

一种基于类中心矫正的层次聚类算法

王 娴, 杨绪兵, 周 宇, 周溜溜

(南京林业大学 信息科学技术学院, 江苏 南京 210037)

摘 要: 为了克服传统层次聚类算法由于两类合并造成的中心点偏移的严重缺陷, 提出了一种基于类中心矫正的层次聚类算法, 从而提高了算法的精确度; 同时继承了传统层次聚类对初始中心点的无依赖性; 经分析, 算法对于已知聚类数和未知聚类数两种情况均有着良好的聚类效果. 通过标准数据测试, 结果表明新算法的聚类性能与层次聚类算法相比有更高的精确度; 并且让新算法用于指导图像分割实验, 证明了算法的有效性.

关键词: 层次聚类; 偏移; 类中心矫正; 精确度

中图分类号: TP39

文献标识码: A

文章编号: 1000-7180(2011)10-0217-04

An Algorithm of Hierarchical Clustering Based on Correcting Class Center

WANG Xian, YANG Xu-bing, ZHOU Yu, ZHOU Liu-liu

(College of Information Science and Technology, Nanjing Forestry University, Nanjing, 210037 China)

Abstract: In order to solve the problem of class center offset caused by merger of class of traditional hierarchical clustering algorithm, an algorithm of hierarchical clustering is proposed based on correcting class center. The accuracy of proposed algorithm is improved. And it inherits the non-dependence of the initial clustering center of the traditional hierarchical algorithm. After analysis, the algorithm obtains a good clustering result for the known and unknown number of clusters. It is proved that the proposed algorithm is not only more precise than traditional hierarchical clustering algorithm, but also the clustering results show that the algorithm is effective in image segmentation.

Key words: hierarchical clustering; offset ; class center correction; accuracy

1 引言

聚类分析方法是一种无监督的学习过程, 按照事物的某些属性, 把事物聚集成类, 使类间相似性尽量小, 类内相似性尽量大^[1]. 目前, 众多学者提出了很多聚类算法^[1-2], 其中层次聚类算法^[3-4]由于不需要固定类别数、对初始中心点的无依赖性以及存在全局最优解而受到广泛的应用. 但是, 层次聚类算法也存在一些问题, 比如层次聚类算法对于海量数据来说, 时空复杂性较高, 使聚类的效率降低、类中心偏移较大等. 为此, 许多学者也提出了许多改进的层次聚类算法. 比如, 文献[5]中提出了通过对类间

距离预排序, 避免合并类时重新计算类间距离的方法, 提高算法运行速度; 文献[6]中介绍了一种基于 90-10 规则的快速层次聚类算法, 缩短了时空复杂性. 针对类中心偏移这个问题, 本文设计出了一种改进的层次聚类算法, 这一算法不仅能够克服传统层次聚类算法由于两类合并造成的类中心偏移这一缺陷, 对聚类中心进行了矫正, 同时也保持了层次聚类算法的优点, 并将算法有效地应用在标准数据集和图像分割领域. 由于所提出的算法思想可以直接推广到其它的快速层次聚类算法中, 因而本文工作主要是针对传统的层次聚类算法.

收稿日期: 2011-04-27; 修回日期: 2011-06-25

基金项目: 江苏省自然科学基金项目(BK2009393); 南京林业大学基金项目(163070053, 163070657)

2 层次聚类算法

层次聚类算法可分为凝聚法(也称自底向上方法)和分裂法(也称为自顶向下方法)。绝大多数层次聚类算法采用的是凝聚法。首先将每个样本作为一个单独的聚类,然后合并相似性最大的聚类为一个大的聚类,逐步聚合成越来越大的类。可见,两个类簇之间相似性度量的计算方式是层次聚类算法的基础。本文主要是基于层次聚类算法的最短距离法^[7],其基本思想是将两个类的距离定义为两类中距离最近的样本之间的距离。

层次聚类算法的过程可描述如下:

- (1) 待聚类样本个数为 N , 视这些样本点独立成类, 得出两两类间距离 $d(i, j)$ 初始化距离矩阵;
- (2) 将距离矩阵中最小取值所对应的两类合并成一个新类;
- (3) 更新距离矩阵;
- (4) 循环步骤(2)、(3), 直到所有样本并为一类或者满足某个设定条件时算法停止。

