对 K-means 及势函数聚类算法的研究与改进①

叶于林[1], 夏秀渝[1], 莫建华[2], 刘 帅[2]

[1](四川大学 电子信息学院, 成都 610065)
[2](中国人民解放军 78438 部队, 成都 610066)

摘 要: 在目前聚类方法中, k-means 与势函数是最常用的算法, 虽然两种算法有很多优点, 但也存在自身的局限性. k-means 聚类算法: 其聚类数目无法确定, 需要提前进行预估, 同时对初始聚类中心敏感, 且容易受到异常点干扰; 势函数聚类算法: 其聚类区间范围有限, 对多维数据进行聚类其效率低. 针对以上两种算法的缺点, 提出了一种基于 K-means 与势函数法的改进聚类算法. 它首先采用势函数法确定聚类数目与初始中心, 然后利用 K-means 法进行聚类, 该改进算法具有势函数法"盲"特性及 K-means 法高效性的优点. 实验对改进算法的有效性进行了验证, 结果表明, 改进算法在聚类精度及收敛速度方面有很大提高.

关键词: 聚类; K-means 聚类算法; 势函数聚类算法

Research and Improvement of K-means and Potential Function Clustering Algorithm

YE Yu-Lin[1], XIA Xiu-Yu[1], MO Jian-Hua[2], LIU Shuai[2]

[1](College of Electronics and Information, Sichuan University, Chengdu 610064, China)
[2](78438 Troops of the Chinese people's Liberation Army, Chengdu 610066, China)

Abstract: In the present clustering method, k-means with potential function is the most commonly used algorithm, although the two algorithms have many advantages, but they also have their own limitations. The clustering number of k-means clustering algorithm cannot be determined, estimate in advance, at the same time sensitive to initial clustering center, and easy to be interfered by abnormal point, the clustering range of potential function clustering algorithm is limited, low efficiency of clustering multidimensional data. In view of the above two algorithms disadvantage, an improved clustering algorithm based on K-means and potential function is proposed in the paper. First, potential function method is used to determine the clustering number and initial center, and then cluster by using K-means method. The improved algorithm has the advantage of blind characteristics of potential function algorithm and also has the advantages of high efficiency of K-means. The experiment verified the validity of the improved algorithm, the results show that the improved algorithm have greatly improved in clustering accuracy and convergence speed.

Key words: clustering; K-means clustering algorithm; potential function clustering algorithm

随着科学技术的进步和社会经济的快速发展, 使得各行业的数据量急剧增加, 如何处理这些海量数据, 并从中提取有效信息为我所用成为了目前各行业的当务之急. 由此, 学者们提出了通过聚类分析[1,2]方法对数据对象根据最大化类内相似、最小化类内间相似的原则进行分组, 并组成多个类或簇, 使得在同一类(簇)的对象之间具有较高的相似性, 而不同类(簇)的对象具有较大的相异性. 其方法效率高、通用性强, 并在经济学、医学、生物学、通信工程、工程技术等领域得到了广泛应用.

目前, 根据各行业的特性, 人们提出了多种聚类算法, 简单可以分为: 基于层次、划分、密度、图论、网格和模型的几大类. 基于划分的聚类算法包括有 K-means 算法[3,4], PAM 算法, EM 算法等, 其中最典型的是 K-means 算法. 它能快速、准确且可对多维数据进行聚类, 不足之处必须提前知道聚类数目及对初始

① 收稿时间:2014-10-10;收到修改稿时间:2014-11-03

聚类中心进行预估, 且容易受到异常点的干扰. 在语音欠定盲分离中, 常采用另外一种聚类方法: 势函数聚类法[5] (potential function clustering algorithm), 它由 Pau Bofill 首先提出, 主要对声源的相对幅度衰减和时间延迟进行聚类估计. 其具有“盲”特性, 即无需提前知道聚类数目, 能自动识别聚类数目且对初始聚类中心没有依赖性, 但是经典势函数聚类算法对多维数据进行聚类有一定的局限性. 因此, 本文提出了一种基于两种算法的优缺点的改进算法, 该改进算法具有势函数算法“盲”特性的优点, 又具有 K-means 算法的快速、高效、准确的对多维数据进行聚类的优点. 实验结果表明, 该改进算法聚类效果优于原始的两种算法并具有较好的稳定性.

