

基于混合核函数的可能性 C-均值聚类算法*

杭欣, 李雷

(南京邮电大学自动化学院, 南京 210003)

摘要: 针对传统的模糊 C-均值算法对于非球形分布的数据聚类效果不理想且易受到噪声数据的影响, 利用可能性 C-均值算法具有良好的抗噪声性能, 将混合核函数引入到该算法中, 提出了一种基于混合核函数的可能性 C-均值(HKPCM)聚类算法。该算法将原空间的待分类样本映射到一个高维的特征空间(核空间)中, 使得样本变得线性可分, 然后在核空间中进行聚类。实验结果证实了 HKPCM 算法的可行性和有效性。

关键词: 聚类算法; 核函数; 模糊 C-均值算法; 可能性 C-均值算法

中图分类号: TP301 **文献标志码:** A **文章编号:** 1001-3695(2012)08-2852-02

doi:10.3969/j.issn.1001-3695.2012.08.013

Possibilistic C-means clustering algorithm based on hybrid kernel function

HANG Xin, LI Lei

(College of Automation, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)

Abstract: Traditional fuzzy C-means algorithm have a bad clustering result for non-spherical data and it is easy to be affected by the noise data. To solve these problems, this paper used the advantage of possibilistic C-means clustering algorithm and combined hybrid kernel function, proposed HKPCM algorithm. The samples in the original space were mapped into a high dimensional space by using this algorithm. So the samples became linearly separable and it was easy to cluster in kernel space. Experiment results indicate that the HKPCM algorithm is feasible and efficient.

Key words: clustering algorithm; kernel function; fuzzy C-means algorithm; possibilistic C-means algorithm

模糊 C-均值^[1]因算法设计简单、解决问题范围广、易于计算机实现, 被用于很多领域。它是一种非常有效的模糊聚类算法。对于大多数无噪声的球形分布的数据当用欧式距离来衡量点的相似性进行聚类时会得到很好的聚类结果。然而, 对于非球形分布的数据, 如环形数据, 就得不到理想的聚类效果, 这是 FCM 算法的主要缺点。为了克服这个缺点, 研究者提出了模糊核函数 C-均值(FKCM)算法^[2]。FKCM 算法首先将原空间的待分类样本映射到一个高维的特征空间(核空间)中, 使得样本变得线性可分, 然后在核空间中进行聚类。如果这个非线性映射是连续和光滑的, 那么在核空间中, 样本的拓扑结构和顺序将会被保持。因此, 在原空间中聚在一起的点, 在核空间中也将会聚在一起, 而且通过非线性映射, 突出了不同类别样本特征的差异, 使得原来线性不可分的样本点在核空间中变得线性可分。可以通过定义一定的相似度量标准和目标函数在核空间中进行聚类。这样, 对那些线性不可分样本以及不对称分布的样本进行聚类时, 利用核方法就可以减少聚类误差, 获得较好的聚类效果^[3-5]。

FCM 与 FKCM 算法的一个共同缺点是当远离聚类边缘的位置有噪声或在聚类之间存在过渡点时极易受到它们的影响。可能性 C-均值聚类(possibilistic C-means clustering, PCM)是由 Krishnapuram 等人^[6]于 1996 年提出的, 它对 FCM 进行了改进, 减弱了噪声的影响, 同时使得算法对初始值和参数的选择不敏感。本文利用 FKCM 和 PCM 算法的优点, 弱化它们的缺点, 在这两种算法的基础上设计了一种新的聚类算法——基于

混合核函数的可能性模糊 C-均值聚类算法。

1 模糊核函数 C-均值聚类

首先利用一个非线性映射 $\Phi: \mathbf{R}^n \rightarrow F, x \rightarrow \Phi(x)$ 将原空间中的样本 x 映射到一个高维的核空间 F 中, 目的是突出不同类别样本之间的特征差异, 使得样本在核空间中变得线性可分, 然后在这个高维的核空间中进行一般的 FCM 聚类。在核空间中, 待分类的样本变为 $(\Phi(x_1), \Phi(x_2), \dots, \Phi(x_n))$, 进行 FKCM 聚类就是最小化下面的聚类误差目标函数:

$$J(X; U, V) = \sum_{j=1}^C \sum_{i=1}^N u_{ij}^m d^2(x_i, v_j) \quad 2 \leq C \leq N \quad (1)$$

