文章编号: 1001-0920(2004)05-0595-03

模糊核聚类的自适应算法

李 侃, 刘玉树
(北京理工大学 计算机科学与工程系, 北京 100081)

摘 要: 针对模糊聚类算法在样本特征不明显时不能取得很好的聚类效果, 以及现有的模糊聚类算法需要事先确定聚类数, 随机性强、容易陷入局部最优等弱点, 将核函数和有效性函数引入到模糊聚类中, 提出了模糊核聚类的自适应算法. 此方法在性能上比经典的聚类算法有了较大的改进, 取得了更好的聚类效果. 实验结果证实了该方法的有效性和可行性.

关键词: 模糊 C 均值; Mercer 核; 特征空间; 有效性函数
中图分类号: TP18 **文献标识码**: A

Fuzzy kernel clustering self-adaptive algorithm

LI Kan, LIU Yu-shu
(Department of Computer Science and Engineering, Beijing Institute of Technology, Beijing 100081, China. Correspondent: LI Kan, E-mail: e-likan@sina.com)

Abstract: Kernel function and validity measure function are introduced to the fuzzy clustering algorithm, and fuzzy kernel clustering self-adaptive algorithm is proposed. The algorithm owns better performance than classical clustering algorithms. Experiment results show the feasibility and effectiveness of the fuzzy kernel clustering self-adaptive algorithm.

Key words: fuzzy C-means; Mercer kernel; feature space; validity measure function

1 引 言

聚类是一种无监督的学习过程. C-均值方法和模糊 C-均值方法[1,2]直接对样本特征进行聚类, 聚类的效果较大程度上取决于样本的分布情况. 这些方法不适合发现有非凸面形状的簇, 或大小差别很大的簇; 而且它们对孤立点数据敏感, 孤立点会极大影响聚类的最终结果. 对此, 本文将Mercer 核[3]引入到模糊 C-均值方法中, 将输入空间的数据隐式地映射到高维特征空间, 在特征空间中特征能被进一步地显示、提取, 从而更好地完成聚类. 此外, 在C-均值方法和模糊 C-均值等方法中, 聚类的数目需事先确定, 容易陷入局部最优[4], 随机性较强. 本文采用了自适应的算法, 利用聚类的有效性函数分析, 在聚类的过程中动态地调整聚类数目, 最终能自适应地确定聚类的数目. 仿真实验验证了方法的有效性和可行性.

2 输入空间聚类

基于距离的衡量标准是很多聚类算法的基础. C 均值和模糊 C- 均值都是以距离来判断聚类问题. 模糊 C- 均值在输入空间内执行. $X = \{x_i \mid i = 1, 2, \cdots, n\}$ 是样本集, k 是聚类数, 聚类中心为 $c_1, c_2, \cdots, c_k$. 目标函数定义为

$$J = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}^{\alpha} \| x_i - c_j \|^2,$$

收稿日期: 2003-03-20; **修回日期**: 2003-07-01.
基金项目: 国防预研项目(114050202).
作者简介: 李侃(1975—), 男, 辽宁大连人, 博士, 从事数据挖掘、人工智能等研究; 刘玉树(1941—), 男, 山东人, 教授, 博士生导师, 从事人工智能、计算机图形学等研究.

$$\sum_{j=1}^{k} u_{ij} = 1, 0 < \sum_{i=1}^{n} u_{ij} < n,$$

$$0 \leq u_{ij} \leq 1, i = 1, 2, \cdots, n, j = 1, 2, \cdots, k. \quad (1)$$

其中: $\alpha > 1$ 为权重指数, u_{ij} 为 X 集中第 j 个数据对第 i 个聚类中心的隶属度

当类间分类边界是非线性时, 采用模糊C-均值算法不能取得很好结果, 而且在样本集很大时, 它不能有效确定聚类数目, 易陷入局部最优. 解决此问题的方法是, 采用非线性映射将输入空间数据映射到高维特征空间. 文献[3]提出的使用核函数的方法可以很好地解决此问题, 同时本文提出的自适应算法可以动态解决聚类数目的问题.

3 特征空间聚类

本文方法是将输入空间的样本非线性地映射到高维特征空间. 采用高斯核函数, 其对应的特征空间是无穷维的, 因而有限样本在特征空间必是线性可分的, 从而可以取得很好的聚类效果.

