# 基于邻域模型的 K-means 初始聚类中心选择算法*)

曹付元[1,2] 梁吉业[1,2] 姜 广[1,2]
(计算智能与中文信息处理省部共建教育部重点实验室 太原 030006)[1]
(山西大学计算机与信息技术学院 太原 030006)[2]

**摘 要** 传统的 K-means 算法由于其方法简单,在模式识别和机器学习中被广泛讨论和应用。但由于 K-means 算法随机选择初始聚类中心,而初始聚类中心的选择对最终的聚类结果有着直接的影响,因此算法不能保证得到一个唯一的聚类结果。利用邻域模型中对象邻域的上下近似,定义了对象邻域耦合度和分离度的概念,给出了对象在初始聚类中心选择中的重要性,提出了一种初始聚类中心的选择算法。另外,分析了邻域模型中三种范数对聚类精度的影响,并和随机选择初始聚类中心、CCIA 选择初始聚类中心算法进行了比较,实验结果表明,该算法是有效的。

**关键词** 邻域模型,初始聚类中心,K-means 聚类,粗糙集

## Initial Cluster Centers Choice Algorithm for K-means Based on Neighborhood Model

CAO Fu-yuan[1,2] LIANG Ji-ye[1,2] JIANG Guang[1,2]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China)[1]
(School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)[2]

**Abstract** The traditional K-means algorithm considered as a simple method has been widely discussed and applied in pattern recognition and machine learning. However, K-means algorithm can not guarantee unique clustering result because initial cluster centers are chosen randomly, moreover, choosing initial cluster centers is extremely important as it has a direct impact on the formation of final clusters. In this paper, concepts of coupling and division are defined by using low approximation and upper approximation of object neighborhood, and importance of objects in the procedure of choosing cluster centers is also given, initial cluster centers choice algorithm for K-means based on neighborhood model is proposed. Compared with choosing initial cluster centers randomly and CCIA algorithms, cluster accuracy affected by three kinds of norm in neighborhood model is analyzed. The experimental results show that the algorithm is effective.

**Keywords** Neighborhood model, Initial cluster centers, K-means clustering, Rough set

聚类分析是数据挖掘研究和应用中的一个重要部分,由于聚类算法不对数据作任何统计假设,在模式识别和人工智能等领域,聚类算法常被称为一种无监督的学习。聚类分析是将数据对象分组成多个类或多个簇,在同一个簇中的对象具有较高的相似度,而不同簇中的对象差别较大[1]。目前聚类分析已被广泛应用于金融欺诈、医疗诊断、图像处理、信息检索和生物信息学等研究领域。

自 20 世纪 60 年代以来,聚类算法被广泛研究并得到了很好的应用[2-5],其中 1967 年 Q. J. Mac 提出的 K-means 聚类算法[6],由于其方法简单,已成为当前最流行的聚类算法之一,特别数据分布呈现类内团聚状,该算法能得到很好的聚类结果。但 K-means 算法只适用于数值型数据,因此许多研究者对 K-means 算法进行了扩展,Z. X. Huang 提出了 K-modes 和 K-prototypes 算法[7]。A. Ahmad 提出了针对混合数据的 K-means 聚类算法[8]。但 K-means 算法是以确定的类别数及选定的初始聚类中心为前提,算法的聚类结果受到取定的类别数及初始聚类中心的影响,聚类结果只能是局部最优,且不能保证得到一个唯一的聚类结果。针对初始聚类中心的选择,许多学者进行了研究。R. O. Duda 和 P. E. Hart 提出了一种初始平均值的回归方法[9]。P. S. Bradley 等提出了一种优化初始点的过程[10]。J. M. Penā 等对 K-means 算法的不同初始方法进行了比较[11]。S. S. Khan 和 A. Ahmad 提出了一种针对 K-means 算法的聚类中心初始化算法(CCIA)[12]。实验结果表明这些算法都优于传统的 K-means 算法,且随机和 Kaufman 初始化方法优于其它的初始化方法,因为它不依赖于对象的序[11]。

