初始聚类中心优化的加权最大熵核 FCM 算法①

许友权[1], 吴 陈[1], 杨习贝[1,2]

[1](江苏科技大学 计算机科学与工程学院, 镇江 212003)
[2](南京理工大学 计算机科学与技术学院, 南京 210094)

摘 要: 针对传统基于最大熵模糊 C 均值聚类算法(MEFCM)仅适用于球状或椭圆状聚类, 为了解决数据分布混乱以及高度相关难以划分的情形, 引入 Mercer 核函数, 使原来没有显现的特征突现出来, 从而使聚类效果更好. 然而在实际问题中, 大多数样本集的样本数据都存在着重要性(权重)不同的现象, 主要针对样本集中各个数据的不同重要程度来设计加权方法, 同时为了克服聚类算法对初始聚类中心选取的敏感性这一弱点, 提出了一个初始聚类中心优化的加权最大熵核模糊聚类算法(WKMEFCM). 通过实验验证, 该算法与原 MEFCM 算法比较, 其聚类结果更加稳定、准确, 从而达到更好的聚类划分效果.

关键词: 核函数; FCM 算法; 特征权重; 最大熵; 初始聚类中心

Maximum Entropy Fuzzy C-Means Clustering Based on Sample Weighting and Initial Cluster Centers

XU You-Quan[1], WU Chen[1], YANG Xi-Bei[1,2]

[1](School of Computer Science & Technology, Jiangsu University of Science and Technology, Zhenjiang 212003, China)
[2](School of Computer Science & Technology, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: This paper aims to demonstrate the traditional maximum entropy fuzzy C-means clustering algorithm (MEFCM) applies to spherical or oval-shaped clusters only. In order to solve the confusion and highly relevant data distribution division of this difficult situation, it introduces Mercer kernel function, so that the original features which do not show can stand out and make the clustering effect better. However, in practical, the majority of sample sets are exist importance (weighting) of different phenomena. The main focus are the samples of different importance to design of each data weighting and in order to overcome the sensitivity of weakness of the initial cluster centers. This paper presents an optimization of the initial cluster centers weighted kernel maximum entropy fuzzy clustering algorithm (WKMEFCM). Experiments show that compared with the MEFCM, the clustering result is more stable, accurate and the clusters division effect is better.

Key words: kernel function; FCM algorithm; feature weighting; maximum entropy; initial cluster centers

聚类分析是人类最基本认知活动之一, 人们常说的“物以类聚”实际上就反映了聚类分析的基本思想. 聚类的基本出发点在于使类内模式间的相似度尽量大, 而类之间的相似度尽量小. 聚类分析应用领域很广泛, 近年来随着其不断发展, 已经被广泛地应用在数据挖掘、图像处理、模式识别等领域中[1].

比较经典的聚类方法有传统的 C 均值方法和模糊 C 均值聚类方法[2]. 后来随着熵理论的不断发展, 研究人员开始将熵理论引入到模糊聚类分析领域中, 提出了基于最大熵的模糊聚类分析方法[3]. 随后在此基础上, 虽然产生了很多该方法的变形[4-6], 但这些方法都没有对样本的特征进行优化, 而是直接利用样本的特征进行聚类. 上述这些方法在适用于样本分布为球状或椭圆状时, 聚类效果很好. 然而在一类样本散布较大, 另一类散布较小的话, 这些方法效果就比较差. 核模糊 C 均值算法 KFCM[7](Kernel Fuzzy C-Means)

①基金项目:国家自然科学基金(61100116);江苏省自然科学基金重点资助项目(BK2011492)
收稿时间: 2013-12-16;收到修改稿时间: 2014-03-24

首先将数据映射到高维空间, 提取并放大有用特征, 从而更好地完成了聚类. 虽然 KFCM 在处理非线性问题上有突出表现, 但存在对初始聚类中心选取敏感的缺陷.

初始聚类中心的选取的方法有随机抽样、距离优化和密度估计三种方法[8]. 传统的 C 均值方法其初始聚类中心的选取就是用随机抽样方法. 文献[9]提出了一种基于最大最小距离法寻找初始聚类中心的方法. 文献[10]用密度函数法求得样本数据空间的多个聚类中心, 并结合小类合并运算, 能很好地避免局部最小.

