

一种模糊加权的改进层次聚类算法研究

李剑英¹, 丁世飞^{1, 2}, 徐 丽¹, 钱 钧¹

(1 中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116;

2 中国科学院 计算技术研究所智能信息处理重点实验室, 北京 100080)

摘 要: 传统层次聚类算法中经常会遇到合并点和分裂点选择的问题, 一旦一组对象被合并或者分裂, 下一步的处理将在新生成类上进行, 已做处理不能撤销, 这样有可能导致低质量的聚类结果. 针对这个问题, 文中提出了一种模糊加权层次聚类改进算法, 每次分层聚类时先计算对象属于这个类可靠度, 然后和阈值进行比较, 当可靠度小于阈值时重新确定对象的归属类, 这样就解决了上述问题. 最后通过实验验证, 该算法确实可行有效.

关键词: 层次聚类; 模糊加权; 可靠度; 阈值

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2011)09-0210-04

An Improved Hierarchical Clustering Algorithm Based on Fuzzy Weighted

LI Jian-ying¹, DING Shi-fei^{1, 2}, XU Li¹, QIAN Jun¹

(1 School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China;

2 Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Science, Beijing 100080, China)

Abstract: There is always the problem of choosing the merge point and split point in traditional hierarchical clustering algorithm. Once a group of objects are merged or split, the next step of the process will be done on the new category, in addition, the already done processing can't cancel, which could result in low quality clustering results. According to this problem, this paper proposes a weighted fuzzy hierarchical clustering algorithm. Every time when we stratify and cluster, we first calculate the reliability of the object to the category, and then compare it with the threshold. If the former is less than the latter, once again we determine its belonging category, which will solve the above problem. In the end, the experimental results show that this algorithm is feasible and effective.

Key words: hierarchical clustering; fuzzy-weighted; reliability; threshold

1 引言

随着数据挖掘研究领域技术的发展, 作为数据挖掘主要方法之一的聚类算法, 也越来越受到人们的关注. 聚类就是按照事物的某些属性, 把事物聚成类, 使类间的相似性尽可能小, 类内相似性尽可能大^[1]. 聚类方法有很多种, 例如基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法、

基于模型的方法等等. 层次聚类分析由于其强壮性和对于输入记录的顺序不敏感及简单易用而成为聚类分析中的一个广泛应用的算法.

基本的层次聚类方法是由 Kaufman 和 Rousseeuw 提出的聚类方法 AGNES 和分裂方法 DIANA^[2]. 这种聚类方法虽然简单, 但经常会遇到合并点和分裂点选择的问题, 而且这种方法不具有很好的可伸缩性. 因此人们提出众多改进的层次聚类算法.

收稿日期: 2011-01-13; **修回日期:** 2011-02-01

基金项目: 国家自然科学基金项目(60975039); 江苏省自然科学基金项目(BK2009093); 中国科学院计算技术研究所智能信息处理重点实验室开放基金(2010-1)

Zhang, Ramakrishnan 和 Linvy 提出的 BIRCH 是一种综合的方法^[3]. Guha, Rastogi 和 Shim 提出的 CURE 算法可以识别复杂形状和大小不同的类,而且能很好地过滤孤立点^[4],之后他们在此基础上又提出了使用于分类数据的 ROCK 方法^[5]. Karypis, Han 和 Kumar 基于对 CURE 和 ROCK 缺点的观察,提出了采用动态模型的 Chameleon^[6]方法,在发现高质量的任意形状的聚类方面有更强的能力.之后大量学者提出了各种改进算法^[7-10]. 这些算法虽然对层次聚类的能力和效率上有了很大的提高,但是对于层次聚类中经常碰到的关于合并点和分裂点选择的问题,都没有提出一个很好的解决方法. 这个问题是这样的,层次聚类过程中,一旦一组对象被合并或者分裂,下一步的处理将在新生成类上进行,已做处理不能撤销,类之间也不能交换对象. 这样的话,如果某一步在做出分裂或合并的决定事没处理的很好,那么可能导致最终的低质量的聚类效果. 针对这个问题,文中在模糊集合思想上提出了一种模糊加权改进层次聚类算法.