T. Y. Lin 提出了邻域模型的概念[13],该模型通过空间点的邻域来粒化论域空间,将邻域理解为基本信息粒子,用来描述空间中的其他概念。Y. Y. Yao 和 W. Z. Wu 分别研究了 1-step 和 k-step 邻域信息系统的性质[14,15]。Q. H. Hu 等利用拓扑空间中球形邻域的概念,构造了基于邻域粗糙集模型的特征选择算法[16]。邻域模型作为数值信息粒度的计算模型已经得到了成功的应用。

M. Meila 和 D. Heckerman 认为初始聚类中心的选择没有一种能普遍接受的方法[17]。邻域的大小与聚类中心选择有着一种必然的联系。本文在邻域模型的基础上,通过对象邻域的耦合度和分离度描述了对象在选择初始聚类中心过程中的重要性,提出了针对 K-means 算法的初始聚类中心确定方法。采用 UCI 国际标准数据验证了该方法的有效性,分析了邻域模型中三种范数对聚类精度的影响。

*)基金项目:国家 863 计划项目(2007AA01Z165),国家自然科学基金(70471003, 60773133),高等学校博士学科点专项科研基金(20050108604),教育部科学技术研究重点项目(206017),山西省重点实验室开放基金(200603023),山西省高校科技开发项目(2007103)和太原市科技局科技兴市专项项目(07010724)。**曹付元** 讲师,博士生,主要研究方向为数据挖掘、机器学习;**梁吉业** 教授,博士生导师,主要研究方向为粗糙集理论、数据挖掘、人工智能等;**姜 广** 助教,主要研究方向为概念格、数据挖掘。

## 1 邻域粗糙集模型基本概念

**定义 1**[16] 给定一个 $N$ 维的实数空间 $R$，$d:R^N\times R^N\to R$，我们称 $d$ 是 $R^N$ 上的一个度量，如果 $d$ 满足

(1)$d(x_1,x_2)\geqslant 0,\forall x_1,x_2\in R^N,d(x_1,x_2)=0$ 当且仅当 $x_1=x_2$；

(2)$d(x_1,x_2)=d(x_2,x_1),\forall x_1,x_2\in R^N$；

(3)$d(x_1,x_3)\leqslant d(x_1,x_2)+d(x_2,x_3),\forall x_1,x_2,x_3\in R^N$；

则称$\langle R,d\rangle$为度量空间。

**定义 2**[16] 给定实数空间上的非空有限集合 $U=\{x_1,x_2,\cdots,x_n\}$，对于 $U$ 上的任意对象 $x_i$，定义其 $\varepsilon$ 邻域为

$$\delta_\varepsilon(x_i)=\{x\mid x\in U,d(x,x_i)\leqslant\varepsilon\},$$

其中 $\varepsilon\geqslant 0$. $\delta_\varepsilon(x_i)$称为由 $x_i$ 生成的邻域信息粒子，简称为 $x_i$ 的邻域粒子。

根据度量的性质有：

(1)$\delta_\varepsilon(x_i)\neq\varnothing$，因为 $x_i\in\delta_\varepsilon(x_i)$；

(2)$x_j\in\delta_\varepsilon(x_i)\Rightarrow x_i\in\delta_\varepsilon(x_j)$；

(3)$\bigcup_{i=1}^{n}\delta_\varepsilon(x_i)=U$。

**定义 3** 设 $S=(U,A,V,f)$是一个数值型信息系统，其中 $U$：对象的非空有限集合，称为论域；$A$：属性的非空有限集合，$A=C\cup D,C\cap D=\varnothing,C$ 为条件属性，$D$ 为决策属性；$V=\bigcup_{a\in A}V_a,V\subset R,V_a$ 是属性 $a$ 的值域；$f:U\times A\to V$ 是一个信息函数，它为每个对象的每个属性赋予一个信息值，即 $\forall a\in A,x\in U,f(x,a)\in V_a$。

