模糊聚类中判别聚类有效性的新指标*)

洪志令[1] 姜青山[2] 董槐林[2] Wang Sheng-Rui[3]
(厦门大学计算机科学系[1] 厦门大学软件学院[2] 厦门361005)
(Department of Computer Science, University of Sherbooke. Quebec, Canada)[3]

摘 要 本文提出了一个在模糊聚类中判别聚类有效性的新指标。该指标可有效地对类间有交叠或有多孤立点的情况做出准确的判定。文中基于模糊C-均值聚类算法(FCM),应用多组的测试数据对其进行了性能分析,并与当前较广泛使用且较具代表性的某些相关指标进行了深入的比较。实验结果表明,该指标函数的判定性能是优越的,它可以自动地确定聚类的最佳个数。

关键词 模糊聚类分析,有效性指标,FCM 算法

A New Cluster Validity Index for Fuzzy Clustering

HONG Zhi-Ling[1] JIANG Qing-Shan[2] DONG Huai-Ling[2] WANG Sheng-Rui[3]
(Deportment of Computer Science[1], School of Software[2], Xiamen University, Xiamen 361005)
(Department of Computer Science, University of Sherbooke, Quebec, Canada)[3]

Abstract In this paper, we propose a new validity index for determining the number of clusters. It is based on a novel way of combining cohesion and discrepancy. Extensive tests of the index in a conventional model selection process (FCM algorithm) have been performed using generated data sets and public domain data sets, and comparison with several existing and important indices has been made. The results obtained show clearly the efficiency of the new index under the condition of overlapping clusters.

Keywords Fuzzy clustering, Validity index, FCM algorithm

1 引言

聚类分析是研究和处理给定对象分类的数学方法,它按一定的距离和相似性测度将数据划分为一系列有意义的子集(或类),每个子集中的数据尽量地"相似"或"接近",而子集与子集间的数据尽可能地有"很大差异"。目前较广泛使用的聚类算法有 K 均值(K-Mean),K-Meriod, 模糊 C-均值聚类算法(FCM)等。其中 FCM 引入了模糊集的概念,可对孤立点、成员关系等更好地进行处理,因此更受关注,成为当前研究热点之一。使用该算法经常会遇到一系列的问题,如聚类初始中心点的选择[1],模糊因子 m 的确定[2]等,这些问题大部分都已经得到妥善解决。

目前,人们关注的焦点是"聚类的有效性问题",即如何聚类才是最合理的,一般采用有效性指标进行评价。迄今为止,已经提出了若干检验聚类有效性的指标,其中,较具代表性的是 Xie 和 Beni[3]于1991年提出的基于"紧凑度"与"分离度"比值的有效性定义。根据该思想,相继提出了一系列的改革与创新,如 Bensaid[4]于1996年针对"紧凑度"定义中对类大小的敏感性问题,提出了改进定义;Kwon[5]于1998年针对 c 趋近于 n 时,指标的失效性问题,提出了惩罚函数定义;Zahid[6]于1999年在考虑数据集几何结构的基础上,加入了对数据集模糊划分的考虑等。

由于这些改进并没有从根本上摆脱整个模型定义的思路,"紧凑度"和"分离度"在数值上的悬殊差距一直没有得到适当的处理,因而先天不足。我们采用多种不同类型数据,进行大量实验测试这些指标,实验结果表明,这些仅在局部上做了优化的指标的判定能力并没有得到有效提高,有时甚至与实际结果偏差更大。更重要的是,该类指标无法准确处理类间有交叠或有多孤立点的情况。

1998年,Rezaee[11]提出了采用线性组合,并通过比例因子对"紧凑度"和"分离度"进行缩放的思想,从而一定程度上弥补了度量差别上的缺陷。虽然该指标整体性能上已经有了很大提高,但结构上非常复杂,并且经常会给出与事实相悖的结果。针对线性组合法的不稳定性,Sun, Wang 和 Jiang[7,8]于2001年,2003年两次对其进行了改进,进一步完善了该指标,极大地提高了算法的稳定性,并且可以准确处理类间有交叠或有多孤立点的情况。

