文章编号: 1673- 9469(2010) 04- 0098- 03

未确知均值聚类

庞彦军, 刘立民, 刘开第
(河北工程大学 理学院, 河北 邯郸 056038)

摘要: 利用未确知系统理论分析特征对样本分类所作贡献, 定义特征的分类权重, 并作为启发性知识用于确定样本与各类间的加权距离及样本属于各类的隶属度, 建立未确知均值聚类算法。IRIS 数据检验表明, 未确知均值聚类算法误判样本数少、收敛速度快、鲁棒性好, 是一种实用、有效的无监督聚类算法。

关键词: 均值聚类; 分类权重; 未确知系统; 隶属度
中图分类号: F224 文献标识码: A

Uncertain means clustering

PANG Yan-jun, LIU Li-min, LIU Kai-di
(School of Science, Hebei University of Engineering, Hebei Handan, 056038, China)

Abstract: In this paper, we firstly defined the classified weight according to the contribution of feature to sample which was used to determined the membership degree of sample to each class and the weighed distance between sample and each class, and then we proposed a new algorithm —the uncertain means clustering. The data of IRIS indicates that the algorithm possesses the better convergence, better robustness and it is an unsupervised clustering algorithm.

Key words: means clustering; classified weight; unascertained system; membership degree

聚类分析[1-2]是多元统计分析的重要方法, 是模式识别的重要工具, 在自动控制、系统辨识、人工智能、故障诊断等领域有重要的应用。基于迭代的动态聚类是最常用的聚类方法。C- 均值聚类[3-4]是一种确定性聚类, 是误差平方和最小意义下的最优聚类, 当存在病态数据和分类不清数据时, 聚类效果不能令人满意。模糊 C 均值聚类[5-6]则将隶属函数引入均值聚类, 能很好的处理分类不清数据, 但当样本存在"野值"时, 效果不是很好。改进的模糊 C 均值聚类[7]等虽解决了"野值"问题, 但迭代算法失去了可解释性。更重要的是, 上述聚类算法没有充分利用输入数据提供的分类信息, 没有体现出不同分类特征对分类作出的"不同贡献"。样本点之所以能被划分为不同类别, 是由于不同样本的同一特征观测值不同。不同样本的某个特征观测值越接近, 则该特征对区分开样本类别做出的贡献越小。样本集关于同一特征取值集中与发散的程度反映了该特征对分类贡献的大小, 这是与分类"同时存在"的客观事实。本文分析特征对样本分类所作贡献, 定义特征分类权重, 给出计算样本关于各类隶属度的迭代算法, 建立一种新的聚类方法—未确知均值聚类。

1 未确知系统理论[8]

未确知性是指由于条件限制, 决策者无法确定事物的真实状态或真实的数量关系, 因而产生的一种主观的、认识上的不确定性。对未确知性的定量描述和处理, 是对人类主观事物范畴的一种探索。

定义 1 设论域 $U=\{x_1, x_2, \cdots, x_n\}$, F 是 U 上的性质空间, E 是 F 上的 σ- 代数, 称 (F, E) 为 U 上的可测空间。

收稿日期: 2010- 10- 10
基金项目: 国家自然科学基金资助(60874116; 60940036); 河北省自然科学基金资助(F2009000857) 特约专稿
作者简介: 庞彦军(1964-), 男, 河北武安人, 教授, 从事不确定信息数学处理方面的研究。

定义 2 如果$\{F_1, F_2, \cdots, F_k\}$满足

$$F = \bigcup_{i=1}^{k} F_i,\ F_i \cap F_j = \phi (i \neq j) \tag{1}$$

令

$$E = \{E_i \mid E_i = \bigcup_{l=1}^{i} G_l,\ G_l \in \{\phi, F_1, F_2, \cdots, F_k\},\ 1 \leq i \leq k\} \tag{2}$$

则称$\{F_1, F_2, \cdots, F_k\}$是 F 的一种有限划分, E 是 F 上的σ代数。

定义 3 设(F, E)为 U 上的可测空间, $\mu_A(x)$为元素 x 具有性质A 的程度, 如果对 $\forall A_l \in E$, $x \in U$, 有

$$0 \leq \mu_A(x) \leq 1 \tag{3}$$

$$\mu_{\cup A_i}(x) = \sum \mu_{A_i}(x),\ (A_i \cap A_j = \phi, i \neq j) \tag{4}$$

$$\mu_F(x) = 1 \tag{5}$$

则称 $\mu_A(x)$ 为可测空间(F, E)上的测度函数, $(U, E, \mu_A(x))$ 为未确知测度空间。

定义 4 设$(U, E, \mu_A(x))$是未确知测度空间, 则以 $\mu_A(x)$ 为隶属函数确定了论域 U 上关于σ代数E 的一个未确知子集G

$$G = \{\frac{\mu_A(x)}{x} \mid \forall x \in U, 0 \leq \mu_A(x) \leq 1, A \in E\} \tag{6}$$

