一种基于几何概率的聚类有效性函数

李晓雯[1)] 毛政元[1)] 李建微[2)]

[1)](福州大学福建省空间信息工程研究中心,空间数据挖掘与信息共享教育部重点实验室,福州 350002)

[2)](福州大学计算机图像图形研究所,福州 350002)

摘 要 聚类有效性是聚类分析中尚未解决的基本问题,最佳聚类数的确定是聚类有效性问题中的主要研究内容。以几何概率为理论依据,针对 2 维数据集提出了一种新的聚类有效性函数,用于确定最佳聚类数。该函数利用 2 维数据集与 2 维离散点集之间存在的对应关系,以 2 维离散点集在特征空间中的分布特征为依据,测度对应数据集的聚类结构,思路直观、容易理解。测度过程中,将点集中的点两两相连生成一个线段集合保存点集的结构信息,通过比较线段集合中线段方向取值与完全随机条件下线段方向取值的相对大小,构造聚类有效性函数。实验结果表明,针对给定的样本数据集,生成该函数的曲线,再根据曲线的形态能够有效地确定 2 维数据集的最佳聚类数,指导聚类算法设计。

关键词 聚类有效性 几何概率 聚类分析 最佳聚类数

中图法分类号:TP301.6 **文献标识码**:A **文章编号**:1006-8961(2008)12-2351-06

A Cluster Validity Function Based on Geometric Probability

LI Xiao-Wen[1)], MAO Zheng-Yuan[1)], LI Jian-Wei[2)]

[1)](*Key Laboratory of Spatial Data Mining and Information Sharing of Ministry of Education, Spatial Information Research Center, Fuzhou University, Fuzhou* 350002)

[2)](*Computer Graphic and Image Institute, Fuzhou University, Fuzhou* 350002)

Abstract Determining optimum cluster number is a key research topic included in cluster validity, a fundamental unsolved problem in cluster analysis. In order to determine the optimum cluster number, this article proposes a new cluster validity function for two dimensional datasets theoretically based on geometric probability. The function uses of the relationship between a two dimensional dataset and the corresponding two dimensional discrete point set to measure the cluster structure of the dataset according to the distributive feature of the point set in the characteristic space. It is designed from the perspective of intuition and thus can be easily understood. During the process of measurement, the structure information of the point set has been stored in a line segment set generated by connecting each pair points in the point set. The cluster validity function is formed by comparing the values of line segment direction in the line segment set with those resulted from completely random condition. In the case study, it is testified that the pattern of the function curve generated with a given example dataset effectively enables the determination of the optimum cluster number of the dataset and supports the design of cluster algorithms.

Keywords Cluster validity, Geometric probability, Cluster analysis, The optimum cluster number

1 引 言

聚类是一种重要的多元统计分析方法,也是数据挖掘中不可替代的分析工具,被广泛应用于各种专题研究之中[1]。完整的聚类分析包括 3 个方面的研究内容:(1)聚类趋势分析;(2)聚类结构提取;(3)聚类结果评价。其中聚类趋势分析的任务是判

基金项目:国家自然科学基金项目(40471113)

收稿日期:2006-09-04;**改回日期**:2007-06-06

第一作者简介:李晓雯(1977~),女。现为福建省龙岩学院助教。2007 年于福州大学获地图学与地理信息系统专业硕士学位。主要研究领域为空间数据挖掘和聚类分析。E-mail: Mosquito-LXW@126.com。

通讯作者:毛政元,E-mail: zymao@fzu.edu.cn

断一个给定的数据集是否具有聚类结构;聚类结构提取就是狭义的聚类分析,即运用某种算法得到聚类结果的过程;聚类结果评价是指根据某种标准判定聚类结果的合理性,即得到的聚类结果(包括类的层次、每一层次上类的数目和类的边界)与数据集中实际存在的聚类结构是否吻合,也称为聚类有效性研究,有效性问题又经常转化为最佳聚类数的自动确定[2],聚类有效性函数是测度聚类结果合理性的基本方法。