虽然经 Sun, Wang 和 Jiang 改进的指标在判定时具有优良的稳定性和准确性,但其复杂度仍然比较高,影响了计算效率。为此,本文提出了一个新的判别聚类有效性的指标函数。该指标也采用线性组合方法,但两个组合项却是通过对 Xie 、Bensaid 、Zahid 等的指标中"紧凑度"与"分离度"的精确定义改进而来,具有一定的直观意义。大量的实验充分证明,该指标不仅具有与 Sun, Wang 和 Jiang 相近的稳定程度,在复杂性上也有了一定程度的改善,它对于有交叠或有多孤立点的情况都能给出准确的判断,并自动给出正确的最佳聚类数。

*)本文的研究受姜青山(教授)校科研启动基金(0630-X01117)资助。洪志令 硕士研究生,主要研究方向为数据挖掘、聚类分析,图像处理、数据库系统、地理信息系统等。姜青山 教授,研究领域包括数据挖掘、图像处理、数学模型、数据库系统、聚类分析、地球信息系统、数字通信、模糊集理论与应用、统计计算等。董槐林 副教授,从事应用软件系统的教学与开发工作。Wang Shengrui 教授,研究领域包括模式识别、人工智能、图像处理和理解、知识采集、信息系统、神经网络、优化等。

本文在结构上将做这样的安排：首先简要地介绍基本的FCM算法及利用FCM算法求解最佳聚类数的过程，然后详细描述我们的新指标，并与当前较广泛使用且较具代表性的某些相关指标进行深入的比较，最后通过实验结果报告，进一步验证指标的优越性能。

2 基本算法

基本FCM聚类算法可表示成下面的数学规划问题：

$$Minimize\ J(X,U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m\|x_j-v_i\|^2 \tag{1}$$

$$u_{ij}=\begin{cases}\dfrac{1}{\sum_{r=1}^{c}(\|x_j-v_i\|/\|x_j-v_r\|)^{2/m-1}} & \text{如果}\|x_j-v_r\|>0(\forall\ 1\leqslant r\leqslant c)\\ 1 & \text{如果}\|x_i-v_i\|=0\\ 0 & \text{如果存在}\ r,r\neq i,\|x_j-v_r\|=0\end{cases} \tag{3}$$

$$v_i=\frac{\sum_{j=1}^{n}(u_{ij})^m\cdot x_j}{\sum_{j=1}^{n}(u_{ij})^m},1\leqslant i\leqslant c \tag{4}$$

2.1 基本FCM聚类算法归纳如下

1. 输入聚类数目 c，模糊因子 m，确定距离函数$\|\square\|$，给定迭代终止条件 δ；

2. 初始化聚类中心 $v_i^0(i=1,2\cdots c)$；

3. 使用式(3)计算；

$u_{ij}(i=1,2,\cdots c;j=1,2\cdots n)$；

4. 使用式(4)计算 $v_i^1(i=1,2\cdots c)$；

5. 如果$\max_{1\leqslant i\leqslant c}(\|c_i^0-c_i^1\|/\|c_i^1\|)\leqslant\delta$，则迭代终止，转步骤6，否则，令 $v_i^0=v_i^1(i=1,2,\cdots c)$，转向步骤3；

6. 输出聚类结果(V,U)。

2.2 利用FCM算法求解最佳聚类数的过程

1. 选择聚类数的范围$[c_{min},c_{max}]$

2. 对于 $c=c_{min}$到 c_{max}

2.1 初始化聚类中心(V)；

2.2 应用基本FCM算法更新 U 和 V；

2.3 判断收敛性，如果没有，则转2.2；

2.4 通过有效性指标函数计算指标值 $V_d(c)$；

3. 比较各有效性指标值，最大(或最小)的指标值 $V_d(c_{i_0})$所对应的 c_{i_0}即所求的最佳聚类数。

整个算法的流程如图1所示。最终聚类结果可通过模糊划分矩阵 U 中的隶属度来确定，即若 $u_{i_0j}=\max_{1\leqslant i\leqslant c}(u_{ij})$，则将 x_j 归入第 i_0类；也可通过 x_j 到各聚类中心的距离来确定，即$\|x_j-c_{i_0}\|=\min_{1\leqslant i\leqslant c}\|x_j-c_i\|$，则将 x_j 归入第 i_0类，其 c_{i_0}是第 i_0类的聚类中心。

