一种新的聚类分析算法

何金国 石青云
(北京大学视觉与听觉信息处理国家重点实验室, 北京 100871)

摘 要 给出了一种新的无监督聚类算法,但这种算法并非是基于目标函数的聚类算法,而是对数据直接设计一种迭代运算,以使数据在保持类特征的情况下进行重新组合最终达到分类的目的. 通过对一类数据的实验表明,该算法在无监督给出类数方面具有较好的鲁棒性;另外,该算法在数据的准确归类、无监督聚类、确定性,以及对特殊类分布的适用性等方面均优于 HCM 和 FCM 算法.

关键词 HCM 算法 FCM 算法 无监督聚类 归类
中图法分类号: TP391.41 **文献标识码**: A **文章编号**: 1006-8961(2000)05-0401-05

A New Algorithm for Clustering Analysis

HE Jing-guo, SHI Qing-yun
(*Center for Information Science, Peking University, Beijing* 100871)

Abstract In this paper, a new algorithm for unsupervised Clustering analysis is proposed, through a new kind of iterative activation the examples of a cluster are moved inside to the center of the gravity of the cluster together. Through this method correct number of clusters could be got. Because each sample moves only in its own cluster while iterating, we can correctly tell which cluster a sample should belong to. The experiments show that the new algorithm has better results in several aspects than HCM and FCM algorithms, such as unsupervised clustering, correct clustering, clustering capability for special data which HCM and FCM algorithms can not cluster. The new algorithm is an unsupervised clustering algorithm but HCM and FCM algorithms need correct number of clusters before iterative activation.

Keywords HCM, FCM, Unsupervised clustering, Iterative activation

0 引 言

K-均值算法是常用的一种基于目标函数的聚类算法,它包括硬 K-均值算法(HCM)和模糊 K-均值算法(FCM). 事实上,HCM 算法是 FCM 算法的一个特例. HCM 算法和 FCM 算法是基于目标函数的算法,由于它们的目标函数存在许多局部极小点[1],如果初始化落在了一个局部极小点附近,就会造成算法收敛到局部极小. 另外,HCM 算法和 FCM 算法都隐含了这样一个归类假设,即判断点 X 属于哪一类(或者最可能属于哪一类)就看它离哪一个聚类中心距离最近. 有时这种假设是不符合实际的. 因此即使求得目标函数的全局最小值也可能不是正确的聚类结果,文献[2]给出了一种解决这一问题的方法,即定义不同的距离测度,但这种方法对这一问题的帮助是有限的. 另外,在实际应用中,例如图象分割中,我们希望无监督给出聚类类数,同时还希望给出正确的归类结果. 基于这种想法,我们设计了一种新的聚类算法,但它不是基于目标函数的. 该算法的主要思想是通过迭代使得数据向类内收缩,从而达到分类的目的. 由于收缩过程可以用很少的内存记录下来,所以迭代完成后,准确的归类结果也有了. 通过对一类数据的试验表明,该算法不仅在无监督

基金项目: 973 项目课题(G1998030606)
收稿日期: 1999-03-09; 改回日期: 1999-10-28

给出类数方面具有较好的鲁棒性,而且在数据的准确归类、无监督聚类、确定性、对特殊类分布的适用性等方面均优于 HCM 和 FCM 算法. 其中,确定性是指该算法的运行结果,不会象 HCM 和 FCM 算法那样因初始值不同而变化,而是能得到确定的结果;另外,有一类数据(例如螺旋形分布) HCM 和 FCM 无法聚类,而新算法则能无监督聚类,我们把这一优越性称为对特殊类分布的适用性.

1 新聚类算法的步骤和解释

本算法的目标是求出每个类重心 $G_i(i=1,2,\cdots,K)$、类数 K,并给出归类结果.