2 传统层次聚类算法分析

根据聚类是自低向上还是自顶向下逐层聚类,层次聚类方法可分为凝聚的和分裂的层次聚类方法. 其中凝聚的层次聚类方法应用更多,该方法首先将每个对象作为一个族,然后合并这些对象成为越来越大的类,直到所有对象都在一个类中或者达到用户定义的希望得到的类的数目为止. 常用的层次聚类方法都属于这一类,只是在类的相似度的定义上有所不同. 常用的几种类间距离度量方法有:最小距离 $d_{\min}$,最大距离 $d_{\max}$,平均值距离 d_{mean} ,平均距离 d_{avg} 等.

假定有 N 个对象要被聚类,例如采用最小距离凝聚的层次聚类方法的基本过程如下:

(1) 初始时,将每个类作为一个对象,计算它们之间的最短距离 d .

(2) 将距离最近的两个类合并为一个新的类.

(3) 重新计算新的类与类之间的距离.

(4) 重复(2)(3)步骤,直到所有对象都在一个类中或者达到用户定义的希望得到的类的数目为止.

从上述处理过程可以看出,两个对象 a 和 b 一旦确定同属于类 1,而第二次聚类时类 1 属于类 2,那么这两个对象 a 和 b 也一定属于类 2. 这样的话,如果某一步在做出分裂或合并的决定时没处理的很

好,那么可能导致最终的低质量的聚类效果. 为此,文中进行以下改进.

3 模糊加权聚类

传统的层次聚类分析是一种硬件划分,它把每个对象严格地划分到某个类中,这种分类的类别界限是分明的. 而在实际中大多数对象并没有如此严格的属性,它们在形态和类属方面存在中介性,适合进行软划分,模糊集合理论的提出为解决这一问题提供了有力的工具.

3.1 基本理论

若对论域(研究的范围) U 中的任一元素 x ,都有一个数 $A(x) \in (0,1)$ 与之对应,则称 A 为 U 上的模糊集, $A(x)$ 称为 x 对 A 的隶属度. 当 x 在 U 中变动时, $A(x)$ 就是一个函数,称为 A 的隶属函数. 隶属度 $A(x)$ 越接近于 1,表示 x 属于 A 的程度越高, $A(x)$ 越接近于 0 表示 x 属于 A 的程度越低. 用取值于区间 $(0,1)$ 的隶属函数 $A(x)$ 表征 x 属于 A 的程度高低^[11]. 隶属度属于模糊评价函数里的概念,其特点是评价结果不是绝对地肯定或否定,而是以一个模糊集合来表示.

3.2 对象属于第一层类的确定

例如在二维平面上有 n 个点 $(a_1(x_1, y_1), a_2(x_2, y_2), \dots, a_n(x_n, y_n))$,确定其聚类情况.

确定对象属于某个第一层类的方法和传统类似,例如用传统方法确定出有 t 个点 $a_1(x_1, y_1), a_2(x_2, y_2), \dots, a_t(x_t, y_t)$ 属于类 F_j . 然后计算这个类的类中心. 类中心 $O_j(m, n)$ 可以这样确定:

$$\begin{aligned} m &= \frac{x_1 + x_2 + x_3 + \dots + x_t}{t} \\ n &= \frac{y_1 + y_2 + y_3 + \dots + y_t}{t} \end{aligned} \quad (1)$$

然后计算这个类中每个点距离这个类中心的距离:

$$d_i = |O_j - a_i| = \sqrt{(x_i - m)^2 + (y_i - n)^2} \quad (2)$$

最远距离: $d_{\max} = \max(d_1, d_2, \dots, d_t)$.