已经实现的聚类结构提取算法很多[3~6],由于同一算法对于不同数据集和不同算法对于同一数据集的适应能力不同,在理论上如何评价一个聚类算法的适应性、在应用中如何针对一个特定数据集选择适用的聚类算法成为一个必须解决的问题,聚类有效性因此成为聚类研究中的热点问题,受到高度重视,相关的研究很多[7~10]。目前各类文献中较有代表性的聚类算法一般要求预先指定最佳聚类数区间,然后利用穷举法选定最佳聚类数,只能适应数据量较少的数据集,在涉及对象数量庞大的空间聚类(如遥感影像非监督分类)中缺乏可行性,本文以几何概率[11]为理论基础提出并实现了一种能够适应大数据量的聚类有效性函数。

2 基于几何概率的聚类有效性函数[12]

2.1 函数的解析表达式

设 2 维数据集 $\boldsymbol{S}$ 中各对象第 1 项属性的最小和最大值分别为 $x_{\min}$,$x_{\max}$,第 2 项属性的最小和最大值分别为 $y_{\min}$,$y_{\max}$,令:

$$a=\min[(x_{\max}-x_{\min}),y_{\max}-y_{\min}]$$

$$b=\max[(x_{\max}-x_{\min}),y_{\max}-y_{\min}]$$

则 $\boldsymbol{S}$ 被映射为以 a,b 为边长的矩形区域上的一个离散点集(仍记为 $\boldsymbol{S}$),运用几何概率,以文献[12]中关于正方形区域内集聚型点模式结构信息指示函数和结构信息提取算法为基础,可以设计出测度点集几何分布结构信息的函数。

将离散点集 $\boldsymbol{S}$ 中的 n 个点两两相连得到一个线段集合

$$\boldsymbol{L}(s_i;r_i;\theta_i;i=1,2,3,\cdots)$$

其中,s_i,r_i,θ_i 分别表示 $\boldsymbol{L}$ 中第 i 条线段的中点坐标、长度及其与纵轴正向所成角,且有

$$0<r_i\leqslant\sqrt{a^2+b^2},0\leqslant\theta_i\leqslant\pi,i=1,2,3,\cdots$$

针对 $\boldsymbol{L}$ 构造函数 H

$$H(\theta,\Delta\theta,r,\Delta r)=\frac{N}{T}\times\frac{1}{P(\theta,\Delta\theta,r,\Delta r)}$$

式中,T 表示 $\boldsymbol{L}$ 中线段的总数,且

$$T=C_n^2=n(n-1)/2$$

N 表示 $\boldsymbol{L}$ 中满足条件

$$\left[\theta-\frac{\Delta\theta}{2},\theta+\frac{\Delta\theta}{2}\right]\bigcap\left[r-\frac{\Delta r}{2},r+\frac{\Delta r}{2}\right]$$

的线段数,可以由统计求得。

函数 $P(\theta,\Delta\theta,r,\Delta r)$ 的含义是在随机条件下,边长分别为 $a,b(a\leqslant b)$ 的矩形区域内线段满足条件

$$\left[\theta-\frac{\Delta\theta}{2},\theta+\frac{\Delta\theta}{2}\right]\bigcap\left[r-\frac{\Delta r}{2},r+\frac{\Delta r}{2}\right]$$

的概率。函数 $H(\theta,\Delta\theta,r,\Delta r)$ 的含义是集合 $\boldsymbol{L}$ 中线段数落入区间

$$\left[\theta-\frac{\Delta\theta}{2},\theta+\frac{\Delta\theta}{2}\right]\bigcap\left[r-\frac{\Delta r}{2},r+\frac{\Delta r}{2}\right]$$

的频率与完全随机情况下线段落入同一区间的概率之比,故其期望

$$E[H(\theta,\Delta\theta,r,\Delta r)]=1$$

当 $r=\frac{\sqrt{a^2+b^2}}{2},\Delta r=\sqrt{a^2+b^2}$ 时,有

$$P(\theta,\Delta\theta,r,\Delta r)=P\left(\theta,\Delta\theta,\frac{\sqrt{a^2+b^2}}{2},\sqrt{a^2+b^2}\right)=U(\theta,\Delta\theta)$$