3.1 模糊核聚类自适应算法(KFCMA)

本文提出了模糊核聚类的自适应方法, 从一个新的角度去完成聚类功能.

输入空间的样本被映射到特征空间 ($x \rightarrow \varphi(x)$). 特征空间聚类中心可表示为

$$c_j = \sum_{k=1}^{n} \gamma_{jk} \varphi(x_k), \quad (2)$$

则

$$J = \sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{\alpha} \| \varphi(x_i) - \sum_{k=1}^{n} \gamma_{jk} \varphi(x_k) \|^2 =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{\alpha} \Big[k(x_i, x_i) - 2\sum_{k=1}^{n} \gamma_{jk} k(x_i, x_k) +$$

$$\sum_{k=1}^{n}\sum_{l=1}^{n} \gamma_{jk}\gamma_{jl} k(x_k, x_l) \Big] =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{\alpha} [k(x_i, x_i) - 2\gamma_j^{\mathrm{T}} k_i + \gamma_j^{\mathrm{T}} k \gamma_j]. \quad (3)$$

其中

$$\gamma_i = (\gamma_{i1}, \gamma_{i2}, \cdots, \gamma_{in})^{\mathrm{T}},$$

$$k = (k_1, k_2, \cdots, k_n),$$

$$k_i = (k_{i1}, k_{i2}, \cdots, k_{in})^{\mathrm{T}}.$$

u_{ij} 和 γ_{jk} 分别表示为

$$u_{ij} = 1 \Big/ \sum_{g=1}^{c} \left(\frac{d_{ij}}{d_{ig}}\right)^{\frac{1}{\alpha-1}},$$

$$\gamma_j = \sum_{j=1}^{n} u_{ij}^{\alpha} k^{-1} k_j \Big/ \sum_{j=1}^{n} u_{ij}^{\alpha}. \quad (4)$$

其中

$$d_{ij} = k(x_i, x_i) - 2\gamma_j^{\mathrm{T}} k_i + \gamma_j^{\mathrm{T}} k \gamma_j.$$

模糊核聚类自适应算法(KFCMA)为:

Step 1: 初始化正的参数 α, ϵ, 设定迭代次数 $m = 1$; 设初值 γ_j, 令聚类数目 $c = 2$;

Step 2: 计算核矩阵, 这里采用的是高斯核 $k(x_i, x_j) = e^{-q\|x_i - x_j\|^2}$;

Step 3: 更新 $u_{ij}^{(m)}$, 重新计算 $\gamma_j^{(m)}$;

Step 4: 如果 $\max \left| u_{ij}^{(m)} - u_{ij}^{(m-1)} \right| \leq \epsilon$, 则迭代结束; 否则 $m = m + 1$, 转向 Step 3;

Step 5: 如果有效性函数达到了最小值, 聚类结束; 否则 $c = c + 1$, 转向 Step 3.

3.2 聚类有效性分析

文献[5]利用在核矩阵对角线表现出来的特征来决定聚类的数目, 并用于C-均值中. 其他学者也将该方法用于模糊C-均值中. 但当样本在核矩阵的对角线特征表现不明显时, 这种方法就不能取得较好的聚类结果, 因为它仅仅是依靠核矩阵进行对聚类数的判断. 本文则采用有效性函数来动态估计聚类数目. 聚类数目不是事先结定的. 首先设定一个聚类的初值, 然后利用有效性函数来动态调整聚类数, 完成聚类功能. 有效性函数是通过紧致性和分离性效果函数来评价类内样本的紧密性、类间样本的独立性. 一个好的聚类其聚类中心的间距应尽可能的大, 而样本与其聚类中心的距离尽可能的小.

紧致性效果函数定义为

$$\mathrm{comp} = \sum_{i=1}^{c}\sum_{j=1}^{n} \lambda_{ij} \frac{k(x_i, x_i) - 2\gamma_j^{\mathrm{T}} k_i + \gamma_j^{\mathrm{T}} k \gamma_j}{n u_{ij}^2}. \quad (5)$$

其中

$$\lambda_j = \begin{cases} 1, u_{ij} > u_{lj}, i \neq l; \\ 0, \text{others} \end{cases}$$

被用于判断孤立点. 如果 $\lambda_j = 0$, 则数据点是孤立点, 将被去掉. 因为孤立点极大影响着最终的聚类效果, 所以在紧致性效果函数中, 加入了对孤立点的判断. 在式(5)中引入 u_{ij}^2, 目的是加强聚类的紧致性.