其中: $m > 1$ 是一个可以控制聚类结果的模糊程度的常数, 模糊划分矩阵 $U = [u_{ij}]_{C \times N}$, u_{ij} 为 x_j 对于类 i 的可能值; 聚类中心矩阵 $P = [v_{ij}]_{C \times N}$, v_i 为第 i 类的聚类中心矢量或聚类原型矢量; C 为聚类个数; N 为聚类数据的数量。约束条件为

$$\sum_{j=1}^C u_{ij} = 1, 0 < \sum_{i=1}^N u_{ij} < N, \forall j \quad (2)$$

d_{ij}^2 为样本点 x_j 与第 i 类聚类原型 v_i 之间的距离度量:

$$d_{ij}^2(x_i, v_j) = \|\Phi(x_i) - \Phi(v_j)\|^2 = K(x_i, x_i) - 2K(x_i, v_j) + K(v_j, v_j) \quad (3)$$

其中: $K(*, *)$ 就是核函数^[7-9]。使用不同的核得到不同的模型。在核学习方法中, 常用的核函数是满足条件 Mercer 的核函数。鉴于 FKCM 算法中主要的运算是进行核空间中样本之间的距离计算, 因此为了减少运算时间, 采用了一种新的可

收稿日期: 2011-12-31; 修回日期: 2012-02-10 基金项目: 国家自然科学基金资助项目(61070234, 61071167)

作者简介: 杭欣(1987-), 男, 江苏徐州人, 硕士研究生, 主要研究方向为数据挖掘、智能计算(hxhangxin@163.com); 李雷(1958-), 男, 教授, 博士, 主要研究方向为模式识别、智能系统、应用数学。

用于核学习的核函数条件正定核函数(conditionally positive definite kernel, CPD 核)。下面采用文献[1]中的一个条件正定核函数:

$$K_{\text{CPD}} = -\frac{1}{2} \|x - y\|^q + 1 \quad 0 < q \leq 2 \quad (4)$$

此函数用于核学习方法中,在核空间中的两个样本之间的距离运算就可以简化为

$$d^2(x_i, v_j) = 2(1 - K(x_i, v_j)) = \|x_i - v_j\|^q \quad (5)$$

在约束条件下求解优化问题 $\min_{(U, V)} \{J_m(U, V)\}$, 可得到:

$$u_{ij} = \frac{(1/d^2(x_i, v_j))^{1/(m-1)}}{\sum_{j=1}^c (1/d^2(x_i, v_j))^{1/(m-1)}} \quad (6)$$

聚类中心在高维空间用隶属度表示为

$$\Phi(v_j) = \sum_{i=1}^N (u_{ij})^m \Phi(x_i) / \sum_{i=1}^N (u_{ij})^m \quad (7)$$

所有匹配的点到各个类的初始隶属度用式(6)计算,匹配点到新的类的距离的更新用式(5)计算。

算法步骤如下:

a) 初始化隶属度

(a) 初始化 C 个聚类中心 v_k , 参数 m , 终止条件 ε ; 迭代 $k=0$;

(b) 计算初始隶属度 $u_{ij}(0)$ 。

b) 最小化目标函数

(a) 令 $k = k + 1$;

(b) 用式(5)计算 $d^2(x_i, v_j)$;

(c) 用式(6)计算隶属度 $u_{ij}(k)$;

(d) 终止条件 $\|u_{ij}(k) - u_{ij}(k-1)\| < \varepsilon$ 。

2 基于混合核函数的可能性C-均值聚类算法

可能性C-均值聚类对FCM进行了改进,减弱了噪声的影响,但是它对算法的初始值和参数的选择很敏感。于是2008年又提出了另一种形式的可能性C-均值聚类^[2]。然而PCM算法虽然能够克服噪声和初始值与参数选择敏感的缺点,但是对于各类样本的边界是线性不可分以及类分布为非Gauss分布或类分布为非椭圆分布时,其聚类效果不理想^[10]。本文提出的FKCM算法虽然能够克服此缺点但极易受到噪声的影响。因此,本文利用PCM与FKCM的优点提出了基于混合核函数的可能性C-均值聚类(HKPCM)算法。下面首先介绍一下可能性C-均值,然后引入混合核函数的概念。