**定义 4** 设 $S=(U,A,V,f)$是一个数值型信息系统，$P\subseteq A$，则 $U$ 关于属性集 $P$ 的距离矩阵 $M_{d_P}=(d_P(x_i,x_j))$是一个 $|U|\times|U|$ 的矩阵，其中任一元素为

$$d_P(x_i,x_j)=\left(\sum_{i=1}^{|P|}|f(x_i,a_i)-f(x_j,a_i)|^\lambda\right)^{1/\lambda}$$

其中 $x_i,x_j\in U,\lambda=1,2,\infty$，在二维实数空间内，基于 1 范数，2 范数和无穷范数的邻域分别对应菱形、圆和正方形区域。

设 $D_{\max}=\max\{d_P(x_i,x_j)\}$为距离矩阵 $M_{d_P}$ 中的最大值，将距离矩阵 $M_{d_P}$ 进行归一化处理，记为

$$M'_{dP}=(d_P(x_i,x_j))/D_{\max}=(d'_P(x_i,x_j)),$$

其中 $d'_P(x_i,x_j)$为矩阵 $M'_{d_P}$ 中的任一元素。

**定义 5** 设 $S=(U,A,V,f)$是一个数值型信息系统，$P\subseteq A,\varepsilon\geqslant 0$，则 $\forall x_i\in U$，定义其 $\varepsilon$ 邻域为

$$\delta_P^\varepsilon(x_i)=\{x\mid x\in U,d_P(x,x_i)\leqslant\varepsilon\}$$

则 $U$ 关于属性集 $P$ 的 $\varepsilon$ 邻域矩阵 $M_{d_P^\varepsilon}=(d_P^\varepsilon(x_i,x_j))$中任一元素为

$$d_P^\varepsilon(x_i,x_j)=\begin{cases}1 & \text{if } d'_P(x_i,x_j)<\varepsilon\\ 0 & \text{if } d'_P(x_i,x_j)\geqslant\varepsilon\end{cases}$$

**定义 6** 设 $S=(U,A,V,f)$是一个数值型信息系统，$X\subseteq U,P\subseteq A,\varepsilon\geqslant 0$，则 $X$ 关于属性集 $P$ 的下近似、上近似和近似精度分别定义为

$$\underline{P_\varepsilon}X=\{x_i\mid\delta_P^\varepsilon(x_i)\subseteq X,x_i\in U\},$$

$$\overline{P_\varepsilon}X=\{x_i\mid\delta_P^\varepsilon(x_i)\cap X\neq\varnothing,x_i\in U\},$$

$$\alpha_{P_\varepsilon}(X)=\frac{|\underline{P_\varepsilon}X|}{|\overline{P_\varepsilon}X|},$$

其中 $0\leqslant\alpha_{P_\varepsilon}(X)\leqslant 1$。

**定义 7** 设 $S=(U,A,V,f)$是一个数值型信息系统，$P\subseteq A,\varepsilon\geqslant 0$，则 $U$ 关于属性集 $P$ 的 $\varepsilon$ 下近似矩阵$\underline{M_{d_P^\varepsilon}}=(\underline{d_P^\varepsilon}(x_i,x_j))$中任一元素为

$$\underline{d_P^\varepsilon}(x_i,x_j)=\begin{cases}1 & \text{if } \delta_P^\varepsilon(x_i)\subseteq\delta_P^\varepsilon(x_j)\\ 0 & \text{otherwise}\end{cases},$$

则 $U$ 关于属性集 $P$ 的 $\varepsilon$ 上近似矩阵$\overline{M_{d_P^\varepsilon}}=(\overline{d_P^\varepsilon}(x_i,x_j))$中任一元素为

$$\overline{d_P^\varepsilon}(x_i,x_j)=\begin{cases}1 & \text{if } \delta_P^\varepsilon(x_i)\cap\delta_P^\varepsilon(x_j)\neq\varnothing\\ 0 & \text{otherwise}\end{cases}$$

**定义 8** 设 $S=(U,A,V,f)$是一个数值型信息系统，$x_i\in U,P\subseteq A,\varepsilon\geqslant 0$，则 $\delta_P^\varepsilon(x_i)$关于属性集 $P$ 的耦合度定义为

$$\beta_P^\varepsilon(x_i)=\frac{|\underline{P_\varepsilon}(\delta_P^\varepsilon(x_i))|}{|\overline{P_\varepsilon}(\delta_P^\varepsilon(x_i))|},$$

