可能性 C-Means 聚类算法的仿真实验*

吕 佳
(重庆师范大学 数学与计算机科学学院, 重庆 400047)

中图分类号:TP18 文献标识码:C 文章编号:1672-6693(2005)03-0129-04

聚类是按照事物间的某种属性将其归入不同的类别,它不需先验知识,而是由机器学习后自动分类,属于无监督的学习方法。聚类分析是多元统计分析的一种,也是非监督模式识别的一个重要分支,在模式识别、图像分割、特征提取等领域中得到了广泛的应用。

传统的聚类分析是一种硬划分,是把每一个待处理的对象严格地划分到某类中,具有"非此即彼"的性质。而客观世界中大多数对象并没有严格的属性,因此这种硬划分并不能真正地反映对象和类的实际关系,适合进行软划分。1981 年 Bezdek 提出了模糊 C-Means(FCM, Fuzzy C-Means Algorithm)[1] 聚类算法,它是一种结合了模糊集理论和 K-Means 聚类算法的用于进行软划分的模糊聚类分析方法。这种算法因其简单、有深厚的数学基础且收敛速度快而得到广泛的应用。但其容易陷入局部最小,且在样本含有噪音时聚类效果差,这是因为 FCM 算法规定了每个样本对各个类的隶属度的和必须为 1,即假定每个数据点对聚类的影响力是相同的,这样就使得当样本中出现噪音和孤立点时,其被赋予了较大的隶属度而被错误地划分到某一类中,故文献提出了可能性 C-Means 聚类算法(PCM, Possibilistic C-Means Algorithm)[2],这种算法克服了 FCM 算法的缺点,有效地解决了样本中包含噪音时聚类效果不好的问题。

本文首先介绍了 FCM 和 PCM 算法,并对两者的性能进行分析比较。通过仿真实验表明,FCM 算法对噪音数据敏感,不能很好地区分噪音和有效数据,而 PCM 算法能很好地解决噪音干扰的问题,具有更优的聚类效果。

1 模糊 C-Means 聚类算法

FCM 算法具有较高的映射精度和分类能力。它是通过最小化聚类准则函数 J_m 把 n 个向量 $x_k(k=1、2、\cdots、n)$ 分成 C 个类别来实现划分的,并求得每个类的聚类中心。其聚类准则函数为

$$J_m(U,V)=\sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik})^m(d_{ik})^2 \tag{1}$$

其中,N 为样本个数;C 为聚类中心数 $2\leqslant C\leqslant N$;$u_{ik}=u_i(x_k)$ 表示第 k 个样本属于第 i 类的隶属度,其要求满足 $\sum_{i=1}^{C}u_{ik}=1,\ \forall k$;$m$ 为权重指数,它的引入是对硬聚类准则函数的推广,表征模糊化程度。要实现模糊聚类就必须选定一个合适的 m,然而最佳 m 的选取目前尚缺乏理论指导。在实际应用中 m 的最佳取值范围为[1.5,2.5][3],推荐使用 $m=2$;v_i 是第 i 条规则输入空间的中心向量,d_{ik} 为第 k 个样本点到第 i 个聚类中心的距离。

$$d_{ik}=\|x_k-v_i\|,\ \forall i,k \tag{2}$$

$$u_{ik}=1/\sum_{j=1}^{C}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)} \tag{3}$$

$$v_i=\frac{\sum_{k=1}^{N}(u_{ik})^m x_k}{\sum_{k=1}^{N}(u_{ik})^m},\ \forall i \tag{4}$$

FCM 算法的实质是寻找一组中心矢量,使各样本到它的加权距离平方和达到最小。其算法步骤为:1)给定聚类类别数 C,样本个数 N,设定终止误差 ε,权重指数 m,最大迭代次数 Num,由随机数产生器产生分类中心矩阵 V,设置迭代计数器 $k=0$;2)用(2)式和(3)式初始化隶属度矩阵 $U\ (k=0)$;3)$k=k+1$,由(4)式更新聚类中心矩阵 V;4)用(2)

* 收稿日期:2004-12-14 修回日期:2005-04-11
资助项目:重庆师范大学研究项目(05XLY003)
作者简介:吕佳(1978-),女,四川达州人,讲师,硕士研究生,主要研究方向为人工智能、计算机网络。