层次聚类算法在合并过程中自始至终均采用两个点的均值作为聚类中心, 如此合并得到的最终聚类中心显然偏离了真正的聚类中心, 从而影响聚类效果。

3 改进算法及其分析过程

3.1 改进算法的基本思想

基于上述的层次聚类类中心偏移的缺陷, 本文提出了改进的层次聚类算法, 它能有效解决这个问题。即每一次合并之前, 我们将当前类簇中所有点的均值作为聚类中心。根据样本点是否属于某个已有聚类(假设现有聚类个数为 k , 即 $\{c_1, c_2, \dots, c_k\}$), 分为以下两种情况:

(1) 点与类簇合并

若聚类 C_j 中已含有个样本点, 当点 x 并入集合 C_j 后, 集合 C_j 中将包含 $l+1$ 个点, 此时, 集合 C_j 的中心点记为 C_{j+1} , 即:

$$C_{j+1} = \frac{1}{l+1} \left(\sum_{x_i \in C_j} x_i + x \right) \quad (1)$$

由式(1)可知, 要计算新聚类的中心点 C_{j+1} , 需要知道点 x , 以及未合并之前各个点的坐标; 而传统层次聚类方法仅仅是由 $C_{\text{new}} = (C_{\text{old}} + x)/2$, 得出新的聚类中心点, 所以会造成最终中心与真实聚类中心之间产生较大的偏移量。为了节省存储空间和时间复杂度, 我们将式(1)化简后得到式(2):

$$C_{j+1} = \frac{l}{l+1} C_j + \frac{1}{l+1} x \quad (2)$$

此时, 我们只需知道点 x 和未合并之前集合的中心点 C_j , 从而简化了迭代计算。

(2) 类与类簇合并

若类 C_i 中含有 m 个点, C_j 中含有 n 个点。当前距离最小的是代表两个类簇 C_i 和 C_j , 则可将 C_i 类和 C_j 进行合并, 此时, 产生的新的类中心点记为 C_{m+n} , 即式(3)所示:

$$C_{m+n} = \frac{1}{m+n} \left(\sum_{x_i \in C_i} x_i + \sum_{x_j \in C_j} x_j \right) \quad (3)$$

由式(3)化简后得到式(4):

$$C_{m+n} = \frac{m}{m+n} C_i + \frac{n}{m+n} C_j \quad (4)$$

可以证明, 按此方法计算的类簇中心可以真实表达数据分布的几何中心。为节省篇幅, 证明略去。此外, 从类中心构造上可以得知, 新算法保持了传统层次聚类的计算复杂度。

3.2 改进算法的实现

下面给出该算法的部分代码的具体描述:

1. origin_point: $1:n \times X = \{x_1, x_2, \dots, x_n\}$
2. $c = \text{DisMat2Row}(x)$;
3. three_unity{ };
4. myset = [];
5. class_index{classnum, 1} = [];
6. update_unity();
7. while(~ isempty(setdiff(origin_point, myset)))
8. if(isempty(intersect(myset, three_unity)))
- &&. (sum(three_unity > n) == 0) end
9. if sum(three_unity > n) == 2 end
10. if sum(three_unity > n) == 1 end
11. update_unity();
12. end

代码的第一行计算任意两个样本点之间的距离, 得到距离矩阵。第三至第五行进行初始化, three_unity 记录样本点对和其对应的距离; myset 记录已获得聚类的样本点, 用来判断算法何时停止; class_index{classnum, 1} 则表示点所在的聚类 classnum。第六行仿照层次聚类算法思想获得首次聚类, 更新距离矩阵。第七行至第十二行运用本文提出的改进算法进行合并, 其中包括三种情况, 即若样本点不属于某个聚类时, 则新开辟一聚类; 反之, 若当前最近的距离对均为代表点时, 则集合与集合合并; 若当前最近的距离对中有一个样本点为代表点,

则点与集合合并;最后每合并一次,则更新距离矩阵.

4 实验分析

实验包括两个部分,第一部分将改进后的算法与传统层次聚类算法分别作用于 Iris 数据集^[8]上,来比较其聚类效果.另一部分是利用改进后的算法对 Matlab 自带图像进行分割处理,证明算法的有效性.