3 新有效性指标

有效性指标 $V_d(c)$衡量了聚类算法结果的好坏，而指标的好坏在于对两种冲突标准的协调能力。标准一要求类内部尽可能地紧凑，即要求数值上越小越好；标准二要求类与类之间的距离尽可能地远离，即要求数值上越大越好。而好的指标正是在类内紧凑度与类间分离度之间找一个平衡点，从而获得最好的聚类效果。本文将首先简要介绍当前较广泛使用且较具代表性的某些指标，然后引入我们的新有效性指标函数，并做具体介绍，最后通过实验，对这些指标进行综合比较，并给出结论。

其中：$X=\{x_1,x_2\cdots x_n\}\subset R^s$ 是欧式空间的 s 维数据集，n 是给定数据集中的数据个数；$U=(u_{ij})_{c\times n}$是模糊划分矩阵，它由样本 x_j 在第 i 类中的隶属度组成，且

$$\sum_{i=1}^{c}u_{ij}=1 \tag{2}$$

$V=\{v_1,v_2\cdots v_c\}\subset R^s$ 是聚类中心集合，c 是聚类中心数；模糊因子 m 是用来决定聚类结果模糊度的权重指数。

$d(x,y)=\|x-y\|,x,y\in R^s$ 是一个距离函数，如欧几里得距离。对(1)式的优化问题结合(2)式的约束条件，应用 *Lagrange* 乘数法求解可得 U,V 的求解公式：

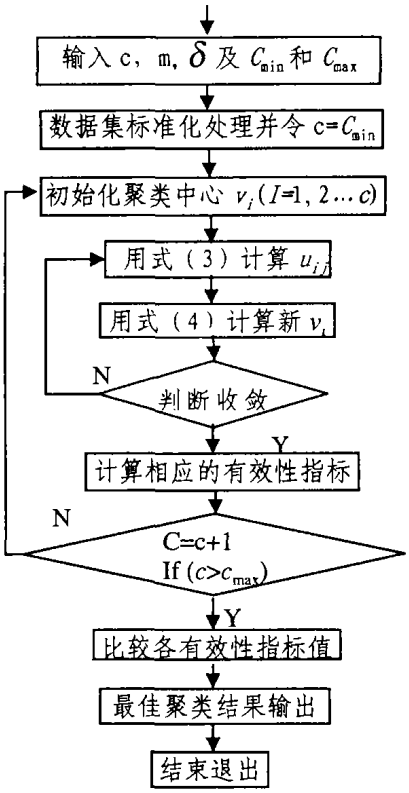

图1 求解最佳聚类数算法流程

3.1 模糊聚类中的有效性指标

3.1.1 Xie-Beni有效性 V_{xie}(1991)[3]

$$V_{xie}(U,V,c)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\|v_i-x_j\|^2}{n\times\min_{i\neq j}\|v_i-v_j\|^2}$$

$V_{xie}(U,V,c)$是类内紧凑度与类间分离度的比例。其中，函数 $J=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\|v_i-x_j\|^2$用来衡量类内的紧凑度，值越小越紧凑。函数 $J_{min}=\min_{i\neq j}\|v_i-v_j\|^2$用来衡量类间的分离度，值越大，分离得越好。$X_{xie}(U,V,c)$就是在类内紧凑度与类间分离度之间找一个平衡点，使其达到最小，从而获得最好的聚类效果。

3.1.2 Amine M. Bensaid有效性 V_{bsaid}(1996)[4]

$$V_{bsaid}(U,V,c)=\sum_{i=1}^{c}\frac{\sum_{j=1}^{n}u_{ij}^m\|x_j-v_i\|^2}{n_i\sum_{i=1}^{c}\|v_i-v_i\|^2}$$

其中 $n_i=\sum_{j=1}^{n}u_{ij}$；相对于 Xie-Beni 的有效性指标，文[4]中提到该指标的优点：(1)、把类间分离度的衡量$\min_{i\neq j}\|v_i-v_j\|^2$替换为$\sum_{i=1}^{c}\|v_i-v_i\|^2$，可对聚类数目相同，分法不同的情况更好地进行比较，因为可能出现不同分法中类间距离最小值相同，而其它类间距相差较大，最终结果却是相同的情况。(2)、把类内紧凑度的衡量由整体总和上的平均替换为各个类中紧凑度的平均和，可使得对各个类的大小不再敏感。实际上，这样的考虑并不全面，我们的新指标将给出新的建议。易知，最小的 $V_{bsiand}(U,V,c)$对应于最好的聚类结果。