则

$$H(\theta,\Delta\theta,r,\Delta r)=\frac{N}{T}\times\frac{1}{U(\theta,\Delta\theta)}=H(\theta,\Delta\theta)$$

且其期望

$$E[H(\theta,\Delta\theta)]=1$$

点集 $\boldsymbol{S}$ 的结构信息保存在集合 $\boldsymbol{L}$ 中,当 $\boldsymbol{S}$ 中的点存在多中心集聚分布时,$\boldsymbol{L}$ 中的线段必然存在集聚分布,即在某些区间上取值大于完全随机条件下的取值,函数 $H(\theta,\Delta\theta)$ 的图形在对应的区间上将取得峰值,峰值与 1 的差值表示集聚的程度,峰值个数表示集聚的方向数。由于点集的几何分布与数据集的特征之间、函数 $H(\theta,\Delta\theta)$ 的图形与点集的结构特征之间存在明确的对应关系,该函数即为聚类有效性函数,能够判断数据集中存在的聚类结构和针对数据集聚类结果的有效性。此过程中的难度在于推导函数 $P(\theta,\Delta\theta,r,\Delta r)$ 的解析表达式。

如图 1 所示,设矩形 $ABCD$ 的边长为 $a,b(a\leqslant b)$,按前文的定义,$P(\theta,\Delta\theta,r,\Delta r)$ 表示在某一给定矩形区域内可能取得的全部线段落入区间

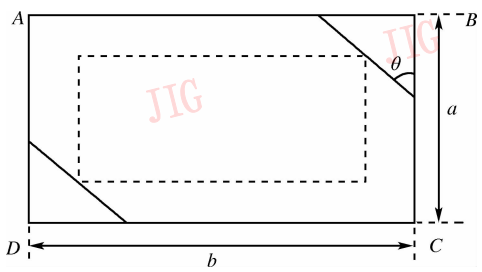

图 1　矩形区域内线段落入指定区间的测度

Fig. 1　The measurement of line segments in the specific intervals and within a rectangle

$$\left[\theta-\frac{\Delta\theta}{2},\theta+\frac{\Delta\theta}{2}\right]\bigcap\left[r-\frac{\Delta r}{2},r+\frac{\Delta r}{2}\right]$$

上的概率,采用线段中点落入区域的面积测度线段总量,对长度为 $r(0<r\leqslant\sqrt{a^2+b^2})$ 且与边 CB 正向成 $\theta(0\leqslant\theta\leqslant\pi)$ 角的线段,其落入矩形 $ABCD$ 区域内的可能性与其中点落入的区域(图 4 中虚线围成的矩形)面积

$$A(r,\theta)=(a-r|\cos\theta|)(b-r|\sin\theta|)$$

成正比。令

$$r_1=r-\frac{\Delta r}{2},r_2=r+\frac{\Delta r}{2};\theta_1=\theta-\frac{\Delta\theta}{2},\theta_2=\theta+\frac{\Delta\theta}{2}$$

则

$$P(\theta,\Delta\theta,r,\Delta r)=$$

$$P\left(\frac{\theta_1+\theta_2}{2},\theta_2-\theta_1,\frac{r_1+r_2}{2},r_2-r_1\right)=\frac{\int_{r_1}^{r_2}\int_{\theta_{\min}}^{\theta_{\max}}A(r,\theta)\mathrm{d}r\mathrm{d}\theta}{B}$$