可能性C-均值的目标函数为

$$J_m(U, P) = \sum_{i=1}^c \sum_{j=1}^n u_{ij} d_{ij}^2 + \sum_{i=1}^c \eta \sum_{j=1}^n (u_{ij} \log u_{ij} - u_{ij}) \quad (8)$$

约束条件: $0 \leq u_{ij} \leq 1, 0 < \sum_{j=1}^n u_{ij} \leq n$ 。 $U = [u_{ij}]_{c \times n}$ 是可能划分矩阵, u_{ij} 为 x_j 对于类 i 的可能性。 $P = [p_i]_{c \times n}$ 是聚类中心矩阵。

$$\eta = \sum_{j=1}^n u_{ij}^m d_{ij}^2 / \sum_{j=1}^n u_{ij}^m \quad (9)$$

在约束条件下求解优化问题 $\min_{(U, P)} \{J_m(U, P)\}$, 可得

$$u_{ij} = \exp(-d_{ij}^2 / \eta) \quad (10)$$

$$p_i = \sum_{j=1}^n u_{ij} x_j / \sum_{j=1}^n u_{ij} \quad (11)$$

在核学习方法中,核函数的选择是非常关键的问题。核函数的好坏不仅影响最终的聚类效果,而且影响算法的速度。KPCM算法中采用的核函数是一个CPD局部核函数,性能与Gauss径向基核函数差不多都是局部性强的核函数,其内推能力随着参数 σ 的增大而减弱。多项式形式的核函数有着良好

的全局性质,具有很强的外推能力,并且多项式的次数越低,外推能力越强。但每种核函数都各有利弊。要想得到学习能力和推广性都很强、兼具良好的外推和内推能力的核函数,就需要把两类核函数结合起来形成混合核函数: $K_{\text{mix}} = \lambda K_{\text{poly}} + (1 - \lambda) K_{\text{rbf}}$ 。这里还要确定最优混合系数 $\lambda \in (0, 1)$, 通过实验得出 λ 一般在 $0.50 \sim 0.99$, 而且当 λ 较大(如为 0.95)时,混合核函数的性能要优良一些,比单独使用多项式核函数或者RBF核函数的效果要好。

常用的核函数有:

a) 多项式核函数

$$K(x_i, x_j) = (x_i x_j + b)^d \quad (12)$$

b) Sigmoid 核函数

$$K(x_i, x_j) = \tanh(\alpha(x_i x_j) + \beta) \quad (13)$$

c) Gauss 核函数

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \quad (14)$$

本文构造如下组合核函数:

$$K(x_i, x_j) = \lambda (x_i x_j + 1)^d + (1 - \lambda) e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \quad (15)$$

算法步骤如下:

a) 初始化隶属度

(a) 初始化 C 个聚类中心 v_k , 参数 m, σ^2 , 终止条件 ε ; 迭代 $k=0$;

(b) 计算初始隶属度 $u_{ij}(0)$ 。

b) 最小化目标函数

(a) 令 $k = k + 1$;

(b) 用式(3)(15)计算 $d^2(x_i, v_j)$;

(c) 用式(9)计算 η ;

(d) 用式(10)计算隶属度 u_{ij} ;

