基于核的模糊熵聚类算法

何巧玲, 叶东毅
(福州大学 数学与计算机科学学院计算机系　福建 福州　350001)

【摘　要】: Dat Tran 等提出的模糊熵聚类算法 FEC 是模糊 C 均值聚类算法 FCM 的一种改进,FEC 在 FCM 的基础上引入熵的概念,对隶属度值分布方面进行算法的优化,但 FCM 与 FEC 二者在非线性可分数据处理时表现并不理想。本文提出一种新的基于核的模糊熵聚类算法 KFEC,结合模糊熵聚类算法和核聚类算法的优点来增强聚类效果。对比实验表明 KFEC 能够处理非线性可分的数据的聚类问题,在一定程度上提高了聚类的质量。

【关键词】: 聚类;模糊熵;Mercer 核;核函数

1. 引言

Dat Tran 等提出的模糊熵聚类算法 (Fuzzy Entropy Clustering)[1]是模糊 C 均值聚类算法(Fuzzy C-Means Clustering)[2][3]的一种改进,FEC 在 FCM 的基础上引入熵的概念, 对隶属度值分布方面进行算法的优化。FCM 与 FEC 二者都没有对数据样本的特征进行优化,而是直接利用样本的特征进行聚类,因此它们的聚类效果在很大程度上取决于数据样本的分布情况。它们都能够较好得处理线性可分数据集的聚类, 但在非线性可分数据或高维数据处理时表现并不理想。继文献[10][11]的支持向量机的聚类算法之后,不少学者提出基于核函数的聚类算法[4]~[9],基于核方法的聚类算法可以分成三种,第一种是核化距离矩阵,即在特征空间用 Mercer 核方法计算距离矩阵[4];第二种基于支持向量的数据描述的核方法[5];第三种用核方法在特征空间实现 K-Means(C-Means)[6]~[9];。

本文提出一种新的基于核的模糊熵聚类算法(Kernel based Fuzzy Entropy Clustering),结合模糊熵聚类算法和核聚类算法的优点来增强聚类效果。KFEC 算法属于基于核方法的聚类算法的第三种。文章最后用仿真数据 Delta 和真实数据 Iris、breast-cancer-wisconsin 对 FCM,FEC,KFEC 进行对比实验, 结果表明 KFEC 能够处理非线性可分的数据的聚类问题, 在一定程度上提高了聚类的质量。

2. 相关算法

2.1 模糊 C 均值聚类算法:

对 D 维输入空间 R^D 的数据集 $X=\{x_1,x_2,...,x_n\}$ 进行聚类,$X_i=\{x_{i1},x_{i2},...,x_{it}\}$,簇 C_i 的中心 $\theta_i=\{\theta_{i1},\theta_{i2},...,\theta_{id}\}$,i=1,2,…,C(C 为簇的个数),常用的相似度度量公式有欧几里德距离

$$d(x_i,\theta_i)=\sqrt{\sum_{d=1}^{D}(x_{id}-\theta_{id})^2}$$

,表示数据 x_i 与簇中心 θ_i 的距离;$U=[u_{it}]$是数据集的隶属矩阵,u_{it} 表示数据 x_i 属于簇 C_i 的隶属度值

$$0<u_{it}<1,\sum_{i=1}^{C}u_{it}=1,0<\sum_{t=1}^{T}u_{it}<T$$

FCM 的目标函数为:

$$OBJ\_FCN_{FCM}(U,\theta;X)=\sum_{i}^{C}\sum_{t}^{T}u_{it}^{m}d^2(x_t,\theta_i)$$

$$s.t.\ \ 0<u_{it}<1,\sum_{i=1}^{C}u_{it}=1,0<\sum_{t=1}^{T}u_{it}<T \qquad (1)$$