其中 $0<\beta_P^\varepsilon(x_i)\leqslant 1$，如果 $\beta_P^\varepsilon(x_i)$越大，则 $x_i$ 的 $\varepsilon$ 邻域的耦合度越大。如果 $\varepsilon=0$，则 $\forall x_i\in U$，都有 $\beta_P^\varepsilon(x_i)=1$，$\beta_P^\varepsilon(x_i)$的计算表达式也可表示为

$$\beta_P^\varepsilon(x_i)=\frac{\sum_{j=1}^{|U|}\underline{d_P^\varepsilon}(x_i,x_j)}{\sum_{j=1}^{|U|}\overline{d_P^\varepsilon}(x_i,x_j)}$$

**定义 9** 设 $S=(U,A,V,f)$是一个数值型信息系统，$\forall x_i,x_j\in U,P\subseteq A,\varepsilon\geqslant 0$，定义 $\delta_P^\varepsilon(x_i)$和 $\delta_P^\varepsilon(x_j)$的分离度为

$$Div(\delta_P^\varepsilon(x_i),\delta_P^\varepsilon(x_j))=\frac{|\delta_P^\varepsilon(x_i)\cap\delta_P^\varepsilon(x_j)|}{|\delta_P^\varepsilon(x_i)\cup\delta_P^\varepsilon(x_j)|}$$

且有 $0\leqslant Div(\delta_P^\varepsilon(x_i),\delta_P^\varepsilon(x_j))\leqslant 1$，如果 $Div(\delta_P^\varepsilon(x_i),\delta_P^\varepsilon(x_j))$越小，则 $x_i,x_j$ 邻域中对象的分离程度越大。如果 $\varepsilon=0$，则 $\forall x_i\in U$，有 $Div(\delta_P^\varepsilon(x_i),\delta_P^\varepsilon(x_j))=0$。设 $\theta$ 为分离度阈值($0\leqslant\theta\leqslant 1$)，如果 $Div(\delta_P^\varepsilon(x_i),\delta_P^\varepsilon(x_j))\geqslant\theta$，则认为 $x_i,x_j$ 属于同一个类内，否则属于两个类。

例 1 设 $S=(U,A,V,f)$是一个数值型数据的信息系统，$U=\{x_1,x_2,x_3,x_4,x_5\},a\in A,f(x,a)$表示对象 $x$ 在属性 $a$ 上的取值，其中 $f(x_1,a)=1.1,f(x_2,a)=1.2,f(x_3,a)=1.6,f(x_4,a)=1.8,f(x_5,a)=1.9$，当指定 $\varepsilon=0.2$ 时，则 $x_1,x_2,x_3,x_4,x_5$ 对应的邻域分别为

$$\delta_{\{a\}}^{0.2}\{x_1\}=\{x_1,x_2\},\delta_{\{a\}}^{0.2}\{x_2\}=\{x_1,x_2\},\delta_{\{a\}}^{0.2}\{x_3\}=\{x_3,x_4\},\delta_{\{a\}}^{0.2}\{x_4\}=\{x_3,x_4,x_5\},\delta_{\{a\}}^{0.2}\{x_5\}=\{x_4,x_5\},$$