式中,$\theta_{\min}=f(r_1,r_2)\geqslant\theta_1,\theta_{\max}=g(r_1,r_2)\leqslant\theta_2$,$f(r_1,r_2)$,$g(r_1,r_2)$ 表示以 r_1,r_2 为自变量的函数,

$$B=\int_0^a\int_0^{\pi}A(r,\theta)\mathrm{d}r\mathrm{d}\theta+\int_a^b\int_{\arccos\frac{a}{r}}^{\pi-\arccos\frac{a}{r}}A(r,\theta)\mathrm{d}r\mathrm{d}\theta+\int_b^{\sqrt{a^2+b^2}}\int_{\arccos\frac{a}{r}}^{\arcsin\frac{b}{r}}A(r,\theta)\mathrm{d}r\mathrm{d}\theta$$

$$=\frac{1}{3}a^3+\frac{1}{3}b^3+ab^2\ln\frac{\sqrt{a^2+b^2}+a}{b}+a^2b\ln\frac{\sqrt{a^2+b^2}+b}{a}-\frac{1}{3}(a^2+b^2)\sqrt{a^2+b^2}$$

2.2　函数分子部分的计算

(1) $0\leqslant r_1<r_2\leqslant a$

当 r_1,r_2 在区间 $[0,a]$ 上取值时,

$$f(r_1,r_2)=\theta_1,g(r_1,r_2)=\theta_2$$

$$\int_{r_1}^{r_2}\int_{\theta_{\min}}^{\theta_{\max}}A(r,\theta)\mathrm{d}r\mathrm{d}\theta=\int_{r_1}^{r_2}\int_{\theta_1}^{\theta_2}A(r,\theta)\mathrm{d}r\mathrm{d}\theta$$

$$=\int_{r_1}^{r_2}\int_{\theta_2}^{\theta_1}(a-r|\cos\theta|)(b-r|\sin\theta|)\mathrm{d}r\mathrm{d}\theta$$

记

$$\int_{r_1}^{r_2}\int_{\theta_1}^{\theta_2}(b-r\sin\theta)(a-r\cos\theta)\mathrm{d}r\mathrm{d}\theta$$

$$=ab(\theta_2-\theta_1)(r_2-r_1)+\frac{1}{6}(r_2^3-r_1^3)\times(\sin^2\theta_2-\sin^2\theta_1)-\frac{1}{2}(r_2^2-r_1^2)\times(b\sin\theta_2-b\sin\theta_1-a\cos\theta_2+a\cos\theta_1)$$

$$=M(a,b,r_1,r_2,\theta_1,\theta_2)$$

此时分子的计算按 θ 取值区间细分为 3 种情况,如表 1。

表 1　P 函数分子的值

Tab. 1　The value of numerator in P Function

角　度	$0<\theta_1<\theta_2\leqslant\frac{\pi}{2}$	$0<\theta_1<\frac{\pi}{2}<\theta_2\leqslant\pi$	$\frac{\pi}{2}\leqslant\theta_1<\theta_2\leqslant\pi$
分子值	$M(a,b,r_1,r_2,\theta_1,\theta_2)$	$M\left(a,b,r_1,r_2,\theta_1,\frac{\pi}{2}\right)+M\left(a,b,r_1,r_2,\pi-\theta_2,\frac{\pi}{2}\right)$	$M(a,b,r_1,r_2,\pi-\theta_2,\pi-\theta_1)$

(2) $a\leqslant r_1<r_2\leqslant\sqrt{a^2+b^2}$

先计算 3 个积分表达式,令

$$0<a\leqslant l_1<l_2\leqslant\sqrt{a^2+b^2},0\leqslant\theta_1<\frac{\pi}{2},$$

$0<\theta_2\leqslant\frac{\pi}{2}$,则

① $\int_{l_1}^{l_2}\int_{\arccos\frac{a}{r}}^{\arcsin\frac{b}{r}}A(r,\theta)\mathrm{d}r\mathrm{d}\theta$

$$=\int_{l_1}^{l_2}\int_{\arccos\frac{a}{r}}^{\arcsin\frac{b}{r}}(a-r\cos\theta)(b-r\sin\theta)\mathrm{d}r\mathrm{d}\theta$$