当$A \in E$ 固定时, 以 $\mu_A(x)$ 为隶属函数确定了论域 U 上的一个未确知子集; 当 $x \in U$ 固定时, 以 $\mu_A(x)$为隶属函数确定了 σ代数E 上的一个未确知子集。所以, $\mu_A(x)$ 是 $U \times E$ 上的二元函数。

2 未确知均值聚类算法

2.1 问题描述

已知 d 维特征空间的N 个训练样本$x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ $(i = 1, 2, \cdots, N)$, 欲将 N 个样本划分为 C 类: $\Gamma_1, \Gamma_2, \cdots, \Gamma_C$。确定出 Γ_i 类的类中心m_i, 则可用最小距离准则确定各样本点及待识样本点的类别。

2.2 基本假设

假设同一类中的样本点在特征空间中彼此应该更“接近”, 并且这种“接近”是欧氏距离或加权欧氏距离意义下的接近[9], 即认为同类样本点在空间呈现超球体分布。如果这种“接近”是指在某个方向上的接近, 将对应“距离”的不同表达方法。

2.3 启发性知识获取

设 $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ 的分量是标称化数据。为了定量描述 d 个特征对初始分类做出的贡献, 令

$$\overline{m} = \frac{1}{C} \sum_{i=1}^{C} m_i = (\overline{m}_1, \overline{m}_2, \cdots, \overline{m}_d) \tag{7}$$

$$\sigma_j^2 = \frac{1}{C} \sum_{k=1}^{C} (m_{kj} - \overline{m}_j)^2,\ 1 \leq j \leq d \tag{8}$$

方差 σ_j^2 的大小反映了 C 个类中心$m_1, m_2, \cdots, m_C$ 在第j 个特征上取值的离散程度。

若 $\sigma_j^2 = 0$, 则 C 个类中心的第j 个向量全部相同, 这时, j 特征对于把C 个类中心“区分开”没起任何作用。亦即j 特征对“区分开” C 个分类的贡献值为零。σ_j^2 越大, 表明 C 个类中心的第j 个分量越分散, j 特征对于把C 个类中心“区分开”所作的贡献“越大”。所以, σ_j^2 的大小反映了特征j 对样本分类所做贡献的大小。

令

$$\omega_j = \sigma_j^2 / \sum_{k=1}^{d} \sigma_k^2 \tag{9}$$

称 ω_j 为特征j 关于给定分类的分类权重。特征分类权重是在给定某种分类下, 特征对“区分开”各类所做“贡献”在所有特征中所占的比例。

2.4 隶属度计算

初始分类给出 C 个聚类中心$m_1, m_2, \cdots, m_C$, 任一训练样本 x_i 关于以m_k 为类中心的Γ_k 类有一个实际上的隶属度 μ_{ik}。显然, μ_{ik}与点x_i 到m_k 的距离及各特征的分类权重有关。当 $\omega_j = 0$ 时, j 特征对分类不起作用, 这时分量$(x_{ij} - m_{kj})^2$ 不应作为距离分量出现在表征 x_i 到m_k 的距离中; 而 ω_j 越大, j 特征对分类贡献越大。所以, 当用 x_i 到m_k 间的距离D_{ik}去表征x_i 关于 Γ_k 类隶属度时, 这种“距离”应是一种加权距离。当样本点 x_i 到类中心m_k 的加权距离越小时, x_i 属于 Γ_k 类的隶属度越大。故

$$D_{ik} = \| x_i - m_k \| = \sqrt{\sum_{j=1}^{d} \omega_j \cdot (x_{ij} - m_{kj})^2} \tag{10}$$

$$\mu_{ik} = (D_{ik} + \varepsilon)^{-1} \cdot \sum_{j=1}^{C} (D_{ik} + \varepsilon)^{-1},\ k = 1, 2, \cdots, C \tag{11}$$