则 $x_1,x_2,x_3,x_4,x_5$ 邻域对应的下近似和上近似分别为

$$\underline{\{a\}_{0.2}}(x_1)=\{x_1,x_2\},$$

$$\underline{\{a\}_{0.2}}(x_2)=\{x_1,x_2\},$$

$$\underline{\{a\}_{0.2}}(x_3)=\{x_3\},$$

$$\underline{\{a\}_{0.2}}(x_4)=\{x_3,x_4,x_5\},$$

$$\underline{\{a\}_{0.2}}(x_5)=\{x_5\},$$

$$\overline{\{a\}_{0.2}}(x_1)=\{x_1,x_2\},$$

$$\overline{\{a\}_{0.2}}(x_2)=\{x_1,x_2\},$$

$$\overline{\{a\}_{0.2}}(x_3)=\{x_3,x_4,x_5\},$$

$$\overline{\{a\}_{0.2}}(x_4)=\{x_3,x_4,x_5\},$$

$$\overline{\{a\}_{0.2}}(x_5)=\{x_3,x_4,x_5\},$$

$x_1,x_2,x_3,x_4,x_5$ 邻域对应的耦合度分别为

$$\beta_{\{a\}}^{0.2}(x_1)=1,$$

$$\beta_{\{a\}}^{0.2}(x_2)=1,$$

$$\beta_{\{a\}}^{0.2}(x_3)=\frac{1}{3},$$

$$\beta_{\{a\}}^{0.2}(x_4)=1,$$

$$\beta_{\{a\}}^{0.2}(x_5)=\frac{1}{3},$$

则有 $\beta_{\{a\}}^{0.2}(x_1)=\beta_{\{a\}}^{0.2}(x_2)=\beta_{\{a\}}^{0.2}(x_3)>\beta_{\{a\}}^{0.2}(x_4)=\beta_{\{a\}}^{0.2}(x_5)$，

则 $x_1$ 作为第一个初始聚类中心，而 $Div(\delta_{\{a\}}^{0.2}(x_1),\delta_{\{a\}}^{0.2}(x_2))=1$，所以 $x_2$ 不能作为第二个中心，又因为 $Div(\delta_{\{a\}}^{0.2}(x_1),\delta_{\{a\}}^{0.2}(x_3))=0$，所以 $x_3$ 为第二个中心，假设分为 2 类，则聚类结果为$\{x_1,x_2\}$和$\{x_3,x_4,x_5\}$。

## 2 初始聚类中心选取算法(Initial Cluster Centers Choice Algorithm, ICCCA)

输入:$S=(U,A,V,f)$,$P\subseteq A$,聚类个数 $k$,范数 $\lambda$,分离度阈值 $\theta$;

输出:$k$ 个初始聚类中心 $Centers$。

步骤 1 初始化 $Centers=\emptyset$,生成 $U$ 关于属性集 $P$ 的距离矩阵 $M_{d_P}=(d_P(x_i,x_j))$ 和归一化矩阵 $M'_{d_P}=(d_P(x_i,x_j))/D_{\max}=(d'_P(x_i,x_j))$,并计算所有对象之间距离的平均值 $\bar{u}$;

步骤 2 在 $[0,\bar{u}]$ 之间选择 $\varepsilon$,生成邻域矩阵 $M_{d_P^\varepsilon}=(d_P^\varepsilon(x_i,x_j))$;

步骤 3 生成 $U$ 中所有对象 $x_i$ 的 $\delta_P^\varepsilon(x_i)$ 的下近似矩阵 $\underline{M_{d_P^\varepsilon}}=(\underline{d_P^\varepsilon}(x_i,x_j))$ 和上近似矩阵 $\overline{M_{d_P^\varepsilon}}=(\overline{d_P^\varepsilon}(x_i,x_j))$,并求出 $\beta_P^\varepsilon(x_i)$;

步骤 4 对 $\beta_P^\varepsilon(x_i)$ 按照由高到低排序,设排序结果为 $x_1'\geqslant x_2'\geqslant\cdots\geqslant x_{|U|}'$,记 $Order=\{x_1',x_2',\cdots,x_{|U|}'\}$;

步骤 5 $x_1'$ 即为第一个初始中心,$Centers=Centers\cup\{x_1'\}$;依次取 $Order[j]$,$2\leqslant j\leqslant|U|$,对 $\forall i$,$1\leqslant i\leqslant|Centers|$,都有 $Div(Centers[i],Order[j])<\theta$ 成立,则 $Order[j]$ 为一个新的初始中心。循环上述操作,当 $|Centers|=k$,算法终止。如果不能选出 $k$ 个初始中心点,则减小 $\varepsilon$ 的取值,转步骤 2。