$$=abl_2\left(\arcsin\frac{b}{l_2}-\arccos\frac{a}{l_2}\right)-abl_1\left(\arcsin\frac{b}{l_1}-\arccos\frac{a}{l_1}\right)+\frac{1}{2}ab^2\ln\frac{l_2+\sqrt{l_2^2-b^2}}{l_1+\sqrt{l_1^2-b^2}}+\frac{1}{2}a^2b\ln\frac{l_2+\sqrt{l_2^2-a^2}}{l_1+\sqrt{l_1^2-a^2}}+$$

$$\frac{1}{2}a(l_2\sqrt{l_2^2-b^2}-l_1\sqrt{l_1^2-b^2})-\frac{1}{2}(a^2+b^2)\times$$

$$(l_2-l_1)-\frac{1}{6}(l_2^3-l_1^3)+\frac{1}{2}b(l_2\sqrt{l_2^2-a^2}-$$

$$l_1\sqrt{l_1^2-a^2})=L(a,b,l_1,l_2)$$

② $$\int_{l_1}^{l_2}\int_{\arccos\frac{a}{r}}^{\theta_2}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta$$

$$=\int_{l_1}^{l_2}\int_{\arccos\frac{a}{r}}^{\theta_2}(a-r\cos\theta)(b-r\sin\theta)\,\mathrm{d}r\mathrm{d}\theta$$

$$=-ab\left(l_2\arccos\frac{a}{l_2}-l_1\arccos\frac{a}{l_1}\right)+$$

$$\frac{1}{2}a^2b\ln\frac{l_2+\sqrt{l_2^2-a^2}}{l_1+\sqrt{l_1^2-a^2}}+\frac{1}{2}b(l_2\sqrt{l_2^2-a^2}-$$

$$l_1\sqrt{l_1^2-a^2})+\left(ab\theta_2-\frac{1}{2}a^2\right)(l_2-l_1)-$$

$$\frac{1}{2}(b\sin\theta_2-a\cos\theta_2)(l_2^2-l_1^2)-$$

$$\frac{1}{6}\cos^2\theta_2(l_2^3-l_1^3)=C(a,b,l_1,l_2,\theta_2)$$

③ $$\int_{l_1}^{l_2}\int_{\theta_1}^{\arcsin\frac{b}{r}}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta$$

$$=\int_{l_1}^{l_2}\int_{\theta_1}^{\arcsin\frac{b}{r}}(a-r\cos\theta)(b-r\sin\theta)\,\mathrm{d}r\mathrm{d}\theta$$

$$=ab\left(l_2\arcsin\frac{b}{l_2}-l_1\arcsin\frac{b}{l_1}\right)+$$

$$\frac{1}{2}ab^2\ln\frac{l_2+\sqrt{l_2^2-b^2}}{l_1+\sqrt{l_1^2-b^2}}+\frac{1}{2}a(l_2\sqrt{l_2^2-b^2}-$$

$$l_1\sqrt{l_1^2-b^2})-\left(ab\theta_1+\frac{1}{2}b^2\right)(l_2-l_1)+$$

$$\frac{1}{2}(b\sin\theta_1-a\cos\theta_1)(l_2^2-l_1^2)-$$

$$\frac{1}{6}\sin^2\theta_1(l_2^3-l_1^3)=S(a,b,l_1,l_2,\theta_1)$$

当 r 在区间 $[a,\sqrt{a^2+b^2}]$ 上取值时，$[f(r_1,r_2),g(r_1,r_2)]$ 可表示为

$$[\theta_1,\theta_2]\bigcap\left(\arccos\frac{a}{r},\pi-\arccos\frac{a}{r}\right)$$

此时，先计算积分表达式

$$\int_a^l\int_{\theta_1}^{\theta_2}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta$$

的值，记

$$R(l,\theta_1,\theta_2)=\int_a^l\int_{\theta_1}^{\theta_2}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta$$