其中 m 为权重指数, 表征聚类的模糊程度,m 的典型值区间为 [1.25,2]。FCM 的聚类准则是最小化约束条件下的目标函数。隶属度和聚类中心的更新方程为:

$$u_{it}=\frac{1}{\sum_{j=1}^{C}(\frac{d(x_t,\theta_i)}{d(x_t,\theta_j)})^{\frac{2}{m-1}}}\qquad \theta_i=\frac{\sum_{t=1}^{T}u_{it}^{m}x_t}{\sum_{t=1}^{T}u_{it}^{m}}$$

2.2 模糊熵聚类(FUZZY ENTROPY CLUSTERING)

在聚类过程中,数据隶属度表征了聚类的模糊程度。若数据对某簇的隶属度远远大于其他类的隶属度时, 聚类的模糊程度低,聚类效果好。由此,在 FCM 的目标函数基础上引入熵函数,得到模糊熵聚类目标函数[1]:

$$OBJ\_FCN_{FEC}(U,\theta;X)=\sum_{i}^{C}\sum_{t}^{T}u_{it}d^2(x_t,\theta_i)+s\sum_{i}^{C}\sum_{t}^{T}u_{it}\log u_{it} \qquad (2)$$

$$s.t.\ \ 0<u_{it}<1,\sum_{i=1}^{C}u_{it}=1,0<\sum_{t=1}^{T}u_{it}<T$$

其中 $s>0$。(2)式的第一项与(1)式相同,第二项为熵函数 与 s 的乘积的负数。

$$E(U)=-\sum_{i=1}^{C}\sum_{t=1}^{T}u_{it}\log u_{it}$$

FEC 的隶属度和聚类中心的更新方程为:

$$u_{it}=\frac{1}{\sum_{j=1}^{C}[\frac{e^{d^2(x_t,\theta_i)}}{e^{d^2(x_t,\theta_j)}}]^{\frac{1}{s}}}\qquad \theta_i=\frac{\sum_{t=1}^{T}u_{it}x_t}{\sum_{t=1}^{T}u_{it}}$$

3. 基于 Mercer 核的模糊熵聚类算法 (KERNEL BASED FUZZY ENTROPY CLUSTERING)

运用核函数方法, 可以把低维空间线性不可分模式通过非线性映射到高维特征空间则实现线性可分。若算法中的矢量间相互作用仅限于内积运算, 则不需要知道非线形变换的具体形式,只要用核函数替换线性算法中的内积,就能得到原输入空间中对应的非线性算法。设非线性映射 $\Phi: x\to\Phi(x)\in H$,将 R^D 空间数据 x 映射到高维特征空间 H 中,原空间的点积在高维特征空间表示为:

$$K(x,y)=<\Phi(x),\Phi(y)>=\Phi(x)^T\Phi(y)$$

满足 Mercer 定理的函数都可以作为核函数 [12]。常用的 Mercer 核函数有高斯核函数、多项式核函数、sigmoidal 核函数等。据研究在缺少先验知识的情况下,用高斯核函数可以基本满足使用要求,因为高斯核函数对应的特征空间是无穷维的,有限的样本在该特征空间肯定是线性可分的。高斯核函数:

$$K(x,y)=\exp(\frac{-\|x-y\|^2}{\sigma})$$

其中 σ 为高斯核函数的宽度;采用高斯核函数,特征空间 H 里的欧几里德距离公式可表示为 :

$$\begin{aligned}&\|\Phi(x_t)-\Phi(\theta_i)\|^2=(\Phi(x_t)-\Phi(\theta_i))^T(\Phi(x_t)-\Phi(\theta_i))\\&=\Phi(x_t)^T\Phi(x_t)-\Phi(\theta_i)^T\Phi(x_t)-\Phi(x_t)^T\Phi(\theta_i)+\Phi(\theta_i)^T\Phi(\theta_i)\\&=K(x_t,x_t)+K(\theta_i,\theta_i)-2K(x_t,\theta_i)\qquad(\because K(x,x)=1)\\&=2-2K(x_t,\theta_i)\qquad(3)\end{aligned}$$

用式(3)代替欧几里德距离公式,得到基于 Mercer 核的模糊熵聚类方法,其聚类准则是使如下的目标函数最小。

基金项目:福建高新计划项目,编号:2005H028