## 3 评估方法

对于一个给定的聚类,如果数据集中包含 $C$ 个类,设 $a_i$ 表示正确分到 $C_i$ 类中的对象数,$b_i$ 表示错误分到 $C_i$ 类中的对象数,$c_i$ 表示排除在 $C_i$ 类中的对象数,则第 $i$ 个类的精度和召回率分别定义为:

$$p_i=a_i/(a_i+b_i)\quad 1\leqslant i\leqslant C,$$

$$r_i=a_i/(a_i+c_i)\quad 1\leqslant i\leqslant C,$$

聚类算法的精度和误分率分别定义为[18]

$$micro\_p=micro\_r=\left(\sum_{i=1}^{C}a_i\right)/n,$$

$$Error=\left(\sum_{i=1}^{C}b_i\right)/n,$$

其中 $n$ 是数据集的对象数,同时有 $\sum_{i=1}^{C}(a_i+b_i)=\sum_{i=1}^{C}(a_i+c_i)=n$。

## 4 实验分析

为了验证该方法的有效性,我们从 UCI 数据集中挑选了 3 组数据,其中 Letter Image Recognition 数据集是从 20000 条记录中的前 16000 条中选出字母为 A 类和字母为 D 类的对象,其中字母为 A 类的对象数有 789,字母为 D 类的对象数有 805,3 组数据描述如表 1 所示。

表 1 数据描述

| | Data Set | Samples | I 类 | II 类 | IV 类 |
|---|---|---|---|---|---|
| 1 | Wine Recognition | 178 | 59 | 71 | 48 |
| 2 | Fisher's Iris | 150 | 50 | 50 | 50 |
| 3 | Letter Image Recognition | 1589 | 789 | 805 | 0 |

在计算各对象的邻域时,为了减少因各属性量纲不一致对结果的影响,我们将所有对象之间的距离都标准化到[0,1]区间,同时可以得到所有对象之间的平均距离 $\bar{u}$。邻域 $\varepsilon$ 的取值是一个非常重要的问题,它决定了邻域的大小,如果邻域太小,则没有其他对象包含在邻域内,如果邻域太大,则可能导致对象之间的分离度降低,不利于聚类。在通常情况下,$\varepsilon$ 取值在 $[0,\bar{u}]$ 之间效果比较好。在三种不同的数据集上取 $\varepsilon=0.1$,$\lambda=2$,$\theta=0.5$,我们分别比较了本文提出的算法、CCIA[12] 算法以及随机算法确定初始聚类中心下的 K-means 聚类算法的误分率,其中随机误分率是 10 次随机聚类结果的平均值,分别如表 2、表 3 和表 4。

表 2 Wine Recognition Data 在三种不同初始聚类中心选择算法下 K-means 聚类算法的误分率

| 实际类别数目 | ICCCA 算法分类结果 I | II | III | ICCCA 误分率 | CCIA 误分率 | 随机误分率 |
|---|---|---|---|---|---|---|
| 59(I) | 59 | 0 | 0 | | | |
| 71(II) | 4 | 64 | 3 | 3.93% | 5.05% | 5.51% |
| 48(III) | 0 | 0 | 48 | | | |
| | 63 | 64 | 51 | | | |

表 3 Fisher's Iris Data 在三种不同初始聚类中心选择算法下 K-means 聚类算法的误分率

| 实际类别数目 | ICCCA 算法分类结果 I | II | III | ICCCA 误分率 | CCIA 误分率 | 随机误分率 |
|---|---|---|---|---|---|---|
| 50(I) | 50 | 0 | 0 | | | |
| 50(II) | 0 | 48 | 2 | 10.67% | 11.33% | 18.13% |
| 50(III) | 0 | 14 | 36 | | | |
| | 50 | 62 | 38 | | | |

表 4 Letter Image Recognition Data 在三种不同初始聚类中心选择算法下 K-means 聚类算法的误分率

| 实际类别数目 | ICCCA 算法分类结果 A | D | ICCCA 误分率 | CCIA 误分率 | 随机误分率 |
|---|---|---|---|---|---|
| 789(A) | 690 | 99 | | | |
| 805(D) | 27 | 778 | 7.90% | 8.55% | 9.26% |
| | 717 | 877 | | | |