本文借助于核方法在 FCM 上的成功运用的思想, 将核方法应用于基于最大熵的 FCM 算法中. 然而, 其仍存在和 KFCM 一样的缺陷, 即对初始聚类中心选取敏感的缺点, 为了克服这一弱点, 本文利用样本点分布密度大小作为权值, 借助数据本身的分布特性, 事先优化初始聚类中心, 提出了一个初始聚类中心优化的加权最大熵核模糊聚类算法(WKMEFCM). 通过 UCI 机器学习数据库数据集的测试, 证实本文所提出的算法与原 MEFCM 算法比较, 在聚类效果上有明显的改善.

1　基于最大熵准则的FCM算法

基于最大熵准则的FCM[3]是在传统FCM的基础上提出的, 是将模糊聚类与熵进行有机结合的一种方法, 它不仅具有数据样本间用熵表示相关信息的优点, 而且具有模糊聚类方法这种软聚类的优质特性, 因而在聚类划分领域有着重要的地位. 假设对于数据集 $\boldsymbol{X}=\{x_1,x_2,...,x_n\}$, $x_j\in\boldsymbol{R}^d$, $j=1,2,...,n$, c为预定的类别数目, $\boldsymbol{v}_i$, $i=1,2,...,c$ 为每个聚类的中心, $u_{ij}\in[0,1]$表示第j个样本对于第i类的隶属度. FCM算法在迭代寻优过程中, 不断更新各类的中心以及所有样本的隶属度, 直到使准则函数

$$J=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2} \tag{1}$$

达到最小. 其中 $\boldsymbol{d}_{ij}^2=\|\boldsymbol{x}_j-\boldsymbol{v}_i\|^2$, 式(1)的约束条件为 $\sum_{i=1}^{c}u_{ij}=1,\forall j$, 运用拉格朗日乘子法, 可得无约束的准则函数

$$L_{FCM}=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2}-\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1) \tag{2}$$

其中模糊熵定义:

$$S_{FE}=-\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\ln u_{ij}\ ,i=1,2,...,c\ ,\ j=1,2,...,n \tag{3}$$

根据文献[11]描述的最大熵聚类原理, 构造最大熵目标函数:

$$\max\ \ J_E=-\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2}-\eta^{-1}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\ln u_{ij} \tag{4}$$

要使(4)式目标函数值最大, 其等价于使下式目标函数值最小:

$$\min\ \ \hat{J}_E=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2}+\eta^{-1}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\ln u_{ij} \tag{5}$$

约束条件为 $\sum_{i=1}^{c}u_{ij}=1,\forall j$, 则应用拉格朗日乘数法, 构造如下的优化准则函数为

$$L=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}d_{ij}^{2}+\eta^{-1}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\ln u_{ij}-\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1) \tag{6}$$

其中 $\eta>0$ 称为差异因子, 最大化目标函数L, 令 $\partial L/\partial u_{ij}=0$ 则得出

$$u_{ij}=\frac{e^{-\eta d_{ij}^2}}{\sum_{i=1}^{c}e^{-\eta d_{ij}^2}} \tag{7}$$

再令 $\partial L/\partial \boldsymbol{v}_i=0$ 得

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}\boldsymbol{x}_j}{\sum_{j=1}^{n}u_{ij}} \tag{8}$$

2　基于核函数的最大熵准则的 FCM 算法

对于数据集 $\boldsymbol{X}=\{x_1,x_2,...,x_n\}$, $x_j\in\boldsymbol{R}^d$, $j=1,2,...,n$, Mercer核[12]非线性映射为 $\phi:\boldsymbol{x}\to\phi(\boldsymbol{x})$, 则由(2)式目标函数变为

$$\min\ \ \hat{J}_E=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}\|\phi(\boldsymbol{x}_j)-\phi(\boldsymbol{v}_i)\|^2+\eta^{-1}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\ln u_{ij} \tag{9}$$

对应的准则函数为

$$\begin{aligned}L=&\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}\|\phi(\boldsymbol{x}_j)-\phi(\boldsymbol{v}_i)\|^2+\\&\eta^{-1}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\ln u_{ij}-\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1)\\=&\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}\left[k(\boldsymbol{x}_j,\boldsymbol{x}_j)-2k(\boldsymbol{x}_j,\boldsymbol{v}_i)+k(\boldsymbol{v}_i,\boldsymbol{v}_i)\right]+\\&\eta^{-1}\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}\ln u_{ij}-\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1)\end{aligned} \tag{10}$$

式(10)中 $\boldsymbol{v}_i$ 为第i类的类中心, $\phi(\boldsymbol{v}_i)$ 为该中心在相应核空间中的像.