式和(3)式计算隶属度矩阵 $U(k)$;5)如果 $\| U(k)-U(k-1) \| < \varepsilon$,则算法停止,进入下一步6),否则转到第3)步;6)输出隶属度类矩阵 U 和聚类中心矩阵 V。

由以上FCM可以看出,整个算法过程就是反复修改聚类中心和分类的过程。FCM算法本质上是一种局部搜索技术,它采用了所谓的爬山技术来寻找最优解,因此容易陷入局部极小值点。并且FCM算法对每个样本都有 $\sum_{i=1}^{C} u_{ik} = 1$ 归一化的约束条件,这使得样本的隶属度不但与该类的中心有关而且受其它类中心位置的影响,共享和为1的隶属度。图1所示为不含噪音的两个类的情形[2]。类1中和该类的聚类中心等距离的两个样本点A和B本应隶属度相同,但由于受到约束条件的限制,因此还要考虑同其它类的关系,样本点A和B到类2的聚类中心的距离不同,故隶属度不一样,从图1可以看出,按照(2)和(3)式,A的隶属度要大于B的隶属度。反之,隶属度相同的两个样本点A和C在类1中的实际位置也相差较远,因此不能真正表征样本属于该类的程度。图2所示为含有两个噪音点A和B的两个类的情形。样本点A和B的位置恰好处于两个类的中心线上,既然是噪音点,其隶属度理应偏小,但由于对隶属度作了归一化的限制,噪音A和B都被赋予了较高的隶属度。相比之下A比B更靠近两个类,依附于两个类的程度更高,隶属度应比B的隶属度偏大,归一化条件却使得A和B的隶属度都为0.5,无法反映出实际情况,从而加大了聚类的误差。

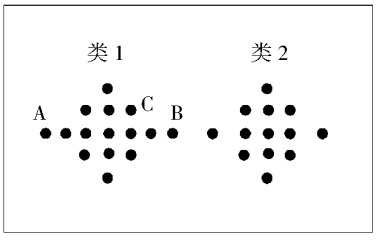

图1 不含噪音的两个类

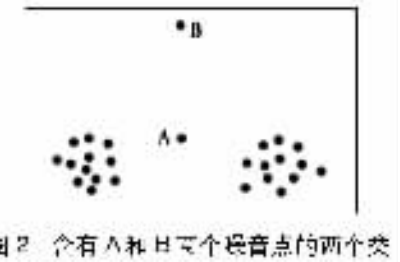

图2 含有A和B两个噪音点的两个类

2 可能性C-Means聚类算法

FCM算法中隶属度归一化的约束条件假定了每个样本的影响力是相同的[4],显然这与实际情况并不总是相符。针对FCM算法中该约束条件的缺陷,Krishnapuram和Keller于1993提出了可能性C-Means聚类算法。该算法放松了样本隶属度的约束,只要满足 $\max_i u_{ij} > 0$ 即可,隶属度不再是对1的共享或者划分,样本点 x_k 的隶属度 u_{ik} 仅表示其在第 i 个类内的典型性或其属于第 i 个类的概率,它只依赖于 x_k 与 v_i 的距离而与其它类中心的位置无关。而典型性正是在模糊集理论的应用中对隶属度最常用的解释。通过放松样本隶属度的约束,能够得到代表样本点特性的隶属度。通常情况下,噪音和孤立点都是代表性比较差的点,应用基于典型性的隶属度可以自动地降低它们的影响从而提高聚类的效果。将聚类准则函数改为

$$J_m(U,V) = \sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik})^m d_{ij}^2 + \sum_{i=1}^{C}\eta_i\sum_{k=1}^{N}(1-u_{ik})^m \quad (5)$$

式中第一项即为FCM的聚类准则函数,意义同FCM一致,即要求各数据到聚类中心的加权距离平方和尽可能小。而第二项要求各个 u_{ik} 尽可能大,以避免无效解。η_i 是惩罚因子,为一合适的正数,决定了聚类时属于一个类别的范围大小,其推荐取值为

$$\eta_i = K\frac{\sum_{k=1}^{N} u_{ik}^m d_{ik}^2}{\sum_{k=1}^{N} u_{ik}^m} \quad (6)$$