由于在二维实数空间内,基于 1 范数、2 范数和无穷范数分别对应菱形、圆和正方形区域,图 1,2,3 分别展示了在三个不同的数据集上 $\varepsilon$ 在 $[0,\bar{u}]$ 且步长为 0.05 所对应的三种范式对聚类精度的影响。

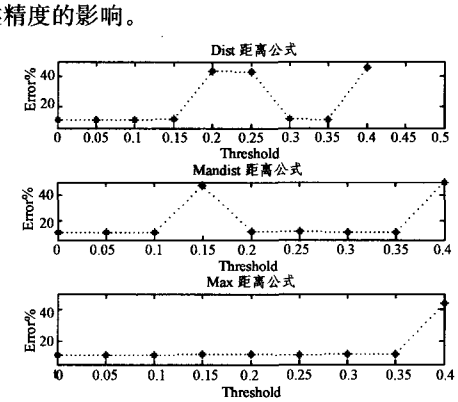


图 1 Fisher's Iris Data 在 $\varepsilon$ 在 $[0,0.4]$ 之间,步长为 0.05,在三种不同范式下对应的误分率(其中距离的平均值为 $\bar{u}=0.4757$)

根据图 1、图 2 和图 3 我们可以看出,相同的数据集在不同的范数下,聚类精度是不同的,所以我们可以根据数据分布的不同,选择不同的范数,使聚类精度达到最优。同时邻域 $\varepsilon$ 的取值在 $[0.05,0.1]$ 之间的效果比较好。

**结束语** 本文基于度量空间的邻域模型,通过对象邻域的耦合度和分离度描述了对象在选择初始聚类中心过程中的重要性,提出了一种初始聚类中心确定的方法,给出了邻域 $\varepsilon$ 的取值范围,分析了邻域模型中的三种范数和邻域 $\varepsilon$ 对聚类

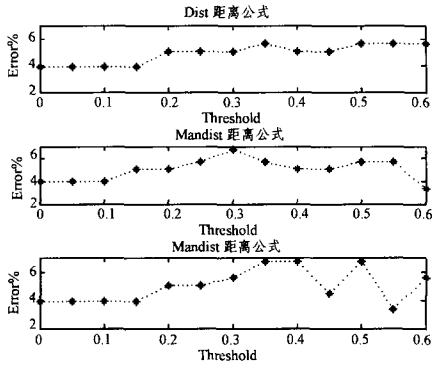


图 2 Wine Recognition Data 在 $\varepsilon$ 在[0,0.6]之间,步长为 0.05,在三种不同范下式对应的误分率(其中距离的平均值为 $\bar{u}=0.6194$)

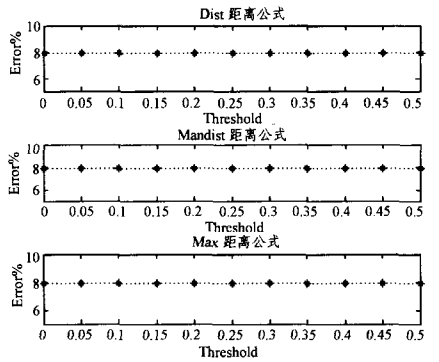


图 3 Letter Image Recognition Data 在 $\varepsilon$ 在[0,0.5]之间,步长为 0.05,在三种不同范式下对应的误分率(其中距离的平均值为 $\bar{u}=0.5671$)

精度的影响。实验分析表明,基于邻域模型的初始聚类中心选择算法优于随机选择初始聚类中心和 CCIA 选择初始聚类中心算法,相对而言参数 $\varepsilon$ 较小的时候聚类精度较高,效果较为理想。

## 参 考 文 献


[1] Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco, US: Morgan Kaufmann, 2001

[2] 张讲社,梁怡,徐宗本. 基于视觉系统的聚类算法. 计算机学报,2001,24(5):496-501

[3] 金阳,左万利. 一种基于动态近邻选择模型的聚类算法. 计算机学报,2007,30(5):756-762

[4] 张敏,于剑. 基于划分的聚类模型. 软件学报,2004,15(6):858-868

[5] 行小帅,潘进,焦李成. 基于免疫规划的 K-means 聚类算法. 计算机学报,2003,26(5):605-610

[6] Mac Q J. Some methods for classification and analysis of multivariate observation // Proceeding 5th Berkley Symposium. On Mathematical Statistics and Probability, 1967, I: 281-297 University of California Press. 1967, Xvii: 666