定义 1 模糊可靠度:

$$B_{ij} = \frac{d_{\max} - d_i}{d_{\max}} + \frac{1}{2} \quad (3)$$

B_{ij} 表示类 F_j 中每个点属于这个类的可能性(加 $\frac{1}{2}$ 是为了避免可靠度为零,使以后计算方便).

这样当第一次聚类的时候,除了确定哪些点属于这个类以外,还确定了这些点属于这个类的模糊

可靠度.

3.3 对象属于第二层类的确定及对象合并后再分裂的确定

定义 2 可靠度累加方法:

$$B_{ik} = B_{ij} \times B_{jk} \quad (4)$$

其中, B_{ij} 表示 a_i 点属于 F_j 的可靠度, B_{jk} 表示类 F_j 属于类 S_k , 那么 B_{ik} 表示 a_i 属于类 S_k 的可靠度. 这样我们由定义 2 可以得到点 a_i 属于类 S_k 的可靠度.

定义 3 阈值的设定:

$$L = p \cdot (0.5)^e \quad (5)$$

其中, L 是设置一个阈值, e 是聚类的层数, p 是一个经验系数, 是一个大于 1 的数.

定义阈值 L 的目的就是来监督可靠度的. 用传统方法再次聚类得到第二层聚类结果, 把它作为预聚类结果. 然后计算每个点属于预定类的可靠度, 当可靠度不小于阈值 L 时就按预备聚类结果来处理对象的归属; 当可靠度小于阈值 L 时要重新计算这个点离其他类中心的距离, 来确定应该归属于哪一类. 这样就可以解决对象合并后的再分裂的问题了.

往后的再次聚类方法和这次的完全相同.

3.4 加权

而实际情况一般是越往上层, 这层的可靠度在总可靠度中起的决定性就越大, 例如第二层到第三层的可靠度在总的可靠度中起的决定性要大于第一层到第二层得可靠度. 为了体现这个现象我们在可靠度累加计算过程中给每个可靠度加权.

定义 4 可靠度加权累加:

$$B_{ik} = q_1 \times B_{ij} \times q_2 B_{jk} \quad (6)$$

其中 q_1, q_2 当是人为的数值, 但为使加权前后可靠度不变, 权值相乘必须等于 1 (要进一步讨论的话, 其实权值是和类间距有关系的一个值, 具体情况以后再讨论).

这样在可靠度累加计算时, 可以采用这种加权累加方法, 使聚类结果更合理.

3.5 改进后的算法描述

改进的算法处理步骤如下:

(1) 根据传统方法进行第一次分类, 然后确定每个类的中心(式 1)以及对象的可靠度(式 3).

(2) 依据传统方法对结果再次分类, 再次计算类中心和对象的属于预定类的可靠度(式 6), 比较可靠度和阈值(式 5)的大小, 小于阈值的对象要重新计算来确定类的归属.

(3) 重复(2), 直到所有对象都在一个类中或者达到用户定义的希望得到的类的数目为止.

4 实验验证

实验 1 所用数据是人为生成的数据两个三行三列的等距矩阵. 其聚类结果应该是分为两类, 并且为矩形.

用最小距离凝聚的层次聚类来处理数据, 过程和结果如图 1 所示.

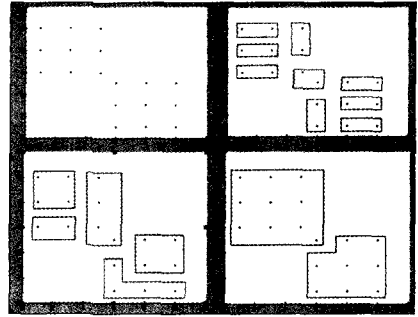


图 1 最小距离凝聚的层次聚类过程及结果

由图 1 可以看到, 两个矩阵相邻两个角上的两点距离最近, 按最小距离凝聚的层次聚类方法, 第一次聚类两点就会分在同一类, 再次聚类的话这两点必然还在一类, 最终导致了如图所示的结果. 而实际上这两点最终应该分在两个不同的类.

用模糊加权层次聚类来处理, 过程和结果如图 2 所示.