对一数据集 $d=\{x_1,x_2,\cdots,x_n\}(D\subset \mathrm{R}^n)$ 做聚类分析:假设 D 的每一样本点 $x_i(i=1,2,\cdots,n)$ 为一个质点,重量为 1;为了便于计算机处理,假设 x_i 的坐标都为非负整数,并假设 D 中存在 K 类,其类重心为 $G_i(i=1,2,\cdots,K)$,记类为 $C_i(i=1,2,\cdots,K)$.

新算法一共分为 4 步,即迭代、选初始类、查表、合并.

1.1 迭代

迭代中,δ 初始置为一个很小的值,$\delta_{\max}$ 为一个给定的大于 δ 的值. α 为一个大于 1,小于 2 的值.

第 1 步 对数据集 D 中的任何一个样本 x_i,求其邻域 $U(x_i,\delta)$ 中的样本数,记作 N_i.

第 2 步 对数据集 D 中的任何一个样本 x_i 的邻域 $U(x_i,\delta)$,令每一个 $x_j\in U(x_i,\delta)$ 的质点重量为 $\tilde{m}_j=\frac{m_j}{N_j}$,并构成一个新系统 $\tilde{U}(x_i,\delta)$,求出 $\tilde{U}(x_i,\delta)$ 的重心 $\tilde{x}$ 及重量 $\tilde{m}$.

第 3 步 由所有的 $\tilde{x}$ 及 $\tilde{m}$ 组成新的数据集 $\tilde{D}$,对 $\tilde{D}$ 进行迭代:如果$\exists\ \tilde{x}_i=\tilde{x}_j$,则合并为一个质点 $\tilde{x}$(重量 $\tilde{m}=\tilde{m}_i+\tilde{m}_j$). $\tilde{D}$ 就是 D 经过一步迭代得到的数据集合. 令 $\delta=\delta\times\alpha$,如果 $\delta>\delta_{\max}$,则 $\delta=\delta_{\max}$,令 $D=\tilde{D}$,重复第 1 步.

本算法可以迭代到数据集 D 的总样本数不再减少,或出现重量比较大的质点,即 $m>\theta_m$(θ_m 为阈值)时停止.

新的聚类算法希望避开聚类中类与类之间互相干扰的问题,第 1 次迭代时,因为 δ 的初值很小,迭代时,类之间互相不干扰,而且类边缘点 x 的邻域 $U(x,\delta)$ 的重心总是偏向类内,所以 $\tilde{D}$ 相对于 D 而言,类出现收缩,收缩的结果必然导致“迭代”第 3 步的合并. 而且,由于类出现收缩,下一步迭代中 δ 可以增大($\delta=\delta\times\alpha$),当然,为了保证类与类之间不互相干扰,$\delta$ 不能太大($\delta\leq\delta_{\max}$). 如果把每一个类作为一个系统,则“迭代”的第 2 步就保证了在迭代过程中的重心位置和重量保持不变,这样一来,从宏观上看就如同给每个质点一个指向重心的“合力”,令质点向类重心方向漂移.

1.2 选初始类

当迭代结束后,则需找出质点重量明显偏大的收缩中心,方法是:将迭代结束后得到的质点按重量从大到小排列为 $m_1>m_2>\cdots>m_{n'}$,定义 $q_i=m_i/m_{i+1}$,$(i=1,2,\cdots,n'-1)$,求 $i'=\min\limits_{q_i>K_G} i$,$K_G$ 为事先给定的一个阈值,则 i' 即为所求的初始类数,与 $m_1,m_2,\cdots,m_{i'}$ 相应的 $x_1,x_2,\cdots,x_{i'}$ 为所求的初始类重心的位置.

若类不为超球形时,则初始类不一定为一个真正的类.

1.3 查表

定义每一个 $U(x,\delta)$ 的重心 x' 为 x 的漂移位置,并记录下这一漂移位置,经过一次迭代后,D 中所有的质点都有了一个漂移位置,则构成一个漂移位置表,以后的每次迭代中需不断更新漂移位置表,迭代结束后即可通过查找漂移位置表,将漂移到 $x_i(i=1,2,\cdots,i')$ 的质点归入第 i 初始类.