同理, 令 $\partial L/\partial u_{ij}=0$ 则得出

$$u_{ij}=\frac{e^{-\eta\left[k(\boldsymbol{x}_j,\boldsymbol{x}_j)-2k(\boldsymbol{x}_j,\boldsymbol{v}_i)+k(\boldsymbol{v}_i,\boldsymbol{v}_i)\right]}}{\sum_{i=1}^{c}e^{-\eta\left[k(\boldsymbol{x}_j,\boldsymbol{x}_j)-2k(\boldsymbol{x}_j,\boldsymbol{v}_i)+k(\boldsymbol{v}_i,\boldsymbol{v}_i)\right]}} \tag{11}$$

再令 $\partial L / \partial \boldsymbol{v}_i = 0$ 得

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} u_{ij}^m K(\boldsymbol{x}_j, \boldsymbol{v}_i)\boldsymbol{x}_j}{\sum_{j=1}^{n} u_{ij}^m K(\boldsymbol{x}_j, \boldsymbol{v}_i)} \tag{12}$$

3 初始聚类中心优化的加权最大熵核 FCM 算法

假定数据集$\boldsymbol{X}$具有n个数据样本, 每一个数据样本可用 x_i 表示, 则第j个数据样本的权重可以按照下式进行计算

$$w_j = \frac{\sum_{i=1, i\neq j}^{n} (1/d_{ij})}{\sum_{j=1}^{n}\sum_{i=1, i\neq j}^{n} (1/d_{ij})} \tag{13}$$

式(13)称为样本点分布密度, 式中的 d_{ij} 代表第j个数据样本 x_j 与第i个数据样本 x_i 之间的远近程度. 对于权重 w_j 的计算公式来说, 其分子部分表示数据样本 x_j 与所有其他数据样本之间距离的倒数总和, 而分母部分则是对数据集$\boldsymbol{X}$中的所有数据样本均按照分子方法所求数值的总和. 这种方法的主要思想是假如某个数据样本与其他数据样本之间的距离函数值越大, 则这个数据样本越疏远、越孤立于其他数据样本, 其所起的作用也就越小, 从而其权重也就越小; 相反, 某个数据样本与其他所有数据样本之间的距离函数值越小, 则说明了这个数据样本其周围的数据样本越多, 从而其所起的作用越大, 权重则也就越大.

将式(9)变成基于特征加权的目标函数

$$\min \ \hat{J}_E = \sum_{j=1}^{n}\sum_{i=1}^{c} w_j u_{ij}^m \| \phi(\boldsymbol{x}_j) - \phi(\boldsymbol{v}_i) \|^2 + \eta^{-1} \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij} \ln u_{ij} \tag{14}$$

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} w_j u_{ij}^m K(\boldsymbol{x}_j, \boldsymbol{v}_i)\boldsymbol{x}_j}{\sum_{j=1}^{n} w_j u_{ij}^m K(\boldsymbol{x}_j, \boldsymbol{v}_i)} \tag{15}$$

数据样本点 x_j 隶属于各个簇的隶属度为

$$u_{ij} = \frac{e^{-\eta w_j \left[k(\boldsymbol{x}_j, \boldsymbol{x}_j) - 2k(\boldsymbol{x}_j, \boldsymbol{v}_i) + k(\boldsymbol{v}_i, \boldsymbol{v}_i)\right]}}{\sum_{i=1}^{c} e^{-\eta w_j \left[k(\boldsymbol{x}_j, \boldsymbol{x}_j) - 2k(\boldsymbol{x}_j, \boldsymbol{v}_i) + k(\boldsymbol{v}_i, \boldsymbol{v}_i)\right]}} \tag{16}$$

上式中初始聚类中心 $\boldsymbol{v}_i$ 一般是随机选取的, 且聚类结果对选取的 $\boldsymbol{v}_i$ 敏感, 本文聚类中心选取的基本思想是首先利用式(13)计算每个样本的密度, 从而得到一个以密度为准则的样本集合D, 然后在集合D基础上进行初始聚类中心的选取和簇的划分. 每划分出一个簇, 就从集合D中删除该簇所包含的样本.