1　K-means聚类算法

1.1 K-means 聚类算法分析

K-means 算法[6]是最经典的聚类算法, 比较常用. 其属于基于距离的聚类算法, 是采用目标函数聚类方法的代表. 目标函数是以数据点到质心点的距离作为优化, 再利用一定函数得到迭代运算的调整规则. 以欧式距离作为相似性度量, 使得目标函数的评价指标 J 最小, J 的数学表达式如 1.1 式所示, 采用了误差平方和作为聚类准则函数, 如果有 N 个数据点需要分为 K 个聚类, K-means 要做的就是最小化 J.

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left\|x_n - \mu_k\right\|^2 \tag{1}$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{2}$$

1.2 K-means 聚类算法的具体步骤

① 确定 K 个中心 μ_k 的初值. 为了得到全局最优解, 可能需要多次选取初始值, 然后从中进行选取.

② 将剩余数据点归类到离它最近的那个中心点的类中.

③ 用公式 $\mu_k = \frac{1}{N_k}\sum_{j\subset cluster_k} x_j$ 计算出每个类的新的中心点.

④ 重复 2, 直到 J 的值相差等于或小于某一个指定阈值为止.

1.3 K-means 聚类算法的优缺点

① 对于一般聚类数据, 即协方差不太大, 具有一定聚类特性的数据, 能快速、有效、准确的进行聚类且能对多维数据进行聚类;

② 但聚类数目需提前给定, 且 K 值的选定难以估计. 有些算法专门提出以确定其分类数目, 比如方差分析理论;

③ 初始聚类中心的选择问题. 初始聚类中心具有非常重要的作用, 如果初始聚类中心选择不当, 直接会影响聚类估计精度, 且将大大影响聚类效果即鲁棒性不够好, 甚至可导致聚类失败.

2　势函数聚类算法

2.1 势函数聚类算法分析

为了解决通过稀疏分量分析法(SCA)[7-9]算法来实现欠定盲分离[10]中信号稀疏性受限的问题. 2001 年, Pau Bofill 首先提出了一种两步盲分离算法, 其第一步就是利用势函数[11]聚类分析混合矩阵, 主要对未知声源的相对幅度衰减和相对时间延迟进行聚类估计. 由此产生了势函数聚类算法. 算法的提出是在双麦克风欠定混合模型下, 混合模型如图 1 所示.

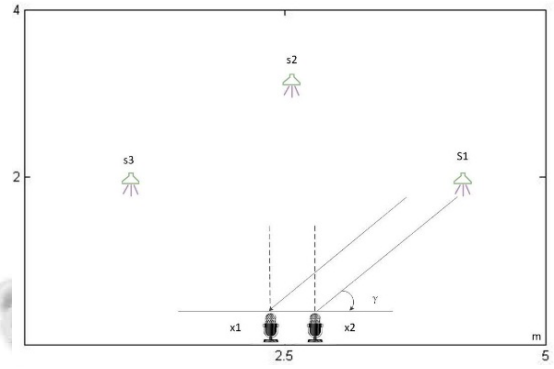

图 1　双麦克风欠定混合模型

图 1 中, n 个源信号通过不同传输路径到达 m 个麦克风($m<n$). 其势函数表达式为:

$$\Phi(\theta,\lambda) = \sum_t l_t \varphi(\lambda(\theta - \theta(t))) \tag{3}$$

$\theta(t)=\tan^{-1}(x_1(t)/x_2(t))$ 作为相对幅度衰减表达式, 这里转换成角度(0~90 度)表示, 如果作为一般聚类函数来讨论, 可以忽略, 仅作为一个普通变量即可. 式(3)中, λ 是需要自己根据聚类的数据情况而设定的可变参数, 它控制 $\varphi(\alpha)$ 中 α 的取值范围, 亦在一定程度上将影响势函数曲线的陡峭程度. $l_t = \sqrt{x_1^2(t)+x_2^2(t)}$ 为能量贡献因子, 在语音处理中可在一定程度上抑制噪