1.4 合并

假设通过新的算法迭代结束后,得到的与 $m_1,m_2,\cdots,m_{i'}$ 相应的 $x_1,x_2,\cdots,x_{i'}$ 为所求初始类的重心位置,将查找漂移位置表归类后得到 i' 个初始类,记为 $C_i'(i=1,2,\cdots,i')$,下面给出“密度”合并算法,算法中 δ_θ、V_θ 为阈值,算法的目的是判断 C_i' 和 C_j' 是否合并:

(1) 对所有的 $x\in D$,求出所有满足边界条件($\exists\ x_i,x_j\in U(x,\delta_\theta),x_i\in C_i',x_j\in C_j'$)的 $U(x,\delta_\theta)$;

(2) 合并第 1 步得到的所有 $U(x,\delta_\theta)$ 为一个边界集 U;

(3) 求 U、C_i' 和 C_j' 的密度,把 U 的密度与 C_i' 和 C_j' 两个初始类的平均密度中较小的一个比较,不妨设 C'_i 的平均密度较小,如果 $\frac{C'_i\text{ 的密度}}{U\text{ 的密度}}<V_\theta$,则合并初始类 C_i' 和 C_j'.

经过合并后的初始类才是所求的类.

算法完成后,就得到了每个类重心 $G_i(i=1,2,\cdots,K)$、类数 K 及归类结果.

2　实验结果和分析

新的聚类算法通过迭代运算, 使类数据收缩成为一个高重量的质点, 从而达到分类的目的, 实验表明, 对于超球形分布的数据, 一般每个类都能收缩到一个质点上; 对于狭长形的类, 有时一个类会收缩到两个以上的高重量质点. 新的聚类算法第 4 步——“密度”合并法的作用就是合并这样一个出现两个以上高质量质点的类. 因为新的聚类算法通过“迭代”、选“初始类”和查表达到以下目的:

(1) 通过查表可确定每个数据归于哪个初始类;

(2) 尽管有时该算法一个类会收缩到两个以上的高重量质点, 但与 FCM 算法陷入局部极值不同, 其收缩到的高质量质点位于类的局部密度极大值处. FCM 算法陷入局部极值时, 类中心会位于如彩色图版I 图 5的两类之间(或彩色图版I 图 6(a) 的两类之间) .

所以, 我们既知道初始类具体的归类结果, 又知道每一个初始类不会跨越两个类, 而只会出现一个由几个相互相连的初始类组成的类, 这样, 就为合并提供了充分的条件, 本文的密度合并法只是一种简单的可行办法. 因为不在一个类中的初始类要么不相连, 要么相连处的密度明显比类内密度小, 否则高密度相连的两类应该属于一个类.

为了实验简便起见, 只选用二维数据集做输入数据, 且对每一个输入数据集, 先用新的算法的第 1 步促成类收缩, 其迭代结束条件为数据集 D 的总样本数不再减少, 然后用算法的第 2 步选出大的质点(初始类) , 并在图中用一小正方形块表示, 第 3 步查找漂移位置表划分初始类, 并将不同的初始类着上不同的颜色, 第 4 步用合并法合并, 将可以合并为一类的收缩质点用线段连接起来, 这样, 对每一个二维输入数据集, 均可得到一个新算法聚类结果图, 从图中可以一眼看出归类结果的好坏以及合并结果的好坏.

为了方便讨论, 先对不同形状数据集进行聚类实验(参数选为 $\delta=1$, $\delta_\theta=3$, $V_\theta=2.0$, $K_G=3$, $\alpha=1.5$, $\delta_{max}=12$) , 然后再对 5 个参数变化时进行聚类实验和讨论, 并与 FCM 算法的实验结果作对比.