则 $R(l,\theta_1,\theta_2)$ 的计算如表 2，其中

$$\alpha=\arccos\frac{a}{b},\beta=\arccos\frac{a}{\sqrt{a^2+b^2}}$$

当 $a\leqslant r_1<r_2\leqslant\sqrt{a^2+b^2}$ 时，分子的值为

$$R(r_2,\theta_1,\theta_2)-R(r_1,\theta_1,\theta_2)$$

(3) $0\leqslant r_1\leqslant a\leqslant r_2\leqslant\sqrt{a^2+b^2}$

分子的值为

$$\int_{r_1}^{r_2}\int_{\theta_1}^{\theta_2}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta=\int_{r_1}^{a}\int_{\theta_1}^{\theta_2}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta+\int_{a}^{r_2}\int_{\theta_1}^{\theta_2}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta$$

$$=\int_{r_1}^{a}\int_{\theta_1}^{\theta_2}A(r,\theta)\,\mathrm{d}r\mathrm{d}\theta+R(r_2,\theta_1,\theta_2)$$

3 实验研究

图 3 是针对 2 维双中心数据集(如图 2 所示)生成的 $H(\theta,\Delta\theta)$ 的图形，它是一条有明显起伏的单峰曲线，其峰值在 $\theta=\pi/2$ 处，说明数据集中存在一个集聚方向，分别沿该集聚方向及集聚方向的法向以一定宽度的条带扫描，记录条带内密度处于峰值的位置，可以分离出 2 个明显的子类，与事实吻合。

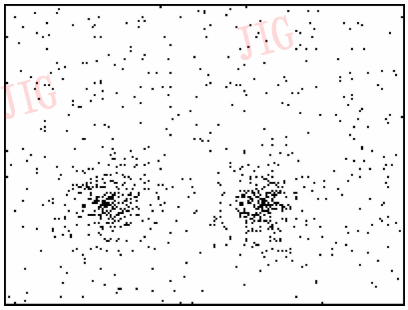

图 2 双中心集聚点集

Fig. 2 A point set with two clusters

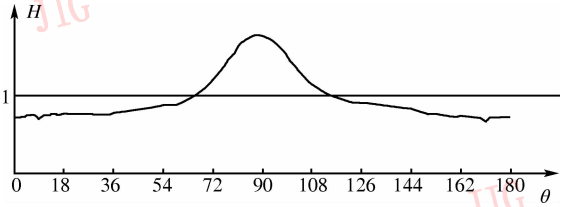

图 3 双中心集聚点集对应的 H 函数图形 $\left(\Delta\theta=\frac{\pi}{10}\right)$

Fig. 3 The graph of H fuction corresponding to the point set with two clusters $\left(\Delta\theta=\frac{\pi}{10}\right)$

图 5 是针对 2 维多中心数据集(如图 4 所示)生成的 $H(\theta,\Delta\theta)$ 的曲线，该曲线上有 3 个明显的峰值，对应于数据集的 3 个集聚方向，以同样的方法对点集扫描可以分离出该数据集 3 个明显的子类，此结果符合数据集的实际情况。

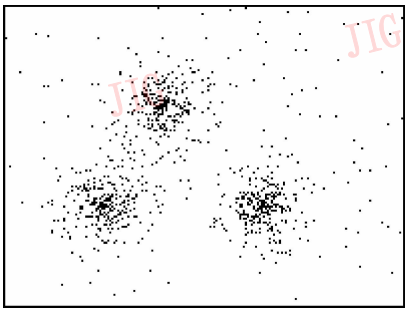

图 4　多方向集聚

Fig. 4　Cluster in multiple directions

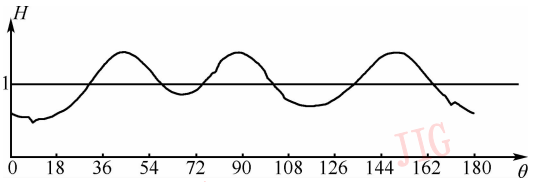

图 5　3 中心多方向集聚点集对应的 H 函数图形 $\left(\Delta\theta = \frac{\pi}{10}\right)$

Fig. 5　The graph of H fuctions corresponding to the point set clustering in multiple directions $\left(\Delta\theta = \frac{\pi}{10}\right)$