根据以上推导, 则聚类中心优化的加权最大熵核 FCM算法详细描述如下:

1)选定核函数 $K(x,y)$, 分别设置模糊指数m, 隶属度精度 ε, 差异因子 η, 最大迭代次数 T, 聚类数 c.

2)计算每一个数据样本的权重 w_j, 并对数据集 $\{w_j\}$ 按从大到小进行排序, 从而得到密度点的集合D.

3)选取当前密度最大的一个数据样本点作为初始聚类中心 $\boldsymbol{v}_i$, $i \in \{1,2,...,c\}$.

4)以 $\boldsymbol{v}_i$ 为簇中心, 以样本平均距离为半径取一圆域, 将此圆域作为一个划分, 并且从集合D中删除被划分的数据点.

5)重复2)~4), 直到选取c个初始聚类中心点.

6)利用获取的聚类中心 $\boldsymbol{v}_i$ 重复下面的运算:

(a): 用当前的聚类中心根据式(16)更新隶属度

(b): 用当前的聚类中心和隶属度根据式(15)更新各个聚类中心

直到隶属度满足 $\max_{i,j} |u_{ij}(t) - u_{ij}(t-1)| < \varepsilon$ 或者 $t > T$, 算法终止.

其中, 样本平均距离公式定义为

$$\bar{d} = \frac{\sum_{i=1, i\neq j}^{n} d_{ij}}{C_n^2} \tag{17}$$

式(17)中 C_n^2 是从n个样本中任取两个样本的组合数.

4 实验结果与分析

为验证本文提出的算法的有效性, 实验采用 UCI 数据集中的 iris、breast 和 wine 作为测试对象, 在 MATLAB 实验环境下进行相关实验分析, 对传统的 FCM 算法、MEFCM 算法、KMEFCM 算法以及 WKMEFCM 算法分别在聚类准确率上和稳定性上进行了比较. 实验中, 采用的核函数为高斯核, 即 $K(x,y) = \exp(-\| x-y \|^2 / \sigma^2)$, 其中 σ 为自定义的参数.

我们设置高斯核函数 $\sigma = 150$, 模糊指数 $m = 2$, 差异因子 $\eta = 100$. 结果表明: 本文提出的初始聚类中心优化的加权最大熵核 FCM 算法在聚类准确率上和稳定性上比另外三种聚类算法都要具有更好的聚类效果. 其中 UCI 数据信息见表 1.

在实验过程中, 我们对每个 UCI 数据集各实验 20 次, 记录每一次的聚类结果, 最后求出平均正确率. 由表 2 可以看出, 采用核函数的 KMEFCM 聚类划分方

法对实际数据集进行聚类划分的效果要好于 FCM 和 MEFCM 的聚类划分效果. 同时, 优化初始聚类中心的 WKMEFCM 方法要比 KMEFCM 方法具有更好的聚类划分效果.

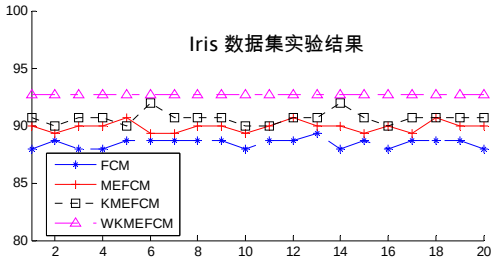
表 1　UCI 数据信息

数据集	样本数目	属性数目	簇数
Iris	150	4	3
Breast	699	9	2
Wine	178	13	3

表 2　FCM、MEFCM、KMEFCM 以及 WKMEFCM 的聚类准确度比较

数据	算法	最高错分个数	最低错分个数	平均正确率
Iris	FCM	18	16	88.47%
	MEFCM	16	14	89.90%
	KMEFCM	15	12	90.63%
	WKMEFCM	11	11	92.67%
Breast	FCM	39	35	94.59%
	MEFCM	32	27	95.67%
	KMEFCM	25	22	96.67%
	WKMEFCM	20	20	97.14%
Wine	FCM	20	18	89.19%
	MEFCM	18	15	90.53%
	KMEFCM	16	10	93.31%
	WKMEFCM	11	11	93.82%

图 1 是四种不同聚类算法对 Iris、Breast 和 Wine 数据集各做 20 次的聚类实验结果. 从图中可以看出, 由于对数据样本进行了加权, 使得聚类划分的性能得到了提升, 对数据样本进行加权的方法也更加有效的促进了核函数方法对于非线性聚类划分的能力. 同时, WKMEFCM 算法的聚类结果比其余三种聚类算法的聚类结果的稳定程度要好, 因此可以看出本文提出的初始聚类中心优化的 KMEFCM 算法其聚类结果在聚类结果上要更加稳定, 同时对数据样本进行加权有助于进一步提高聚类的性能. 聚类中心不会随测试次数的增加而改变, 并且准确率也相对较高, 可见该算法具有很好的性能.