3.1.3 S. H. Kwon 有效性 V_{kwon}(1998)[5]

$$V_{kwon}(U,V,c)=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|v_i-x_j\|^2+\frac{1}{c}\sum_{i=1}^{c}\|v_i-\bar{v}\|^2}{\min_{i\neq j}\|v_i-v_j\|^2}$$

其中 $\bar{v}=\frac{1}{n}\sum_{j=1}^{n}X_j$；该指标是对 Xie-Beni 的又一改进。主要体现在引入惩罚函数$\frac{1}{c}\sum_{i=1}^{c}\|v_i-\bar{v}\|^2$，即分子中的第二项。在 c 变得很大，甚至接近 n 的情况下，Xie-Beni 指标将趋近于0，渐渐失去其判定的能力，而经过这样的处理后，它可以有效地缩减指标值下降的趋势，从而使该指标仍保持有效。详细论述请参见文[5]。最小的对应于最好的聚类结果。

3.1.4 N. Zahid 有效性 V_{zahid}(1999)[6]

$$V_{zahid}=SC_1(U,V;X)-SC_2(U)$$

其中：

$$①SC_1(U,V;X)=\frac{\sum_{i=1}^{c}\|v_i-\bar{v}\|_A^2/c}{\sum_{i=1}^{c}\left(\sum_{k=1}^{n}(u_{ik})^m\|x_k-v_i\|_A^2/n_i\right)}$$

这里 $n_i=\sum_{k=1}^{n}u_{ik}$，易知，越大 SC_1对应越佳分类。SC_1考虑了数据集的几何结构。

$$②SC_2(U)=\frac{FS}{FC}=\frac{\text{模糊类间分离度}}{\text{模糊类内内聚度}}。$$

这里，$FS=\sum_{i=1}^{c-1}\sum_{r=1}^{c-i}\left(\sum_{k=1}^{n}(\min(u_{ik},u_{jk})^2)/n_{ij}\right)$，

其中 $j=r+i$，$n_{ij}=\sum_{k=1}^{n}\min(u_{ik},u_{jk})$；

$FC=\sum_{k=1}^{n}(\max_{1\leqslant i\leqslant c}u_{ik})^2/n_{\cup}$，其中 $n_{\cup}=\sum_{k=1}^{n}\max_{1\leqslant i\leqslant c}u_{ik}$；

SC_2考虑了数据集的模糊划分，其值越小，其分类法越佳。显然，对于综合的有效性指标函数 V_{Zahid}，最大值对应于最佳的分类数。

3.1.5 H. Sun-S. Wang-Q. Jiang 有效性 V_{wsj}(2003)[7]

$$V_{wsj}(U,V,c)=Scat(c)+\frac{Sep(c)}{Sep(C_{\max})}$$

其中：

$$①Scat(c)=\frac{\frac{1}{c}\sum_{i=1}^{c}\|\sigma(v_i)\|}{\|\sigma(X)\|},$$

这里$\|x\|=(x^Tx)^{\frac{1}{2}}$，$\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k$ 且

$$\sigma(X)=\{\sigma(X)^1,\sigma(X)^2,\cdots,\sigma(X)^s\}^T,\sigma(X)^P=\frac{1}{n}\sum_{k=1}^{n}(x_k^P-\bar{x}^P)^2$$

$$\sigma(v_i)=\{\sigma(v_i)^1,\sigma(v_i)^2,\cdots,\sigma(v_i)^s\}^T,\sigma(v_i)^P=\frac{1}{n}\sum_{k=1}^{n}u_{ik}(x_k^P-v_i^P)^2$$

$(P=1,2,\cdots,S)$

$$②Sep(c)=\frac{D_{\max}^2}{D_{\min}^2}\sum_{i=1}^{c}\left(\sum_{j=1}^{c}\|v_i-v_j\|^2\right)^{-1}),$$

这里 $D_{\min}=\min_{i\neq j}\|v_i-v_j\|(i,j\in[1,c])$；$D_{\max}=\max_{i,j}\|v_i-v_j\|(i,j\in[1,c])$ $Scat(c)$表示类内紧凑度，$Sep(c)$表示类间分离度，权重因子 $\alpha=Sep(C_{\max})$用来补偿它们的度量差别，它实际是一个关于类内紧凑度和类间分离度的线性组合：$\alpha\cdot Scat(c)+\beta\cdot Sep(c)$，它的本质就是如何确定 α,β，或 α 与 β 的比值，以补偿它们之间的度量差别。最小的 $V_{wsj}(U,V,c)$对应于最佳的聚类数。