表 2　$R(l,\theta_1,\theta_2)$ 的值

Tab. 2　The value of $R(l_1,\theta_1,\theta_2)$

角　度	l_1、l_2 取值	l 的范围	$R(l,\theta_1,\theta_2)$ 的值
$0 < \theta_1 < \theta_2 < \beta$	$l_1 = a/\cos\theta_1$ $l_2 = a/\cos\theta_2$	$a < l \leqslant l_1$	$M(a,b,a,l,\theta_1,\theta_2)$
		$l_1 < l < l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + C(a,b,l_1,l,\theta_2)$
		$l \geqslant l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + C(a,b,l_1,l_2,\theta_2)$
$0 < \theta_1 < \alpha < \beta \leqslant \theta_2 \leqslant \frac{\pi}{2}$	$l_1 = a/\cos\theta_1$ $l_2 = b/\sin\theta_2$	$a < l \leqslant l_1$	$M(a,b,a,l,\theta_1,\theta_2)$
		$l_1 < l < l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + C(a,b,l_1,l_2,\theta_2)$
		$l \geqslant l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + C(a,b,l_1,l_2,\theta_2) + L(a,b,l_2,l)$
$\alpha \leqslant \theta_1 \leqslant \beta < \theta_2 \leqslant \frac{\pi}{2}$	$a\sin\theta_2 - b\cos\theta_1 > 0$ $l_1 = b/\sin\theta_2$ $l_2 = a/\cos\theta_1$	$a < l \leqslant l_1$	$M(a,b,a,l,\theta_1,\theta_2)$
		$l_1 < l < l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + S(a,b,l_1,l,\theta_1)$
		$l \geqslant l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + S(a,b,l_1,l_2,\theta_1) + L(a,b,l_2,l)$
	$a\sin\theta_2 = b\cos\theta_1$ $l_1 = a/\cos\theta_1$ $l_2 = b/\sin\theta_2$	$a < l \leqslant l_1$	$M(a,b,a,l,\theta_1,\theta_2)$
		$l \geqslant l_1 = l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + L(a,b,l_2,l)$
	$a\sin\theta_2 - b\cos\theta_1 < 0$ $l_1 = a/\cos\theta_1$ $l_2 = b/\sin\theta_2$	$a < l \leqslant l_1$	$M(a,b,a,l,\theta_1,\theta_2)$
		$l_1 < l < l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + C(a,b,l_1,l,\theta_2)$
		$l \geqslant l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + C(a,b,l_1,l_2,\theta_2) + L(a,b,l_2,l)$
$\beta < \theta_1 < \theta_2 \leqslant \frac{\pi}{2}$	$l_1 = b/\sin\theta_2$ $l_2 = b/\sin\theta_1$	$a < l \leqslant l_1$	$M(a,b,a,l,\theta_1,\theta_2)$
		$l_1 < l < l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + S(a,b,l_1,l,\theta_1)$
		$l \geqslant l_2$	$M(a,b,a,l_1,\theta_1,\theta_2) + S(a,b,l_1,l_2,\theta_1)$
$0 < \theta_1 < \frac{\pi}{2} < \theta_2 \leqslant \pi$			$R\left(l,\theta_1,\frac{\pi}{2}\right) + R\left(l,\pi-\theta_2,\frac{\pi}{2}\right)$
$\frac{\pi}{2} \leqslant \theta_1 < \theta_2 \leqslant \pi$			$R(l,\pi-\theta_2,\pi-\theta_1)$

4　结　论

聚类有效性是聚类研究中尚未解决的瓶颈问题[10, 13~16],具有重要的理论意义和实用价值。本文以几何概率为理论基础,针对 2 维数据集提出一种新的聚类有效性函数,直接根据 2 维数据集在样本特征空间的结构特征合理确定数据集的子类(最佳聚类数),无需引入参数,不依赖统计假设,思路新颖直观。实验结果表明,本文所实现的聚类有效性函数不仅能够确定最佳聚类数,而且适应性强,对类间边界模糊、容易误判的数据集和大数据集同样有效。H 函数实现算法的时间复杂度取决于其中的一个二重循环,该循环体内的计算量为 $O(1)$,二重循