声，增强源信号语音的强度，作为一般聚类函数应用时可以置 1. 基函数 $\varphi(\alpha)$ 的数学表达式为:

$$\varphi(\alpha)=\begin{cases}1-\dfrac{\alpha}{\pi/4} & |\alpha|<\pi/4\\ 0 & \text{其它}\end{cases} \tag{4}$$

如需要求得势函数的“势能”，需要先对自变量 θ 进行取值. θ 的取值范围为 $[\pi/2L,\pi]$，即把 $[0,\pi]$ 的值做 L 等分，第一个取值为 $\pi/2L$，L 决定了取值的数目，即决定了聚类的精度. 在依次求出 L 个 θ 值对应的势函数 $\Phi(\theta,\lambda)$ 值的大小之后，就可以画出一条势函数曲线. 对于势函数曲线我们可以求出它的极大值以及极大值的个数，其中极大值的个数就是源信号的个数，极大值对应的 θ 就是混合信号的数据最集中的地方. 在求出所有的 θ 后，根据 θ 就可以求出相对幅度衰减和相对时间延迟.

2.2 势函数聚类算法的优缺点

① 势函数法对分布在 $[\pi/2L,\pi]$ 内的具有一定聚类特性的数据具有很好的聚类估计效果.

② 其中 L 决定聚类估计精度，λ 决定聚类估计中心. 因此，势函数法相比 K-means 法更具有“盲”特性，即不需要提前知道其聚类数目，能自动识别，也不受聚类初始中心选择的影响.

③ 聚类区间范围受限，只能在接近于 0 到 π 的范围内，如果需要聚类二维及多维数据，便不能再直接对其进行聚类.

3 改进的聚类算法及实验分析

3.1 改进的聚类算法

为了能够快速准确的对多维数据进行“盲”聚类，本文针对 K-means 与势函数聚类算法各自的优缺点，结合两种算法的特性实现了一种改进聚类算法[12]. 其实现方法为: 我们首先可以利用势函数聚类算法对原始数据进行初始聚类，“盲”识别出聚类数目与初始中心；然后再把相关聚类数目与初始中心传递给 K-MEANS 算法进行聚类，从而实现了 K-MEANS 算法对多维数据的“盲”聚类，具体实现方法流程如框图 2:

该改进聚类算法，通过实现流程图我们可以看出分两步来实现，第 1 步是通过势函数法对原始数据进行聚类，通过计算公式(1)与(2)来实现；第 2 步再利用 K-means 对数据进行聚类，借助度量公式(1)与(1.2)来实现.

第 1 步其关键主要是获取聚类数目及初始聚类中心，由于势函数法的实现主要是针对语音的欠定盲分离提出的，其实现前提是信号具有稀疏性，所谓稀疏性[13]是指信号在时域或者其变换域中的大多数的采样时刻的取值等于零或者十分接近于零，而极少数的采样时刻的取值明显不为零，即从信号的时频域看，在较多时频点，每个时频点的能量主要由一个源信号贡献. 自然情况下信号在时域的稀疏性并不一定很明显，但是这些信号在其变换域中稀疏性就可能比较明显. 这就需要对信号进行一些变换，比如短时傅立叶变换，小波变换或者 Gabor 变换等. 因此，利用信号具有稀疏性特性，通过势函数法可以获取信号的势能曲线，从势能曲线中通过观察波峰的个数就可以获取信号相应的聚类数目，波峰位置对应的数值获得相应的初始聚类中心. 在通过势函数法获取势能曲线过程中，这里 I_t 能量贡献因子置 1，对 θ 的取值范围进行 L(自定义)等分，再依据原始势函数公式依次求出 L 个 θ 值对应的势函数 $\Phi(\theta,\lambda)$ 的势能，进而可获得势能曲线.