3 未确知均值聚类迭代算法

对 Γ_k 类的类中心m_k 赋予质量 μ_{ik}, 令 Γ_k 类的新类中心向量$m_k^{(1)}$ 为

$$m_k^{(1)} = \sum_{i=1}^{N} \mu_{ik} \cdot x_i / \sum_{i=1}^{N} \mu_{ik},\ k = 1, 2, \cdots, C \tag{12}$$

以新类中心 $m_k^{(1)}$ 替代初始类中心向量, 可以建立求类中心的迭代算法。

步骤 1 对 N 个训练样本 x_i (i = 1, 2, …, N) 的观测数据实施标称化变换, 标称化后的无量纲数据记为 x_i = (x_{i1}, x_{i2}, …, x_{id}) ; 给定分类数 C。

步骤 2 对 N 个样本初始分类, 得到 C 个聚类的初始类中心 $m_1^{(0)}$, $m_2^{(0)}$, …, $m_C^{(0)}$。

步骤 3 由(7) 、(8) 、(9) 式, 得分类权重向量 $\omega^{(0)}$ = ($\omega_1^{(0)}$, $\omega_2^{(0)}$, …, $\omega_C^{(0)}$) 。

步骤 4 由式(10) 与式(11) , 得隶属度向量 μ_{i1}, μ_{i2}, …, μ_{iC}) (i = 1, 2, …, N) 。

步骤 5 由式(12) 得新聚类中心 $m_1^{(1)}$, $m_2^{(1)}$, …, $m_C^{(1)}$。

步骤 6 若 $\max_i \| m_i^{(1)} - m_i^{(0)} \| < \delta$, 其中 $\delta > 0$ 是预先给定得小正数, 则迭代停止, 所求的 C 个聚类中心为 $m_1^{(1)}$, $m_2^{(1)}$, …, $m_C^{(1)}$。

步骤 7 令 $m_i^{(0)} = m_i^{(1)}$ (i = 1, 2, …, C) , 返回步骤 3。

4 有效性检验

对 3 类共 150 个样本的 IRIS 数据, 采用密度法确定 3 个初始类中心, 结合本文算法经 10 次迭代后求出 3 个聚类中心, 然后对 150 个训练样本按“最小加权距离准则” 重新归类。经 15 次重复实验, 平均误识率为 1. 3%, 表明本文算法稳定、实用、鲁棒性较好。

5 结论

1) 未确知均值聚类根据样本关于各类隶属度与类中心间的内在联系, 直接用迭代法求聚类中心, 避开了构造准则函数, 使得算法的每一步涉及的类中心与隶属度具有物理的可解释性。

2) 未确知均值聚类充分利用了输入数据提供的关于分类的启发式信息, 构造的隶属度严格满足测量准则。

3) IRIS 数据检验表明, 未确知均值聚类算法较模糊均值聚类算法误判样本数少且收敛速度快, 是一种实用、有效的无监督聚类算法。

参考文献:

[1] MARQUES DE SA J P. 模式识别—原理、方法及应用[M]. 北京: 清华大学出版社, 2002.

[2] 顾洪博, 赵万平. 基于 MMD 聚类算法及在高校成绩分析中的应用[J]. 河北工程大学学报(自然科学版) , 2010, 27 (1) : 96- 98.

[3] 周巧萍, 潘晋孝, 杨 明. 基于核函数的混合 C 均值聚类算法[J]. 模糊系统与数学, 2008, 22(6) : 148- 151.

[4] 高新波, 裴继红, 谢维信. 模糊 C- 均值聚类算法中加权指数 m 的研究[J]. 电子学报, 2000, 28(4) : 80- 83.

[5] 刘蕊洁, 张金波, 刘 锐. 模糊 C 均值聚类算法[J]. 重庆工学院学报, 2008, 22(2) : 139- 141.

[6] 陈佳妮, 段文英, 丁 徽. 模糊 C- 均值聚类分析在基因表达数据分析中的应用[J]. 森林工程, 2010, 26 (2) : 54- 58.

[7] 刘坤朋, 罗 可. 改进的模糊 C 均值聚类算法[J]. 计算机工程与应用, 2009, 45(21) : 97- 98.

[8] 刘开第, 曹庆奎, 庞彦军. 基于未确知集合的故障诊断方法[J]. 自动化学报, 2004, 30(5) : 747- 756.

[9] 王 鑫, 颜 炎, 杨睿嫦, 等. 多批次测试数据建模新方法[J]. 黑龙江科技学院学报, 2010, 20(3) : 227- 229.

(责任编辑 马立)