4.1 数据集实验

Iris 数据集包含 150 个样本数据,分别取自三种不同类别的鸢尾花,共包含 3 类且均匀分布(每类分别有 50 个样本),每个样本数据含有 4 个属性,即萼片长度、萼片宽度、花瓣长度和花瓣宽度,单位是 cm.实验利用聚类平均精确度 (avg) 以及中心点偏移距离 (disoff) 作为评价标准.

(1) 平均精确度:设原数据集有 k 个类,用 c_i 表示第 i 类, n_i 表示 c_i 中样本的个数, m_i 为聚类正确的个数,则 m_i/n_i 表示第 i 类中的精确度,则其平均聚类精度的数学计算公式如式(5):

$$\text{avg} = \frac{1}{k} \sum_{i=1}^k m_i/n_i \tag{5}$$

(2) 中心点偏移距离:设原数据集有 k 个类, c_i 是算法得到的各类簇的中心点的坐标, o_i 为各类簇的实际中心点的坐标,其数学计算公式如式(6):

$$\text{disoff} = \sum_{i=1}^k (c_i - o_i)^2 \tag{6}$$

将层次聚类算法和本文的算法分别进行实验比较,得出的聚类结果如表 1 所示.

表 1 两种算法的实验对比结果

聚类算法	聚对样本数 (样本总数)	平均 精确度/%	偏移 距离(disoff)
层次聚类	102(150)	68.000	0.435 4
改进算法	124(150)	82.667	0.040 6

如表 1 所示,两种聚类算法中,在聚类精确度以及类中心偏移距离两方面综合考察,本文设计的算法均优于传统的层次聚类算法.正如表中数据显示,层次聚类算法最终的聚类中心与实际产生的聚类中心存在较大偏差,而本文算法则纠正了这一缺陷,其与实际的中心点的偏移距离较小.为了考察本文算法对偏移距离的纠正是否仅仅是由于聚类精确度所造成,因此利用部分 Iris 数据集(取 60%数据集)提高了传统层次聚类算法精度,并在此基础上再次得

出了类中心的偏移程度,结果如表 2 所示.

表 2 两种算法的实验对比结果(部分 Iris 数据集)

聚类算法	聚对样本数 (样本总数)	平均精确度%	偏移距离 (disoff)
层次聚类	76(90)	84.44 4	0.361 8
改进算法	124(150)	82.667	0.0406

从表 2 分析可以得出如下结论:传统层次聚类伴随着样本数的减少,准确度自然得到了提升,在已超出本文算法精确度的情况下其类中心的偏移距离仍然要远远大于本文所设计的算法;综合分析,本文所提出的新算法不仅提升了传统层次聚类的聚类精确度,并且使得聚类结果的类中心与实际类中心偏移极小,并通过实验证明了该性能的改进并非是由于精确率的提升所造成.

4.2 图像分割实验

为了验证本文提出的算法在图像分割应用中的有效性,通过实验对图像进行分割测试,如图 3 所示:图 1(a)为分割测试图像为 matlab 自带的校准图像“cameraman”.利用对图像像素特征统计的先验知识来确定类别数 c ,如果假设图像中的每个类别在像素特征上服从高斯分布,则在图像的统计直方图上的每个比较明显的峰值对应着每个类的均值,因此找到了这些明显峰值也就得到了聚类数目 c ,即 c 取 2.图 1(b)为本文提出的聚类算法分割结果.

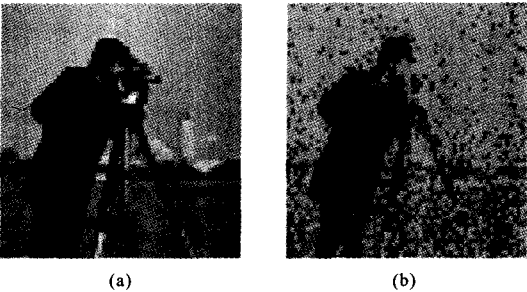


图 1 实验测试图及算法效果图

从内存开销方面来看,层次聚类与本文算法均需要全局计算出距离矩阵,图 2(a)表示的是 64×64 的测试图像,而当利用传统层次聚类直接进行处理时发生了内存溢出问题,因此将原始测试图像先进行预处理:将图像分为 4 块(每块大小为 32×32),再分别对其子块进行聚类,完成聚类后再拼接成 64×64 的图像,利用这种分而治之的策略分割效果如图 2(b)所示;从图中可以看出,联结处(子块的边缘)的分割效果并不好.