2.1　对不同形状数据集的实验结果

(1) 对类形状为球形的二维数据集做实验, 如彩色图版I 图 1, 该数据集是从文献[4] 中扫描得到, 新算法第 1 步迭代结束后, 得到的收缩质点按重量大小排列如表 1.

表 1

质点	x_1	x_2	x_3	x_4	…
重量	610.5	590.1	583.6	10.2	…

所以, 根据新算法第 2 步很容易选出如彩色图版I 图 1 中小正方形块所标记的 3 个大质点, 由图中结果可以看出, 对于这种类形状为圆形(多维情况下为球形) 的数据集, 新算法可以省略“密度”合, 并初始类与最终分类一致.

彩色图版I 图 2 的数据是在彩色图版I 图 1 中加入一个圆形干扰类得到的. 聚类结果: 初始类与最终分类一致. 新算法可以省略“密度”合并, 而且图中 x_3 和 x_4 类没有出现互相干扰, 与设计该算法的初衷相吻合.

(2) 对于形状不是超球形的数据集进行聚类实验, 如彩色图版I 图 3, 两个形状不为圆形的类初始类与最终分类不一致, 但经过新算法第 4 步合并后可得到准确的分类结果. 图中 x_4 和 x_5 类形状不规则, 一个类出现两个以上的初始类, 不过, 初始类重心一般出现在类密度的局部极大值处, 而且也有初始类的具体归类结果(如图不同颜色属于不同“初始类”) , 因而为合并创造了条件.

(3) 对螺旋形数据集进行聚类实验

如彩色图版I 图 4 所示, 两个螺旋形的类分布是用 Photoshop 手工描点产生的, 从彩色图版I 图 4 可见, 聚类结果准确. 螺旋形的类分布是聚类分析中一个很难的课题, FCM 和 HCM 算法无法均对这种分布聚类.

2.2　6 个参数变化时的聚类实验与讨论

参数 δ 是第 1 步“迭代”的邻域半径, 因为是离散数据, δ 只能取正整数, 且迭代过程中, δ 以 $\delta=\delta\times\alpha$ 的速度不断增大. 当 δ 取 1 或较大的数时, 只对迭代的次数, 即对算法的速度有影响, 而对聚类结果基本无影响, 但 δ 初值不能过大, 否则类之间数据会相互干扰; 若类之间距离较大, 则 δ 初值可以取大一些, 这样才有利于算法速度加快, 如果类之间距离无法估计, δ 取 1 总是可行的. 实验证实确实如此.

参数 K_G 反映的是收缩的高重量初始类重心与非初始类重心的比值, 从表 1 可以看出, 高重量初始类重心与非初始类重心是很容易区分的. 由所做的聚类实验结果可见, K_G 取2.5~ 20都是可行的, 而且对聚类结果无影响. 甚至表 1 的数据 K_G 取 2~ 50, 聚类结果都无变化. 参数 K_G 实际上反映的是初始

类与非初始类小质点的重量之比, 如果聚类有特殊要求, 例如希望去除一些相比之下太小的类, 可以使用较小的 K_G(例如 1.5).

参数 α 取一个大于 1, 小于 2 的数即可, α 用于控制 δ 的增大速度, 从而进一步控制算法的速度, α 之所以取大于 1, 小于 2 是根据每次迭代所能达到的类收缩量制定的. 实验中 α 取 1.5.

参数 δ_θ、V_θ 是用于"密度"合并的, 其中 V_θ 是指初始类密度与初始类相连处的密度之比不能太大, 实验显示, 一般同视觉上相吻合的 V_θ 为 1.5~ 2.0.

参数 δ_θ 太小时, 聚类数据稀疏, 使 $U(x, \delta_\theta)$ 邻域内多数样本 x 成为孤立点, 对我们所做的实验数据类型, 参数 δ_θ 可以简单地取成 $\max(\delta, 3)$. 为增强算法适应性, 参数 δ_θ 可以用如下方法自适应确定: δ_θ 从 1 开始逐渐增加, 当 δ_θ 使大多数样本的 $U(x, \delta_\theta)$ 邻域内包含有两个以上的样本时即为所求.