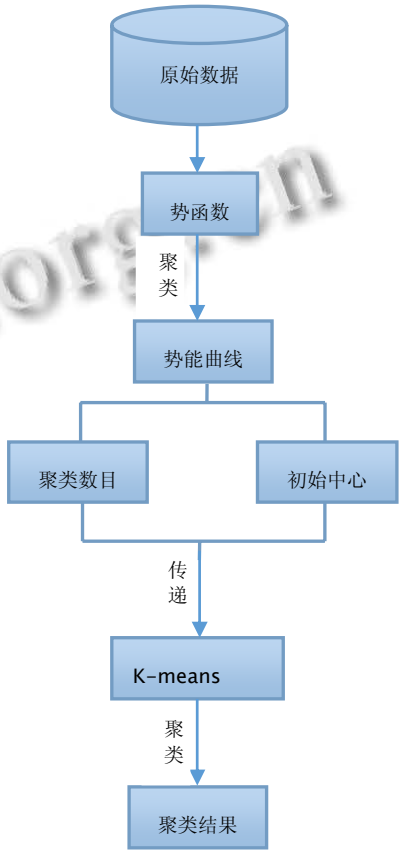

图 2 改进聚类算法实现框图

第 2 步通过 K-means 对数据进行聚类, 其主要是把通过势函数法获取的聚类数目及初始聚类中心传递给 K-means 进行聚类, 这样就避免了 K-means 聚类算法需要对聚类数目提前预估及对初始聚类中心的依赖. 从而提高了收敛速度及聚类精度.

3.2 改进的聚类算法的优点

① 具有势函数聚类算法的“盲”特性的优点. 即在不需要提前知道其聚类数目, 就能自动识别, 同时避免了对聚类初始中心的依赖;

② 具有 K-means 聚类算法能对多维数据进行聚类以及收敛速度快的优点. 即避免了对多维数据聚类范围的局限性;

③ 具有更好的聚类稳定性、可控性, 控制参数变化范围小, 且聚类精度也较高.

3.3 实验仿真分析

根据提出的算法流程图, 对算法进行了实验仿真进行验证. 为了对相关的聚类算法进行实验仿真对比, 首先利用 MATLAB 的 mvnrnd 函数产生服从正态分布的随机数, 然后对随机数进行聚类估计, 以验证改进的聚类算法的性能. 实验条件为: 随机产生的均值为 1.0, 协方差为 0.5 的正态分布随机数的 100 个数据样点, 数据采样点图如 3 所示.

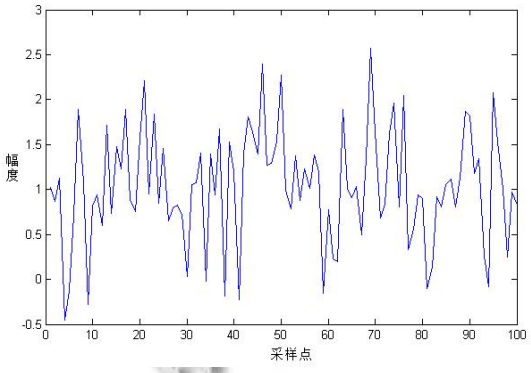

图 3　数据样点图

① 直接利用原始 K-MEANS 算法进行聚类, 聚类数目取 1. 其聚类估计结果如表 1 所示.

表 1　K-MEANS 算法聚类估计结果

均值	协方差	估计值	时间(s)
1.0	0.5	0.9863	0.2873

② 利用改进的聚类算法进行聚类分析, 实验条件同上.

第 1 步, 利用势函数聚类算法进行聚类分析确定其聚类数目和初始中心. 首先给定 L (为 250)和 λ (为 0.2)的值, 然后进行聚类. 其聚类估计的势函数曲线如图 4 所示.