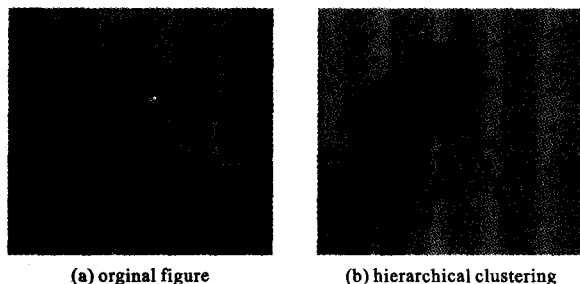


图2 原始图像与传统层次聚类分割效果图

5 结束语

文中对层次聚类算法中心点存在偏移这一不足之处进行了分析研究,并在矫正类中心过程中优化了距离计算方法,最后通过实验证明了传统的层次聚类算法与改进后的算法在标准数据集和图像分割上的有效性.比较发现,改进算法的聚类效果更加准确,符合客观实际.但是,两种算法计算时间与内存开销相当,对于较大数据集存在局限性,因而如何在有限内存开销范围内处理更大规模数据集有待进一步研究.

参考文献:

- [1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008,19(1):48-51.

- [2] 史变霞,张明新. 一种改进的层次聚类算法[J]. 微电子学与计算机, 2010,27(12):55-56.
- [3] 胡晓庆,马儒宁,钟宝江. 层次聚类算法的有效性研究[J]. 山东大学学报:工学版, 2010,40(5):146-147.
- [4] Kantardzic M. Data mining concepts, models, methods and algorithms[M]. 北京:清华大学出版社, 2003.
- [5] 段明秀,杨路明. 对层次聚类算法的改进[J]. 湖南理工学院学报:自然科学版, 2008,21(2):28-30.
- [6] 郭晓娟,刘晓霞,李晓玲. 层次聚类算法的改进与分析[J]. 计算机应用与软件, 2008,25(6):243-244.
- [7] Beyer K S, Goldstein J, Ramakrishnan R, et al. When is 'nearest neighbor' meaningful? [C] // Proceedings of the 7th International Conference On Data Theory. London, UK: Springer-Verlag, 1999:217-235.
- [8] Flsher A. Iris plants database[EB/OL]. [2011-04-16]. <http://archive.ics.uci.edu/ml/datasets/Iris>. Html. Authorized license.

作者简介:

王 娟 女, (1988-), 硕士研究生. 研究方向为模式识别、图像处理.

杨绪兵 男, (1973-), 博士, 副教授. 研究方向为模式识别、神经计算和图像处理.

周 宇 女, (1972-), 硕士, 副教授. 研究方向为图像处理、信息融合、模式识别.

周溜溜 男, (1987-), 硕士研究生. 研究方向为数据挖掘.

(上接第 216 页)

5 结束语

提出了一个新的抗噪特征,该特征的提取过程为:先计算语音信号的自相关函数,然后对自相关函数进行线性预测,最后对线性预测后的自相关函数用线性美尔尺度进行变换.用该特征在 GMM 模型中进行说话人识别实验.结果表明,在各种信噪比(SNR)环境中,该方法的精度高于线性预测倒谱系数(LPCC)和美尔倒谱系数(MFCC).

参考文献:

- [1] 赵力. 语音信号处理[M]. 北京:机械工业出版社, 2003.
- [2] 李建文,张晋平. 基于改进语音特征提取方法的语音识别[J]. 微电子学与计算机, 2009,26(007):230-233.
- [3] 朱周华,期望最大(EM)算法及其在混合高斯模型中的应用[J]. 现代电子技术, 2003(24):16-18

作者简介:

黄亚娟 女, 高级工程师. 研究方向为电子设备检测, 模式识别;

郑 林 男, (1984-), 硕士研究生. 研究方向为语音识别.