参数 δ_{max} 越大, 初始类出现越少, 大到一定程度, 即使是狭长分布的类也不会出现两个以上的初始类, 但太大会使所有的类合并为一体; δ_{max} 越小, 初始类出现越多, 实际使用时, 可取 δ_{max} 为估计的类重心间的最小距离.

综合起来, 一个缺省的取值方案如下:

δ_{max}= 类重心间的最小距离(估计), 可小于类重心间的最小距离; δ= 1; α= 1.5; δ_θ 可用如上所述的自适应算法求得. 实验显示, δ_θ 小范围改动对聚类性能影响不大;

V_θ、K_G 两个参数是算法为了适应不同聚类要求而设的, 如前所述, V_θ 是指初始类密度与初始类相连处的密度之比, 对彩色图版I 图 2 的 x_3 类和 x_4 类, 当 V_θ 大于 3 时就合为一类了, 事实上, 对于 x_3 和 x_4 这样的两类, 有时的确需要作为一类看待, 所以 V_θ 是界定两类之间相连部分松散程度("密度")的一个值, 必须由实际应用决定, 实验显示, 与视觉上一致的 V_θ 值在 1.5~ 2.0.

K_G 的情况与 V_θ 类似, 对彩色图版I 图 6(a) 中的 x_1 的类, 特殊情况下可不作为一类, 此时可以通过缩小 K_G 的办法来实现. 前面的讨论显示, 即使 K_G 在很大范围内变化也对聚类结果都无影响. 因为收缩的初始类重量与少数离散的孤立点重量之比差别是很明显的.

通过以上算法的讨论和实验可知, 之所以有这么多参数正是由于实际复杂聚类的需要. 由于新的聚类算法可以分为收缩和合并两个相对独立的部分, 每一部分的参数选择可以独立研究. 另外, 由于本算法迭代时, 数据只在类内重新组合, 类间数据互不影响, 因此只要对几种分布聚类成功, 则将这几种数据分布合在一起进行排列组合, 也一定能聚类成功, 这一特点大大简化了实际应用时的参数训练工作. 本文给出的聚类实验用参数都是 δ= 1, δ_θ= 3, V_θ= 2.0, K_G= 3, α= 1.5, δ_{max}= 12, 且聚类结果都很好; 若把本文的这些类数据合在一起排列组合也一样能聚类成功. 由于迭代中类间数据互不干扰, 所以聚类成功与否由主要类内的形状、分布决定, 故将研究聚类的最小单位定为一个类, 从而简化了研究过程.

2.3 与 HCM 和 FCM 及其它算法的对比

HCM 和 FCM 算法是基于目标函数的聚类算法, 因而容易陷入局部极小; HCM 和 FCM 算法隐含如下假设, 即判断点 X 属于哪一类(或者最可能属于哪一类)就看它离哪一个聚类中心距离最近, 但这种假设在处理实际数据是很难满足, 因而聚类后数据的准确归类就有困难; HCM 和 FCM 算法需要有监督给出聚类类数. 类数 K 的确定问题长期以来一直是在聚类分析中一个研究课题, 实践中也经常需要一种聚类算法对输入数据集能无监督地给出准确的分类结果. 在 K 均值算法(HCM 和 FCM)中, 选择合适的 K 是很困难的, 而且选择不好, 对聚类结果影响很大[3,4]. 通过新的聚类算法迭代运算使类内数据收缩到一起, 而类之间数据迭代中互不干扰, 从原理上避免了 HCM 和 FCM 算法聚类的不确定性, 由于收缩过程可以记录, 且数据的分类不是通过最近距离法, 而是通过查找漂移位置表来进行; 另外, 还由于类之间数据互不干扰, 数据只在类内漂移, 因而通过查表就能实现数据准确分类. 实验显示, 新的聚类算法具有较好的鲁棒性. 此外 HCM 和 FCM 算法在类数给定值与实际不符时, 对聚类结果影响很大, 将无法得到正确的聚类结果.