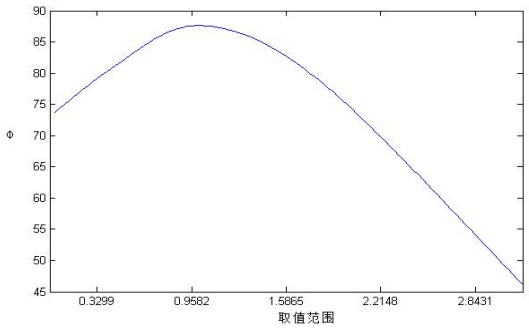

图 4　一维数据势函数曲线图

从图 4 我们可以明显看出有一个波峰, 即聚类数目为 1. 同时“势能”在接近 1 附近达到最大, 此时我们便可根据势能最大点的横坐标确定其估计值, 即为初始聚类中心.

第 2 步, 根据势函数聚类分析确定的聚类数目和初始中心, 直接传递给 K-MEANS 算法, 再利用 K-MEANS 算法进行聚类. 其聚类结果如表 2 所示. 实验所得值均是 100 次实验的统计平均值.

表 2　基于势函数聚类确定聚类数目和初始中心的 K-MEANS 算法聚类结果

均值	协方差	L	λ	估计值	时间(s)
1	0.5	250	0.2	0.9906	0.0683

通过对表 1 和表 2 的实验结果进行对比, 我们明显看到改进的聚类算法的聚类估计结果优于原始的 K-MEANS 算法聚类, 即估计精度更高; 同时聚类估计收敛时间也有大大缩短. 由此, 我们得出: 对于具有一定聚类特性的数据, 改进的聚类算法能够对数据进行很好的聚估计, 聚类结果精度较好; 同时也比 K-means 聚类算法更具有“盲”特性. 同样在对多维数据进行数据聚类时, 须先通过势函数法单独提取每维变量的势能曲线, 以确定聚类数目和初始中心, 再利用 K-means 聚类算法就可对多维数据进行聚类了.

4　结语

K-means 与势函数聚类算法在实际的聚类分析中, 各有优缺点. 本文主要结合两种聚类方法的优点提出

了一种改进的两者相结合的聚类算法, 弥补了两种聚类方法的不足. 其主要克服了 K-means 算法需提前知道聚类数目及对初始中心预估的缺点, 利用势函数法的“盲”特性自动识别聚类数目, 以及对初始聚类中心选择没有依赖性; 同时克服了势函数法聚类范围受限, 不能进行多维数据聚类的缺点, 利用 K-means 算法可对多维数据聚类的优点, 可快速高效的进行聚类. 通过实验证明该改进方法比原始的两种算法聚类更具有稳定性、准确性、可行性.

参考文献

1 姜园,张朝阳,仇佩亮,周东方.用于数据挖掘的聚类算法.电子与信息学报,2005,27(4): 655–662.

2 郭军华.数据挖掘中聚类分析的研究[硕士学位论文].武汉:武汉理工大学,2003.

3 步媛媛,关忠仁.基于 K-means 聚类算法的研究.西南民族大学学报(自然科学版),2009,35(1):198–200.

4 李卫平.K-Means 聚类算法研究.中国西部科技,2008,7(8): 52–53.

5 Bofill P, Zibulevsky M. Underdetermined blind source separation using sparse Representation. Signal Processing, 2001, 81(11): 2353–2362.

6 张玉芳,毛嘉莉,熊忠阳.一种改进的 K-means 算法.计算机应用,2003,23(8):31–33.

7 李昌利.欠定盲源分离的稀疏分量分析方法.广东海洋大学学报,2009,29(4):70–74.

8 邱天爽,毕晓辉.稀疏分量分析在欠定盲源分离问题中的研究进展及应用.信号处理,2008,24(6): 966–970.

9 李白燕,郭水旺,李应生.基于两步法稀疏分量分析的欠定盲源分离.电声技术,2010,34(9):64–67.

10 高波.欠定盲源分离及其应用[硕士学位论文].大连:大连理工大学,2009.

11 代勇,夏秀渝,陈林.一种改进的势函数聚类算法.电子技术应用,2013,39(11):107–110.

12 杨静,张玉洁,李宏伟.基于 K-均值聚类和势函数法的欠定盲分离.电信科学,2012,28(1):98–101.

13 何昭水,谢胜利,傅予力.信号的稀疏性分析.自然科学进展,2006,16(9):1167–1173.