3.2 一个新的有效性指标函数

在实验的过程中，我们发现，当前使用的有效性指标对于类间有交叠或有多孤立点情况的处理还存在缺陷，对此类情况无法做出准确判别。然而在实践应用中，类间有交叠或有多孤立点情况是经常出现的，这引发了我们的研究。综合考虑前面各指标的优缺点，并结合实验进行深入考察与验证后，我们提出了如下的指标函数：

$$V_{hong}(U,V;X)=Coh(c)+\alpha\cdot Dis(c)$$

其中：

$$①Coh(c)=\sum_{i=1}^{c}\left(\sum_{k=1}^{n}(u_{ik})^m\|x_k-v_i\|_A^2/n_i\right)$$

$$②Dis(c)=1/\min_{i\neq j}\|v_i-v_j\|^2$$

$$③\alpha=\frac{B(C_{\max})}{S(C_{\max})}=\frac{\sum_{i=1}^{c_{\max}}\|v_i-\bar{v}\|_A^2/c_{\max}}{\sum_{i=1}^{c_{\max}}\left(\sum_{k=1}^{n}(u_{ik})^m\|x_k-v_i\|_A^2/n_i^{1/2}\right)}$$

这里 $\bar{v}=\frac{1}{c}\sum_{i=1}^{c}v_i$，$n_i=\sum_{k=1}^{n}u_{ik}$。

$Coh(c)$表示类内紧凑度，这样的定义可以把各个类的个性体现出来，其值越小越好。$Dis(c)$表示类间分离度，当类间分离得较好时，类间的最小距离将大些，通过取倒数，将类内与类间数值的矛盾消除，从而达到预期的效果。

该指标的核心在于比例因子的计算。首先要注意的是，比例因子是关于 $C_{\max}$的函数，这是我们经过分析比较并结合实验得出的结论：在各种不同的聚类数中，计算指标值必须有一个共同的参照，我们把参照选定为关于 $C_{\max}$的函数。其次，我们考虑 $Coh(c)$与 $Dis(c)$度量值的差距，参照标准化处理方法，并结合类比的思想，把每一个类及其所属成员想象为一个大成员，由这些大成员又组成一个新的“大类”，如图2所示。

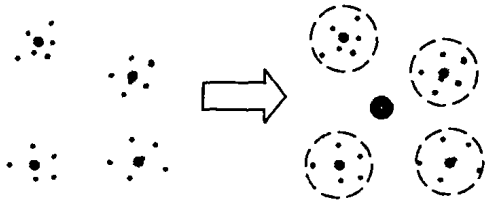

图2

这样，通过“大类”与原来“小类”的相似性，很容易把“大类”的度量缩放到小类中去。基于此思想并结合实验，我们提出了如上指标函数。比例因子中的 $B(C_{\max})$代表“大类”中各成员到“大”聚类中心的平均距离，$S(C_{\max})$代表“小类”中各成员到“小”聚类中心的平均距离。最小的 $V_{hong}(U,V;X)$对应于

最佳聚类数。

该指标的另一创新之处在于对 $S(C_{max})$ 的定义，比例因子作为一个共同的参照，必须或尽可能地做到对各种不同分类法，类的不同大小不敏感，也就是说要归一到统一的共性上。很容易看到它与 $Coh(c)$ 定义的区别，我们对 $n_i=\sum_{k=1}^{n}u_{ik}$ 做了开方处理，即 n_i^g，$g=1/2$。

为什么做这样的处理？考虑一般情况，假设在分类数 C 较小的时候，某个类有50个成员，它们到该类聚类中心的距离总和为 D_1，随着 C 的变大，该类成员数变少，假设现在剩下10个成员，它们到聚类中心的距离总和为 D_2，那么 D_1/D_2 会等于5吗？显然不会，究其原因，它们的聚类中心已经发生了偏移，D_1/D_2 将小于5。为了尽量屏蔽这种情况引起的误差，我们引入了一个新的参数 g，暂且称它为"泛化因子"，这里取 $g=1/2$。同时，为了保持各种分类法中各个分类的个性，$Coh(c)$ 不做这样的处理。