通常 K 取值为1。为使聚类准则函数值达到最小,隶属度的更新公式改为

$$u_{ik} = \frac{1}{1+\left(\frac{d_{ik}^2}{\eta_i}\right)^{\frac{1}{m-1}}} \quad (7)$$

从上式可以看出当 $\eta_i = d_{ik}^2$ 时,$u_{ik} = 0.5$,故 η_i 越大,在迭代时属于该类的点越多,这样对于聚类趋势不是很好,即样本点比较分散的类就能够有更大的范围去搜索该类的点,并最终稳定在样本点密集、聚类趋势好的类附近。隶属度 u_{ik} 只依赖于 x_k 与 v_i 的距离,而与其它类中心的位置无关。另外还要说明的是,权重指数 m 在FCM和PCM中的解释完全不同。在FCM中 m 的增加表示数据集中的所有点在类间的模糊性增加了,而在PCM中 m 的增加表示数据集中的所有点完全属于指定的一个类的概率增加

了,m 过大会导致聚类产生一致的类中心,m 一般取值 1.5。

PCM 本质上是一个穷举型搜索算法[5],每个类是独立于其它类的,一个样本点每次只能看到其所属类而不是所有类,故算法需要进行适当的初始化才能收敛到全局最小点。一般推荐用 FCM 算法的结果作为 PCM 算法的初始划分。其算法流程如下:1)确定聚类中心数 C,权重指数 m,最大迭代次数 Num 及迭代终止误差 ε,设置迭代计数器 $k=0$,并用 FCM 聚类的结果初始化其隶属度矩阵 $U(k=0)$ 和初始聚类中心矩阵 $V(K=0)$;2)按(2)和(6)式计算 η_i;3)按(7)式更新隶属度矩阵 $U(k)$;4)按(4)式重新计算聚类中心矩阵 $V(k)$;5)如果 $\| U(k)-U(k-1) \| <\varepsilon$,则算法停止进入下一步 6),否则转到第 3)步;6)输出隶属度矩阵 U 和聚类中心矩阵 V。

3　仿真实验

著名的 IRIS 数据是用来检验聚类算法和聚类有效性函数性能的标准样本集。该样本集由 150 个四维样本组成的,总共分成 3 个类别,每一个类别有 50 个样本。分别用 FCM 算法和 PCM 算法对 IRIS 数据样本集进行测试,FCM 的初始化部分中权重指数 m、最大迭代次数 Num 和终止误差 ε 分别取值为 2、5 000 次和 0.001。PCM 的 m、Num 和 ε 分别取值为 1.5、5 000 次和 0.000 1。每种算法的实验各进行 3 次,得到的聚类结果见表 1。表 1 分别列出了 FCM 和 PCM 在对 IRIS 数据样本集聚类之后的聚类中心。FCM 迭代次数 5 000 次,而 PCM 平均迭代次数为 38 次。FCM 的聚类结果作为 PCM 的初始划分,为其提供了准确的聚类。

表 1　IRIS 数据样本集聚类结果对比

算法	聚类中心
FCM	(5.88,2.76,4.36,1.39) (6.77,3.05,5.64,2.05) (5.00,3.40,1.48,0.25)
PCM	(5.01,3.41,1.18,0.25) (6.72,3.05,5.65,2.02) (5.91,2.76,4.42,1.38)

众所周知,在 IRIS 数据样本集中没有明显的孤立点。故为了比较两者对噪音的处理能力,人为地添加两个噪音点 A(0,0,0,0)和 B(8,8,8,8)。除了样本数变为 152 外,其余的参数仍然采用前面的初始值。表 2 列出了噪音点 A 和 B 在 FCM 和 PCM 中的隶属度。FCM 中噪音点 A 和 B 的的隶属度之和分别为 1,且隶属度较大。PCM 中 A 和 B 的隶属度之和都不超过 0.1。FCM 中对隶属度之和为 1 的条件非常严格,无论样本点的性质如何,必须要满足这一约束条件,故尽管 A 和 B 是噪音点,对聚类没有任何贡献,但是仍然要被赋予较大的隶属度,造成样本的错分。PCM 克服了 FCM 这一缺陷,解除了隶属度归一化的约束条件,每个样本点的隶属度仅与该类有关,代表其在某类的典型性,不涉及到其它的类别。这样,A 和 B 的隶属度就实际反映了它们对各个类的聚类中心的位置关系,恰当地体现出它们作为噪音其隶属度应有的特点。表 3 列出了在添加噪音点后的 IRIS 数据样本集上分别利用 FCM 和 PCM 所得到的新的聚类中心和算法准确率。

