j=i+1,i+2,\cdots,c}\left\{\frac{d(c_i,c_j)}{\min\limits_{k=1,2,\cdots,c}diam(c_k)}\right\}\right\}$$

该指标包含了紧密度和分离度,但是由于计算量大和对噪音数据点的敏感而限制了其应用。

基于 Dunn 的思想,采用不同的方式来定义紧密度和分离度就得到了不同的聚类有效性函数，其中有著名的 Xie-Beni 指标[6],其定义的分离度就是聚类中心间距离的最小值,即 $s=\min\limits_{1\leqslant p,q\leqslant c\ p\neq q}\|V_p-V_q\|^2$。但是此分离度指标仅考虑了各聚类中心的几何位置关系,对于每一类的数据分布情况并不能进行很好的描述。

使用的分离度定义[13]为 $s=\|x_j-v_i\|$,其中 $i^*=\arg\min\limits_{k=1,2,\cdots,c}u_{ij}$,即数据 x_j 到其隶属度最小的聚类中心 v_i 的欧式距离作为分离度的度量方法。

采用此分离度定义方案的优点是,可以有效地将某一样本点对类的隶属程度和样本的分布结构联系起来,对类之间有交叠和多孤立点的情况都能做出正确的评价。当采用距离度量法时,一个样本对某一类的隶属度越小就表示它到这此聚类中心的距离越大,相应的类之间的分离度就越明显,效果越好。

3.3　聚类有效性函数

基于上述关于紧密度和分离度的分析,定义聚类有效性函数为 $F_{DS}=\dfrac{\max\limits_{1\leqslant k\leqslant c}D_k(\boldsymbol{U};c)}{\min\limits_{1\leqslant j\leqslant n}\|x_j-v_i\|}$,显然较小的 F_{DS} 对应较好的聚类结果。

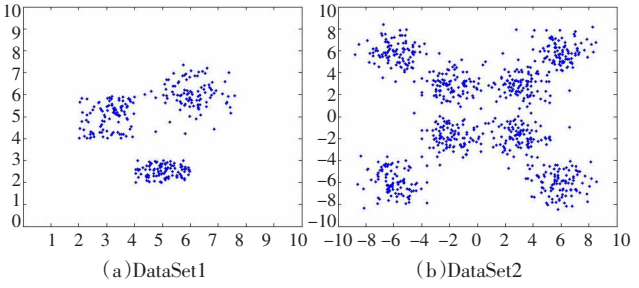
明显地,越小的聚类模糊度对应越好的聚类效果,越大的分离度对应越好的聚类效果,所以分子 $\max\limits_{1\leqslant i\leqslant c}D_i(\boldsymbol{U};c)$ 和分母

$\min_{1\leqslant j\leqslant n}\|x_j-v_i\|$ 都是对应最坏聚类情况下的指标,即聚类结果最不合理状态。这里,并未采用平均紧密度和平均分离度作为分子、分母是因为平均指标会使得聚类有效性函数的敏感性降低,弱化其评价效果。

4 实验及结果分析

4.1 实验数据集

人工数据集 Dataset1 和 Dataset2 分别如图 1(a)和(b)所示。Dataset1 由 300 个点组成,其中两类为矩形域均匀分布,另一类为正态分布,分为 3 类,每类 100 个点;Dataset2 由均值分别为(3,3),(3,-2),(-2,-3),(-2,-2),(-6,6),(6,-6),(6,6)和(-6,-6),各维方差均为 1 的正态分布点构成,每类有 100 个点,共 8 类。



(a)DataSet1 (b)DataSet2

图 1 人工生成的 2 个数据集

选自机器学习数据库的真实数据集 IRIS 由 4 维空间的 150 个样本组成,分为 3 类,第一类与其他两类是线性分离的,后两类由一定的重叠,所以聚类的结果类别数为 2 或 3 较为合理。该数据集常被用来检验聚类算法和聚类有效性函数的性能。

4.2 实验参数设置

基于 FCM 算法对上述 3 个数据集进行聚类。算法终止迭代的条件为 $\varepsilon\leqslant 0.000\,1$ 或迭代次数大于 400 次。由于目前对 c 的选取没有明确的指导性意见,所以在实验中进行了灵活选取(对 Dataset1 选 $2\leqslant c\leqslant 8$,Dataset2 选 $2\leqslant c\leqslant 12$,IRIS 数据集选 $2\leqslant c\leqslant 10$)。由于权重系数 m 对聚类结果影响较大,一般认为 m 在[1.5,2.5]之间选取较恰当,对 m=1.5,m=2,m=2.5 的情况均进行了实验。

4.3 实验结果分析

在 MATLAB 7.0 下通过 3 个数据集对提出的聚类有效性函数 F_{DS} 进行测试:

(1)Dataset1 的实验结果数据及分析(表 1,图 2)。

从实验结果数据看出,在不同的m 值条件下,有效性函数 F_{DS} 均在 c=3 时取到最小值,符合实际情况。

表 1 Dataset1 的 F_{DS} 值

c	F_{DS}			c	F_{DS}		
	m=1.5	m=2.0	m=2.5		m=1.5	m=2.0	m=2.5
2	0.177 8	0.286 5	0.291 0	6	0.090 8	0.120 1	0.157 6
3	**0.020 6**	**0.056 6**	**0.090 5**	7	0.084 7	0.136 1	0.151 6
4	0.081 7	0.112 9	0.125 7	8	0.077 9	0.116 4	0.157 5
5	0.092 1	0.123 4	0.130 0				

(2)Dataset2 的实验结果数据及分析(表 2,图 3)。

表 2 Dataset2 的 F_{DS} 值

c	F_{DS}			c	F_{DS}		
	m=1.5	m=2.0	m=2.5		m=1.5	m=2.0	m=2.5
2	0.051 9	0.111 8	0.193 0	8	**0.013 4**	**0.025 8**	**0.038 4**
3	0.038 5	0.075 5	0.089 3	9	0.022 5	0.036 5	0.041 9
4	0.020 5	0.038 2	0.051 6	10	0.018 2	0.034 8	0.046 0
5	0.037 0	0.055 6	0.063 8	11	0.021 2	0.034 1	0.045 8
6	0.035 6	0.048 8	0.059 7	12	0.019 8	0.034 0	0.044 3
7	0.024 5	0.041 2	0.046 0				

对 Dataset2,不同的 m 值条件下,F_{DS} 均在 c=8 时取到最小值,从图 3 可以看出各曲线同时又在 c=4 时出现极值,这和实际情况是相符的。从数据的几何分布可以看出,数据可以每 2 类归为一类,这样沿 45°,135°,225°,315°四个方向聚成 4 类也较为合理。

(3)IRIS 数据集的实验结果数据及分析(表 3,图 4)。

表 3 IRIS 数据的 F_{DS} 值

c	F_{DS}			c	F_{DS}		
	m=1.5	m=2.0	m=2.5		m=1.5	m=2.0	m=2.5
2	**0.062 0**	**0.100 4**	**0.141 9**	7	0.098 5	0.139 0	0.180 6
3	0.071 2*	0.121 2*	0.158 9*	8	0.096 2	0.152 5	0.178 7
4	0.076 7	0.129 5	0.167 4	9	0.089 2	0.150 4	0.179 8
5	0.084 8	0.136 0	0.185 7	10	0.086 8	0.146 4	0.174 3
6	0.084 5	0.146 3	0.171 7				

对 IRIS 数据集,对不同 m 值 F_{DS} 均在 c=2 时取得最小值,又同时在 c=3 时取得次小值,这与实际分类情况(第二类和第三类有重叠)是一致的。针对 IRIS 数据集,还使用多个有效性指标与 F_{DS} 进行对比测试(表 4),显示出 F_{DS} 的合理性。

通过实验可以看出,在不同 m 值情况下,有效性函数 F_{DS} 均能在合理的聚类类别数下取得较为优化的值,说明 F_{DS} 对 m 具有良好的鲁棒性。

5 结束语

提出了一个新的聚类有效性函数,在聚类基本特性的基础

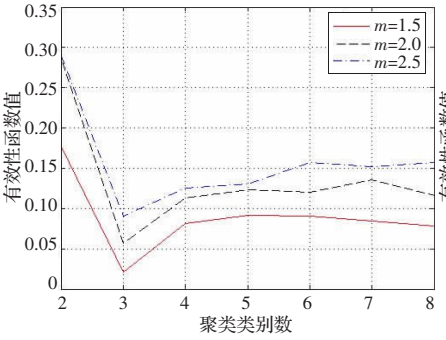


图 2 Dataset1 的 F_{DS} 值随聚类类别数变化曲线

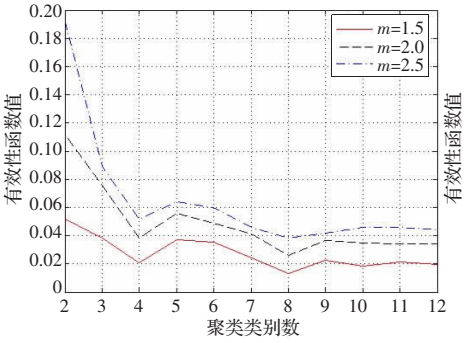


图 3 Dataset2 的 F_{DS} 值随聚类类别数变化曲线

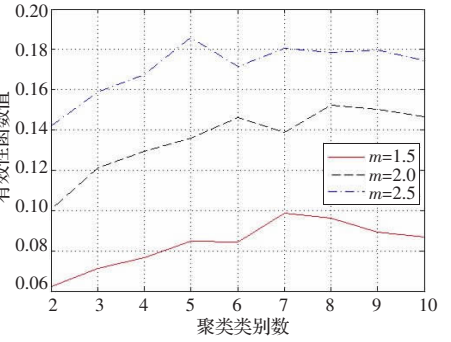


图 4 IRIS 的 F_{DS} 值随聚类类别数变化曲线

(下转 132 页)