4 实验结果比较分析

为说明该新指标的性能，以下我们将对实验结果做一个报告。实验中我们测试了多组各种各样的数据，现仅以较具代表性的四组来做一个说明。其中两组数据来自于公众领域，另两组由混合高斯分布生成。这几组数据已经在文[7]得到验证。实验中，统一取 m=2，$\delta=0.001$，$C_{min}=2$，$C_{max}=10$，$\|\square\|$ 定义为欧几里得距离。初始化方法采用 Greedy 算法。为确定最佳聚类数，将 V_{xie}，V_{bsaid}，V_{kwon}，V_{zahid}，V_{wsj} 与新指标 V_{hong} 进行了综合比较。

4.1 测试数据集一(X_{30})

该数据集来源于文[9]，它由30个2-维的样本组成，分成3个类，每类含10个样本，如图3所示。对于 $c=2$ 到10的计算结果如表1所示。对于这个单的数据集，V_{bsaid}，V_{kwon}，V_{Zahid} 不能给出正确的聚类数。

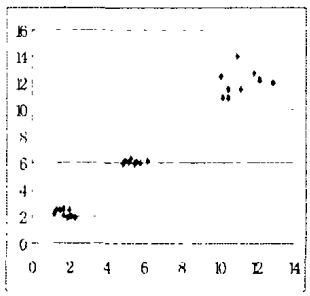

图3 X_{30}

表1 X_{30} 的各种有效性指标值

聚类结果输出

result	Vxie	Vbsaid	Vkwon	VZahid	Vwsj	Vhong
2	0.0399	0.0715	*1.4710	3.7056	0.1551	9.2580
3	*0.0228	0.0090	1.8210	16.203	*0.0343	[illegible]
4	0.0809	0.0054	12.898	26.870	0.1214	14.634
5	0.0466	0.0035	12.010	41.114	0.1069	12.491
6	0.0845	0.0024	26.408	54.772	0.2692	25.934
7	0.2736	0.0018	96.368	51.968	1.1353	108.39
8	0.2335	0.0014	104.42	69.030	1.1754	108.44
9	0.2025	0.0009	104.55	89.743	1.3192	108.57
10	0.1057	*0.0006	101.23	*108.54	1.0014	108.70

4.2 测试数据集二(IRIS)[1,10]

这是一组关于三种花的生物统计数据，通常称为 IRIS 数据集。它包含150个样本，每个样本有4-维，分别为花瓣与花萼的长和宽。它分为3类，每类包含50个样本。IRIS 是数据分析中最经常使用的基准测试数据之一。图4与图5分别是它的1-3-4维和1-2-3维投影图。

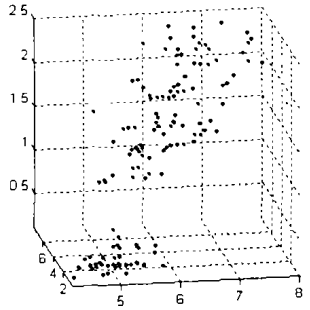

图4 IRIS：1-3-4维

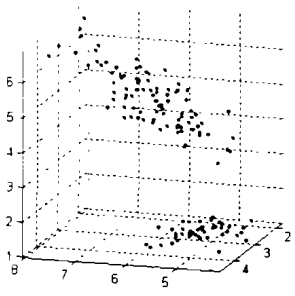

图5 IRIS：1-2-3维

虽然该数据集不大，但却包含着大数据集中经常遇见的重要特征。在它的三个类中，其中两个是互相重叠的。重叠性使得以前定义的很多指标不能对聚类的数目做出准确的估计。对于 $c=2$ 到10的计算结果如表2所示。在这组数据集里，由于类间重叠的原因及孤立点的干扰，很多指标都不能给出正确的聚类数，而仅有 V_{wsj} 和我们的指标 V_{hong} 能给出准确的判定。