表 2　引入噪音点 A 和 B 后 FCM 和 PCM 计算得到的隶属度

算法	A	B
FCM	(0.30,0.21,0.50)	(0.33,0.45,0.22)
PCM	(0.01,0.02,0.04)	(0.01,0.01,0.02)

表 3　加入两个噪音点后的 IRIS 数据样本集聚类结果对比

算法	聚类中心	算法准确率/%
FCM	(5.89,2.77,4.38,1.42) (6.79,3.09,5.67,2.10) (4.98,3.39,1.48,0.26)	89.3
PCM	(5.01,3.41,1.19,0.25) (6.72,3.05,5.65,2.01) (5.91,2.75,4.42,1.41)	92.1

观察表 1 和表 3,可以看出 FCM 前后聚类中心有变化,这是因为增加了噪音样本点造成的。由于 FCM 算法中存在每个样本点对各个类的隶属度的和为 1 的限制,这样每个样本点的隶属度的确定不仅与其所属的类中心有关,而且还受其它类中心位置的影响,导致一些性质不是很好的点,如噪音被赋予了较高的隶属度,反过来影响了聚类中心的确定。而 PCM 前后聚类中心的变化幅度非常小,几乎可以看成没有变化。从表 2 就可看出 A 和 B 被赋予了极小的隶属度,故而对聚类中心产生的影响相对较小,聚类中心较为稳定。FCM 和 PCM 对第一类的聚类效果都很好,没有出现错分样本。表 3 中统计的算法准确率来源数据被错分的样本数全部出自于第

二类和第三类。对第二类和第三类的聚类效果PCM明显要优于FCM。IRIS的第二类和第三类之间数据有重叠,关系到聚类结果的有效性。这些相互交叉的样本点本应只与其所属的类中心有关,但由于受到隶属度归一化条件的约束,使得都要考虑受其它两个类的聚类中心位置的影响,一定程度上削弱了其真正属于该类的程度,从而妨碍了有效的聚类。而PCM放松了样本的归一化约束,并引入了惩罚因子 η_i,样本点 x_k 的隶属度 u_{ik} 仅表示其在第 i 个类内的典型性或其属于第 i 个类的概率,而与其它类中心的位置无关。故能更好地反映各样本点与每个类的隶属关系。聚类效果比FCM更好。

4 结论

本文利用FCM和PCM算法对数据聚类分析进行了研究比较。从理论上分析了PCM的可行性,从技术上进行了实际验证。FCM算法要求隶属度归一,造成对噪音数据敏感,聚类效果不好。而PCM算法突破了隶属度和为1的限制,能较好地处理噪音,且运算速度快,能收敛到全局最小值。通过仿真实验表明,PCM比FCM算法在聚类分析中具有更加优越的性能。

参考文献:

[1] BEZDEK J C. Pattern Recognition with Fuzzy Objective Function Algorithms[M]. New York:Plenum Press,1981.

[2] RAGHU K,JAMES M K. A Possibilistic Approach to Clustering[J]. IEEE Trans on Fuzzy Systems,1993,1(2):98-110.

[3] PAL NR,BEZDEK J C. On Cluster Validity for the fuzzy C-Means model[J]. IEEE Trans. on Fuzzy Systems,1995,3(3):370-379.

[4] 张敏,于剑. 基于划分的模糊聚类算法[J]. 软件学报,2004,15(6):858-868.

[5] YANG M S,WU K L,YU J. A Novel Fuzzy Clustering Algorithm[A]. In Proc. of the 2003 IEEE Int'l Symp. On Computational Intelligence in Robotics and Automation [C]. Kobe,Japan:IEEE,2003. 647-652.

(责任编辑 游中胜)