(e) 终止条件 $\|u_{ij}(k) - u_{ij}(k-1)\| < \varepsilon$ 。

3 实验结果

本章将给出一些例子来说明本文提出的聚类方法的实验结果与FCM、FKCM算法的对比,例子包括环形分布的数据与加噪声的环形分布的数据。

3.1 环形数据实验

本例中聚类结果为一个环形分布的数据,数据由一个中心是一个类周围有一个环形类包含的数据组成。数据分为两类分别用符号‘·’与‘+’表示。实验结果如图1、2所示。

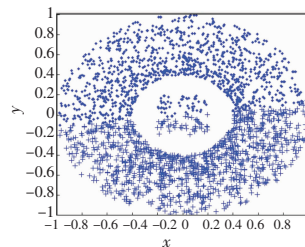


图1 无噪声环形数据 FCM 聚类算法结果

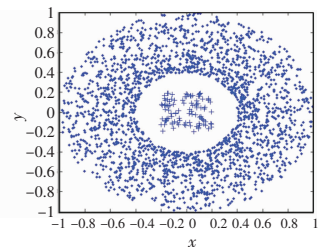


图2 无噪声环形数据 FKPCM、HKPCM 聚类算法结果

在图1中实验结果表明对于环形分布的数据FCM聚类算法不能得到令人满意的效果,而图2FKPCM、HKPCM聚类算法通过引入核函数能够得到比较理想的聚类效果。

3.2 加噪声的环形数据

图3~5在环形数据中加入了噪声来对比FCM、FKPCM、HKPCM算法的聚类性能。

对于加入噪声的环形数据由图3可以看到,FCM算法仍无法得出好的聚类结果;图4FKCM算法虽然(下转第2885页)

(上接第 2853 页)能够大致分开,但圆形内部聚类结果不够理想;图 5 HKPCM 算法则能够得到理想的聚类结果。

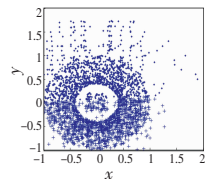


图 3 加噪声的
环形数据 FCM
聚类算法结果

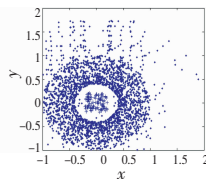


图 4 加噪声的
环形数据 FKCM
聚类算法结果

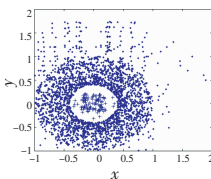


图 5 加噪声的
环形数据 HKPCM
聚类算法结果

4 结束语

本文提出了一种基于混合核函数的可能性模糊 C-均值聚类算法,对于分布复杂和有噪声的数据都能得到理想的聚类效果。首先针对 FCM 算法的缺点在 FCM 算法中引入了核函数的概念,提出了一种模糊核函数 C-均值聚类算法。然而,当远离聚类边缘的位置有噪声或在聚类之间存在过渡点时,FKCM 同 FCM 一样极易受到它们的影响。针对 FKCM 算法的不足,本文提出了基于混合核函数的可能性 C-均值聚类算法。实验结果证实了该算法的可行性与有效性。

参考文献:

[1] 袁玉波,杨传胜,黄延祝,等. 数据挖掘与最优化技术及应用[M].

北京:科学出版社,2007:112-115.

- [2] 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.
- [3] 张莉,周传达,焦李成. 核聚类算法[J]. 计算机学报,2002,25(6):587-590.
- [4] 高新波. 模糊聚类分析及其应用[M]. 西安:西安电子科技大学出版社,2004:23-26.
- [5] 朱明. 数据挖掘[M]. 合肥:中国科学技术大学出版社,2002:138-140.
- [6] KRISHNAPURAM R, KELLER J M. The possibilistic C-means algorithm: insight and recommendations[J]. IEEE Trans on Fuzzy Systems, 1996, 4(3):385-393.
- [7] GIROLAMI M. Mercer kernel based clustering in feature space [J]. IEEE Trans on Neural Networks, 2002, 13(3):780-784.
- [8] ZHANG Jiang-she, LEUNG Y W. Improved possibilistic C-means clustering algorithms [J]. IEEE Trans on Fuzzy Systems, 2004, 12(2):209-217.
- [9] GAN W Y, LI D E. Hiertarchical clustering based on kernel density estimation [J]. Journal of System Simulation, 2004, 16(2):302-309.
- [10] FRANK C H R, KIL-SOO C, BYUNG-IN C. Kernel approach to possibilistic C-means clustering [J]. International Journal of Intelligent Systems, 2009, 24(3):272-293.