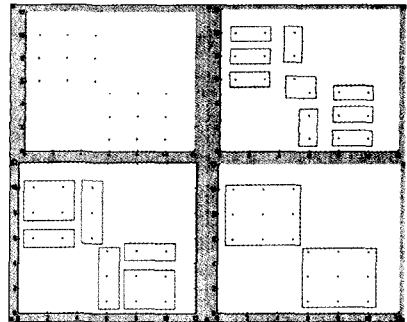


图 2 模糊加权层次聚类过程及结果

由图 2 可以看到, 两个矩阵相邻两个角上的两点第一次聚类是分在了一个类中. 但是第二次聚类时, 这个类和其他任何类合并, 都会发现有个点距离要合并成的这个类的类中心很远, 即这个点属于要合并成的这个类的可靠度很低, 所以重新确定这个点属于哪个类. 这样就可能使原来的两个点处于不同的类中了.

实验 2 选取专门用于测试分类、聚类算法的国际通用的 UCI 数据库中的 IRIS 数据集. IRIS 数据集包含 150 个样本数据, 分别取自三种不同的鸢尾属植物 *setosa*、*versicolor* 和 *virginica* 的花朵样本.

实验结果如表 1 所示.

表 1 两种层次聚类算法的实验对比结果

聚类方法	聚错	运行时间	平均准确率
	样本数	/s	/%
最小距离凝聚的层次聚类	51	0.138	66
模糊加权层次聚类	19	4.85	87

通过实验结果表明,模糊加权层次聚类确实能有效解决对象的合并后再分裂的问题,使聚类结果的准确率大幅提高.

5 结束语

文中针对传统的层次聚类算法中对象一旦合并或者分裂下一步的处理将在新生成类上进行已做处理不能撤销这个问题,提出了模糊加权层次聚类算法.该算法要计算可靠度和设置阈值来监督聚类,当可靠度低于阈值时要重新计算对象到底归属哪一类,这样就解决了上述问题.通过实验验证,这种方法有效可行,有效的解决了对象合并后再分裂或者分裂后再合并的问题.

但是该算法中阈值和权值的设定具有很大的人为性,并且该算法时间复杂度比其他方法要大.如何更合理的设置阈值和权值以及如何降低该算法的复杂度将是进一步的研究方向.

参考文献:

[1] 刘同明. 数据挖掘技术与应用[M]. 北京: 国防工业出版社, 2001.

[2] Kaufman L, Rousseeuw P J. An introduction to cluster analysis[C]// Finding Groups in Data. New York: John Wiley and Sons, 1990:458—462.

[3] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases [C]// SIGMOD'96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data . New York: ACM's flagship publication, 1996: 144—154.

[4] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases[J]. Information Systems, 2001,26(1):35—58.

[5] Guha S, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes [J]. Information Systems ,2000,25(5): 345—366 .

[6] Karypis G, Han E H, Kumar V. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling [J]. COMPUTER, 1999,32(8): 68—75

[7] 唐旭清, 朱平, 程家兴. 基于模糊商空间的聚类分析方法[J]. 软件学报, 2008,19(4):861—868.

[8] 李玲玲, 方帅, 辛浩. 改进的基于层次聚类的模糊聚类方法[J]. 合肥工业大学学报:自然科学版,2010,33(6): 859—862.

[9] 朱娴, 马卫. 一种基于层次聚类的双聚类算法[J]. 微计算机应用, 2009.30(5): 12—17.

[10] 李凯, 王兰. 层次聚类的簇集成方法研究[J]. 计算机工程与应用, 2010,46(27): 120—123.

[11] 蒋泽军. 模糊数学教程[M]. 北京: 国防工业出版社, 2004.

作者简介:

李剑英 男,(1985—),硕士研究生. 研究方向为层次聚类算法、粒度计算.

丁世飞 男,(1963—),博士,教授. 研究方向为数据挖掘、人工智能、粒度计算.