表2 IRIS 的各种有效性指标值

聚类结果输出

result	Vxie	Vbsaid	Vkwon	VZahid	Vwsj	Vhong
2	*0.0545	0.1024	*8.4571	2.4064	0.1768	1.6876
3	0.1346	0.0576	21.605	2.9892	*0.1245	[illegible]
4	0.1934	0.0442	31.809	2.7665	0.1859	1.8765
5	0.2270	0.0338	38.668	*3.0437	0.2797	2.1930
6	0.3194	0.0312	54.749	2.1976	0.4909	2.9160
7	0.3739	0.0273	64.858	1.7067	0.6768	3.5445
8	0.4277	0.0262	74.428	0.9887	0.9409	4.1961
9	0.2770	0.0163	52.173	1.3894	0.6826	3.7841
10	0.3529	*0.0147	68.277	1.0424	-1.0174	4.9660

4.3 测试数据集三(6类)

用真实数据测试聚类算法的不利之处在于聚类数目与聚类中心的真实值是未知的。因此也不可能据此去计算聚类算法引起的误差。而这正是我们使用生成数据评价聚类算法的原因。这里，我们首先使用混合高斯分布生成具有重叠性的数据集，并预先已知其最佳聚类数，接下来，就可用它来评价各种有效性指标的优越性了。

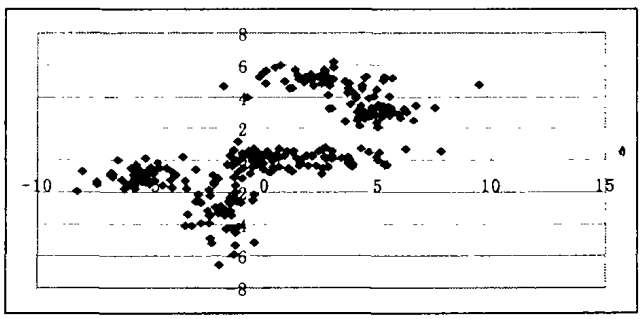

图6 数据集三：有6个类的2-维数据，类间高度重叠

数据集三由混合高斯分布生成，有300个2-维的样本，分成6个类，每个类50个样本，如图6。类1、3，类2、4，类5，6是重叠的。由于重叠性，加上含有一定数量的孤立点，这个数据集是相对较难聚类的。表3是使用各种有效性指标计算所得的指标值。

在这组数据集里，结果比较令人满意，仅有 V_{bsaid} 不能给出正确的聚类数。

表3　数据集三的各种有效性指标值

聚类结果输出

result	Vxie	Vbsaid	Vkwon	VZahid	Vwsj	Vhong
2	0.1058	0.2117	32.019	0.9088	0.3018	12.517
3	0.1275	0.1075	38.983	1.1627	0.1809	9.9923
4	0.1323	0.0652	40.782	1.5114	0.1686	8.5391
5	0.1269	0.0378	39.973	1.8549	0.1565	7.1937
6	*1.0949	0.0286	*30.247	*1.9045	*0.1530	*6.2436
7	0.3105	0.0235	100.56	1.7325	0.4370	6.3545
8	0.2685	0.0203	87.918	1.2583	0.4449	5.2661
9	0.5583	0.0163	186.51	1.2412	1.0255	6.5316
10	0.4840	*0.0136	165.04	1.0927	1.0266	6.3613

4.4　测试数据集四(5类)

与数据集三一样，数据集四也是采用混合高斯分布生成的。它有250个3-维的样本，分成5个类，每个类50个样本，如图7所示。在这个数据集里，类1和3重叠。表4是使用各种有效性指标计算所得的指标值。可以看到，我们的指标 V_{hong}，还有 V_{wsj} 都可以给出正确的结果。