Breast 数据集实验结果

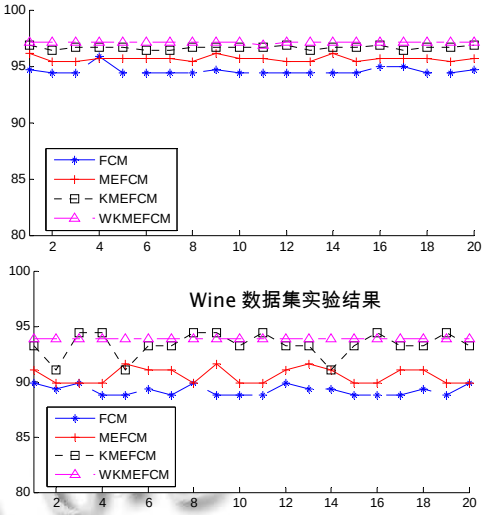


图 1 Iris、Breast 和 Wine 数据集各 20 次聚类实验结果

5　结语

基于最大熵模糊聚类算法是一种被广泛应用的算法, 但是它的聚类结果会随预先给定的初始聚类中心的不同而产生波动, 同时该算法很难解决数据分布混乱以及高度相关难以划分的这种情况. 文采将密度的思想和核方法相结合, 提出了一种初始聚类中心优化的加权最大熵核 FCM 算法, 实验证明了该算法的有效性. 采用基于密度的方法, 客观地反映了数据的分布状况, 在此基础上选取初始聚类中心, 提高了算法的稳定性. 同时引入核方法, 使原来没有显现的特征突现出来, 从而进一步提高了聚类的质量, 获得了更好的聚类效果. 但是本文的算法只是对样本数目较少和维数较小的样本集进行了实验分析, 同时高斯核函数的参数和差异因子的选取都是基于经验性的, 因此还需要作进一步的研究. 下一步我们将研究和的选取, 以及利用该算法对文本这种海量、高维的数据集进行应用研究, 同时在优化算法降低时间复杂度上做改进.

参考文献

1 Mirkin BG. Clustering: A Data Recovery Approach. CRC Press, 2012.

2 Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Wiley. com, 2009.

3 Li RP, Mukaidono M. A maximum-entropy approach to fuzzy clustering. Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on

Fuzzy Systems and The Second International Fuzzy Engineering Symposium. Proc. of 1995 IEEE International Conference on. IEEE. 1995, 4. 2227–2232.

4 Pavlov DY, Pennock DM. A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains. Advances in Neural Information Processing Systems. 2002. 1441–1448.

5 Shitong W, Chung KFL, Zhaohong D, et al. Robust maximum entropy clustering algorithm with its labeling for outliers. Soft Computing, 2006, 10(7): 555–563.

6 Jenssen R, Eltoft T, Girolami M, et al. Kernel maximum entropy data transformation and an enhanced spectral clustering algorithm. Advances in Neural Information Processing Systems. 2006. 633–640.

7 Wu Z, Xie W, Yu J. Fuzzy c-means clustering algorithm based on kernel method. Proc. of the Fifth International Conference on. Computational Intelligence and Multimedia Applications. IEEE. 2003. 49–54.

8 HeJ L. Initialization of cluster refinement algorithms: are-view and comparative study. Proc. of International Joint Conference on Neural Networks. 2004. 297–298.

9 周涓,熊忠阳,张玉芳,等.基干最大最小距离法的多中心聚类算法.计算机应用,2006.

10 Cao F, Ester M, Qian W, et al. Density-based clustering over an evolving data stream with noise. Proc. of the 2006 SIAM International Conference on Data Mining. 2006. 328–339.

11 H-L Eng, K-K Ma. Unsupervised image object segmentation over compressed domain. Proc. of the IEEE Int. Conf. Image. 2000, 3. 758–761.

12 Girolami M. Mercer kernel-based clustering in feature space. IEEE Trans. on Neural Networks, 2002, 13(3): 780–784.