[7] Huang Z X. Extensions to the k-Means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2: 283-304

[8] Ahmad A, Dey L. A K-means clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering, 2007, 63: 503-527

[9] Duda R O, Hart P E. Pattern Classification and Scene Analysis. John Wiley and Sons. NY, 1973

[10] Bradley P S, Mangasarian O L, Street W N. Clustering via concave minimization. In: M. C. Mozer, M. I. Jordan, T. Petsche, Eds. Advances in Neural Information Processing System, MIT Press, 1997, 9: 368-374

[11] Penā J M, Lozano J A, Larrañaga P. An empirical comparsion of four initalization methods for the K-means algorithm. Pattern Recognition Letter, 1999(20): 1027-1040

[12] Khan S S, Ahmad A. Cluster center initialization algorithm for K-means clustering. Patter Recognition Letters, 2004, 25: 1293-1302

[13] Lin T Y. Granular Computing on binary relations I: data mining and neighborhood systems. In: rough sets in knowledge discovery, Skoworn A and Pokowshi L(eds) Physica-Verlag, 1998: 107-121

[14] Yao Y Y. Relational interpretation of neighborhood operators and neighborhood systems. Information Sciences, 1998, 111 (198): 239-259

[15] Wu W Z, Zhang W X. Neighborhood operator systems and approximations. Information Sciences, 2002, 144(1/4): 201-217

[16] Hu Q H, Yu D R, Xie Z X. Neighborhood classifiers. Expert Systems with Applications, 2007(in Press)

[17] Meila M, Heckerman D. An experimental comparison of several clustering methods. Microsoft Research Report MSR-TR-98-06. Redomnd, WA, 1998

[18] Yang Y M. An evaluation of statistical approaches to text categorization. Journal of Information Retrieval, 1999, 1 (1/2): 67-88


(上接第 77 页)

有较好的应用价值。

## 参 考 文 献


[1] Gritzalis S, Spinellis D, Georgiadis P. Security protocols over open networks and distributed systems: formal methods for theiranalysis, design, and verification. Computer Communications, 1999, 22(8): 695-707

[2] Cervesato I, Durgin N A, Lincoln P D, et al. Relating strands and multiset rewriting for security protocol analysis // Proceedings of the 13th IEEE Computer Security Foundations Workshop. Cambridge, England, 2000: 35-52

[3] Thomas Y C W, Simon S L. A semantic model for authentication protocols // Proceedings of the 14th IEEE Symposium on Research in Security and Privacy. Oakland: IEEE Computer Society Press, 1993: 178-194

[4] Woo T Y C, Lam S S. A semantic model for authentication protocols // Proceedings of the IEEE Symposium on Research in Security and Privacy. Oakland, CA, 1993: 178-194

[5] Thayer F, Herzog J C, Guttman J D. Strand space: why is a security protocol correct // Proceedings of the 1998 IEEE Symposium on Security and Privacy. 1998: 160-171

[6] 范红,冯登国. 安全协议理论与方法. 北京:科学出版社,2003

[7] 卿斯汉. 安全协议 20 年研究进展. 软件学报,2003,14(10): 1740-1752

[8] Schneier B. Applied Cryptography 2nd Edition. New York: John Wileysons, 1996

[9] Halpern J Y, Fagin R. Modelling konwledge and action in distributed systems. Distributed Computing, 1989, 3(4): 159-179

[10] Marrero W, Clarke E, Jha S. Verifying security protocols with Brutus. ACM Transactions on Software Engineering and Methodology, 2000, 9(4): 443-487

[11] Stoller S D. A bound on attacks on payment protocols // Proceedings of the 16th Annual IEEE Symposium on Logic in Computer Science(LICS). Boston, Massachusetts, 2001: 61-70