环的总次数为 $O(n^2)$，即该算法的时间复杂度为 $O(n^2)$。对于高维数据集，可以先通过降维简化数据集，再应用本文提出的聚类有效性函数；也可将数据集拆分为一组 2 维数据集，分别处理后再汇总分析结果。如何扩展该聚类有效性函数使其适应高维数据集是下一步研究的内容。

参考文献(References)

1 Han J, Kamber M. Data Mining: Concepts and Techniques [M]. Los Altos, CA, USA: Morgan Kaufmann, 2001.

2 Bezdek J C, Pal N R. Some new indexes of cluster validity [J]. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybercetics, 1998, **28**(3): 301 ~ 305.

3 Backer E, Jain A K. A clustering performance measure based on fuzzy set decomposition [J]. IEEE Transactions on PAMI, 1981, **3**(1): 66 ~ 95.

4 Windham M P. Cluster validity for the fuzzy c-means clustering algorithm [J]. IEEE Transactions on PAMI, 1982, **4**(4): 357 ~ 363.

5 Al-Sultan K S, Selim S Z. Global algorithm for fuzzy clustering problem [J]. Pattern Recognition, 1993, **26**(9): 1357 ~ 1361.

6 Jain A K, Murty M N, Flynn P J. Data clustering: a review [J]. ACM Computing Surveys, 1999, **31**(3): 265 ~ 323.

7 Nakamura E, Kehtarnavaz N. Determining number of clusters and prototype locations via multi-scale clustering [J]. Pattern Recognition Letters, 1998, **19**(14): 1265 ~ 1283.

8 Pakhiraa M K, Bandyopadhyay S, Maulik U. Validity index for crisp and fuzzy clusters [J]. Pattern Recognition, 2004, **37**(3): 487 ~ 501.

9 Kim M, Ramakrishna R S. New indices for cluster validity assessment [J]. Pattern Recognition Letters, 2005, **26**(15): 2353 ~ 2363.

10 Kim D W, Lee K H, Lee D. On cluster validity index for estimation of optimal number of fuzzy clusters [J]. Pattern Recognition, 2004, **37**(10): 2009 ~ 2025.

11 Seneta E, Parshall K H, Jongmans F. Nineteenth-century developments in geometric probability: J. J. Sylvester, M. W. Crofton, J. É. Barbier, J. Bertrand [J]. Archive for History of Exact Sciences, 2001, **55**(6): 501 ~ 524.

12 Mao Zheng-yuan, Li Lin. The Measurement of Spatial Patterns and Its Application [M]. Beijing: Science Press, 2004. [毛政元, 李霖. 空间模式的测度及其应用[M]. 北京: 科学出版社出版, 2004.]

13 Wu S, Chow T W S. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density [J]. Pattern Recognition, 2004, **37**(2): 175 ~ 188.

14 Gao Xin-bo, Xie Wei-xin. Advances in theory and applications of fuzzy clustering [J]. Chinese Science Bulletin, 1999, **44**(21): 2241 ~ 2251. [高新波, 谢维信. 模糊聚类理论发展及应用的研究进展[J]. 科学通报. 1999, **44**(21): 2241 ~ 2251.]

15 Yu Jian. Cluster Validity and Its Application [D]. Beijing: School of Mathematical Sciences of Beijing University, 2000. [于剑. 聚类有效性及其应用[D]. 北京: 北京大学数学科学学院, 2000.]

16 Yu Jian. On the Fuzziness Index of the FCM Algorithms [J]. Chinese Journal of Computers, 2003, **26**(8): 968 ~ 973. [于剑. 论模糊 C 均值算法的模糊指标[J]. 计算机学报, 2003, **26**(8): 968 ~ 973.]