分离性效果函数表示为

$$\mathrm{sep} = \min_{i,j} (\gamma_i^{\mathrm{T}} k \gamma_i - 2\gamma_i^{\mathrm{T}} k \gamma_j + \gamma_j^{\mathrm{T}} k \gamma_j). \quad (6)$$

模糊聚类的有效性函数定义为紧密性与分离性之比. 最小化有效性函数, 则模糊聚类的目标函数(3)也可达到最小. 有效性函数表示为

$$s = \frac{\mathrm{comp}}{\mathrm{sep}} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} \lambda_j \frac{k(x_i, x_i) - 2\gamma_j^{\mathrm{T}} k_i + \gamma_j^{\mathrm{T}} k \gamma_j}{n u_{ij}^2}}{\min_{i,j} (\gamma_i^{\mathrm{T}} k \gamma_i - 2\gamma_i^{\mathrm{T}} k \gamma_j + \gamma_j^{\mathrm{T}} k \gamma_j)}. \quad (7)$$

4　实验结果

实验数据采用的是 iris 数据. 在 iris 数据库中随机地选取 30 个数据, Versicolor, Virginica, Setosa 各选取 10 个数据. 为了显示聚类结果, 图 1 所示 FCM 算法(图 1) 得到的聚类结果聚类中心容易陷入到局部最优的状态, 同时数据位置也极大影响着聚类中心和聚类的最终结果. 而图 2 采用 KFCMA 算法, 聚类结果得到了明显的提高, 聚错率也极大减小, 且聚类数目是以自适应的方法得到.

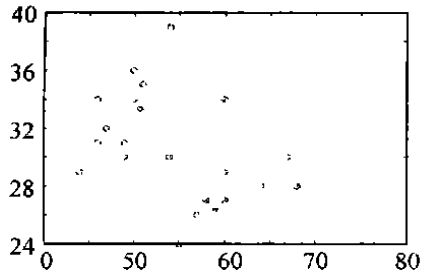

图 1　FCM 聚类

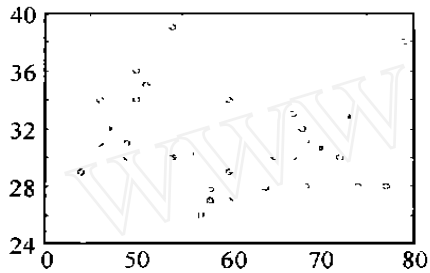

图 2　KFCM 聚类

5　结　　论

聚类算法的有效性依赖于样本间的特征. 当类间的特征不明显或有交迭时, 模糊聚类不能取得很好的聚类效果. 本文采用了 Mercer 核的模糊聚类自适应算法. 通过 Mercer 核, 把输入空间的样本映射到高维特征空间, 从而将各类样本的特征差别加大, 可以取得更好的聚类结果. 为了避免聚类陷入局部最优解、随机性强等弱点, 提出了将有效性函数作为动态调整聚类数的标准, 利用有效性函数来动态确定聚类的数目. 实验结果证明了模糊核聚类自适应算法比经典的模糊聚类算法的聚类效果好.

参考文献(References):

[1] MacQueen J. Some methods for classification and analysis of multivariate observations [A]. *Proc 5th Berkeley Symposium in Mathematics, Statistics, Probbability* [C]. California, 1967. 281-297.

[2] Bezdek JC. *Pattern Recognition with Fuzzy Objective Function Algorithms* [M]. New York: Plenum Press, 1981.

[3] Scholkopf B, Mika S, Burges C, et al. Input space versus feature space in kernel-based methods[J]. *IEEE Trans on Neural Networks*, 1999, 10(5): 1000-1017.

[4] Kamel S Mohamed. New algorithms for solving the fuzzy c-means clustering problem [J]. *Pattern Recognition*, 1994, 27(3): 421-428.

[5] Girolami M. Mercer kernel based clustering in feature space[J]. *IEEE Trans on Neural Network*, 2002, 13(3): 780-784.

下　期　要　目