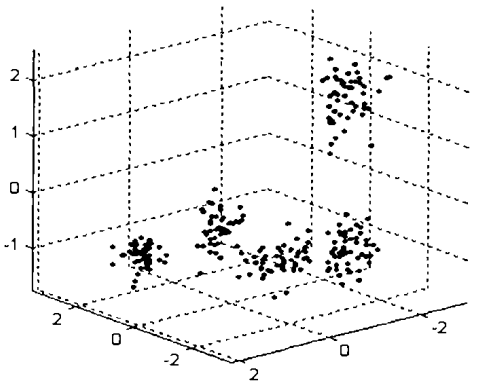

图7　有5个类的3-维数据

表4　数据集四的各种有效性指标值

聚类结果输出

result	Vxie	Vbsaid	Vkwon	VZahid	Vwsj	Vhong
2	0.1634	0.3225	41.131	0.6015	0.3845	3.3840
3	0.0861	0.0837	22.071	1.9007	0.1806	2.1539
4	*0.0796	0.0357	*21.074	3.4266	0.1118	1.5093
5	0.0872	0.0216	23.957	4.4221	*0.1044	*1.4165
6	0.3501	0.0168	99.166	*4.7911	0.3444	3.3189
7	0.6284	0.0147	181.45	4.7400	0.6368	5.6191
8	0.8699	0.0122	255.39	4.1523	0.9274	8.4019
9	0.7301	0.0112	218.03	3.5774	1.0180	8.4418
10	0.6922	*0.0101	210.56	3.4615	1.0217	8.4648

综上，新的有效性指标显著地改善了应用 FCM 算法确定最佳聚类数的性能。在测试的多组数据集中，我们的有效性指标都能给出正确的聚类数。V_{xie}，V_{kwon} 和 V_{zahid} 虽然有时对于类间交叠且多孤立点的情况可以给出正确的判定，但它们不稳定，经常会出现难以预料的结果，如对于简单的数据集 X_{30}，K_{kwon}，V_{Zahid} 竟然不能给出正确的结果。V_{bsaid} 虽说是 V_{xie} 的改进，但从上面的实验结果可以看出，该指标仅在局部上对定义进行了细化，但整体性能上并没有提高。值得注意的是指标 V_{wsj}，对测试的多组数据，它也都给出了很准确的判定。在接下来的工作中，我们将重点与 V_{wsj} 做更深入的比较分析，如从时间复杂性、空间复杂性等。

总结与展望　本文主要是提出了一个新的确定聚类个数的有效性指标。该指标 V_{hong} 实际是原始数据集 X、模糊划分矩阵 U、聚类中心 V 的函数。实验结果表明，该新指标不管是对于类间无重叠、重叠，或者具有多个孤立点等情况都能给出正确的判定，而现存的指标却可能给出难以预料的结果。很有必要去从理论上证明为什么新指标可以产生这样令人满意的结果，但据我们所知，当前文献中对有效性指标的评估都是从实验结果入手的。结合统计领域的相关知识，并采用恰当的数学模型可对指标的深层次行为有更进一步的了解。

我们也已证明，新指标应用于 K-Mean 算法(硬划分)中也具有同样的优越性能。FCM 算法实质上是 K-Mean 的自然推广，K-Mean 是 FCM 算法在模糊因子的特例。实际上，新指标不做任何修改就可适用于 K-Mean，这一点已经通过了实验的验证。指标中"泛化因子"将是下一步研究的主要内容，我们将从理论和实验上对"泛化因子"做更深入的考察。

参 考 文 献

1　Gonzalez T. Clustering to Minimize and Maximum Intercluster Distance. Theoretical Computer Science, 1985,38:293～306

2　Pal N R, Bezdek J C. On Cluster Validity for the Fuzzy C-Mean Model. IEEE Transactions on Fuzzy Systems [J], 1995. 370～390

3　Xie X, Beni G. A Validity Measure for Fuzzy Clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 1991, 13(8):841～847

4　Bensaid A M. Validity-Guided (Re) Clustering with Applications to Image Segmentation. IEEE Transactions on Fuzzy Systems, 1996,4(2)

5　Kwon S H. Cluster validity index for fuzzy clustering. ELECTRONICS LETTERS, 1998,34(22):2176～2177

6　Zahid N, Limouri M, Essaid A. A New cluster-validity for fuzzy clustering. Pattern Recognition Letters, 1999,32:1089～1097

7　Sun H, Wang S, Jiang Q. A New Validation Index for Determining the Number of Clusters in a Data Set. IJCNN'01, Washington DC, July 2001. 14～19

8　Sun H, Wang S, Jiang Q. FCM-Based Model Selection Algorithms for Determining the Number of Cluster. By Pattern Recognition, 2003

9　Bezdek J C. Chapter F6: Pattern Recognition in Handbook of Fuzzy computation. IOP Publishing Ltd, 1998

10　Anderson E. The Iris of the Gaspé Peninsula. Bulletin of American Iris Society, 1935,59: 2～5

11　Rezaee M, Letlieveldt B, Reiber J. A new cluster validity index for the Fuzzy c-means. Pattern Recognition Letters, 1998,19:237～246