我们用彩色图版I 图 2 的数据做聚类实验, 对比 FCM 算法和新的聚类算法的性能, 发现 FCM 算法以大约以 0.4 的概率陷入局部极小(如彩色图版I 图 5).

对彩色图版I 图 4 的螺旋形分布数据, FCM 算法无法聚类. 这是由 FCM 算法的原理决定的; 彩色图版I 图 7是 FCM 算法在 K= 2 时的聚类结果, 实际上, 对于这种分布, 无论两个聚类中心落在何处, 都无法作为正确的聚类结果; 可是新算法用于对这种数据聚类, 其结果却很令人满意(彩色图版I 图 4).

对彩色图版I 图 6 的数据, 新的聚类算法无监督即能准确聚类, 而 FCM 算法有监督也不能聚类, 其运行 15 次, 有 14 次结果如彩色图版I 图 6(a) 所示. 原因是 FCM 算法的全局极小在这种情况下不是解.

彩色图版I 图 5 和图 6 聚类用的参数仍然是 $\delta = 1, \delta_\theta = 3, V_\theta = 2.0, K_G = 3, \alpha_r = 1.5, \delta_{max} = 12$.

3 总 结

本文给出了一种新的无监督聚类算法, 即通过迭代使数据重新组合得以方便分类, 由于迭代过程中, 类内数据收缩到一点, 而类间数据互不干扰, 从而可以通过记录类数据收缩过程来达到准确分类. 另外, 由于算法迭代时类间数据互不干扰, 因此只要对几种分布聚类成功, 则这几种数据分布合在一起进行排列组合, 也一定能聚类成功, 这一特点大大简化了实际应用时的参数训练工作. 由于聚类成功与否由类内形状、分布决定, 故研究聚类的最小单位定为一个类, 从而简化了研究过程. 通过参数的讨论和实验证明, 实际应用时, 参数的确定是比较容易的, 可以通过实际聚类要求方便地训练参数. 由于实际应用复杂多变, 有时要求忽略太小的类, 有时要求将松散相连的两类(低密度相连) 合成一类; 有时聚类数据非常复杂(例如螺旋形数据), 之所以有这么多参数正是为了适应实际复杂聚类的需要. 该算法在无监督聚类、准确归类、适用范围、确定性等方面均优于 HCM 和 FCM 算法.

参 考 文 献

1 Selim S Z, Ismail M A. K-MeansIType Algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Trans Pattern Analysis and Machine Intelligence, 1984, PAMI-6(1): 81~ 87.

2 Erigui II, Krishnapuram R. Clustering by competitive agglomeration. Pattern Recognition, 1997, 30(7): 1109~ 1119.

3 边肇祺等编著. 模式识别. 北京: 清华大学出版社, 1986.

4 Lei Xu. Rival penalized competitive learning for clustering analysis, RBF Net, and curve detection. In IEEE Trans. On Nearal Networks, 1993, 4(4): 636~ 649.

何金国 1995 年毕业于北京大学数学系, 现为北京大学数学学院信息科学系博士研究生. 研究方向为模式识别、医学图像处理等.

石青云 中国科学院院士, 1957 年毕业于北京大学数学力学系. 现任北京大学数学科学学院教授, 应用数学专业博士生导师, 北京大学信息科学中心学术委员会主任, 国际模式识别委员会理事. 研究方向为模式识别、图象数据库、图象数据压缩等.

通　　告

为适应我国信息化建设需要, 扩大作者学术交流渠道, 本刊已加入《中国学术期刊(光盘版)》和"中国期刊网". 作者著作权使用费与本刊稿酬一次性付给. 如作者不同意将文章编入该数据库, 请在来稿时声明, 本刊将做适当处理.