基于分布模型的层次聚类算法

叶 茂[1]，陈 勇[2]

(1. 电子科技大学计算机科学与工程学院 成都 610054; 2. 深圳大学经济学院 广东 深圳 518060)

【摘要】提出了一种新的层次聚类算法，先对数据集进行采样，以采样点为中心吸收邻域内的数据点形成子簇，再根据子簇是否相交实现层次聚类。在层次聚类过程中，重新定义了簇与簇之间的距离度量，并以此为基础建立堆结构。利用估计数据点总体分布的思想，证明该算法将逼近最优解。实验结果表明，算法的聚类效果大大优于现有的聚类算法。

关 键 词 聚类; 数据挖掘; 模式识别; 分布

中图分类号 TP301 文献标识码 A

Hierarchical Clustering Algorithm Based on Distribution Model

Ye Mao[1]，Cheng Yong[2]

(1. School of Computer Science and Engineering, UEST of China Chengdu 610054;

2. Department of Economics, University of Shenzhen Guangdong Shenzhen 518060)

Abstract A novel agglomerative method is proposed. This algorithm consists of three steps, first samples the dataset, then form the subcluster by absorbing the points in the å neighborhoods of sample points, at last final clusters are constructed by combining the subclusters. The distance measure of two clusters is redefined. Based on this concept, heap structure is constructed. Formally a theoretical explanation of the algorithm is given using the method approaching the actual distribution. Experimental results show the quality of ADA is much better than very many well-known algorithm CURE.

Key words clustering; data mining; pattern recognition; distribution

将数据点分组成多个类或簇称为聚类，在同一个簇中的点之间具有较高的相似度，而不同簇中的点差别较大。在欧氏空间中，一般采用距离来刻画相似度。聚类分析已经广泛地应用于许多领域[1~4]，包括模式识别、数据分析、图像处理和市场研究。通过聚类，能够识别密集和稀疏区域，同时发现全局的分布模式和数据属性之间的相互关系。传统的层次聚类算法根据簇间相似度的定义分为single-link，complete-link，group average，Ward’s method，BIRCH和CURE等。传统的层次聚类算法通常只能处理特殊的数据分布，如single-link处理长直的数据集，complete-link与group average处理大小一致的紧超球，Ward’s method处理等方差高斯分布，BIRCH算法处理球形并且大小一致的数据，CURE算法能部分处理任意形状的数据集，但对带状数据聚类结果不令人满意。本文提出一种基于分布模型的层次聚类算法，能处理一般的数据分布。

1 相关定义

传统的距离度量不能反映簇与簇之间的相似度，设C_i、C_j表示两个簇，m_i、m_j分别表示C_i和C_j簇中心，传统簇间距离的定义如下

收稿日期：2003 – 07 – 10

作者简介：叶 茂(1973 –)，男，博士，讲师，主要从事数据挖掘，模式识别，智能计算等方面的研究；陈 勇(1969 –)，男，博士，讲师，主要从事数据挖掘，金融数学等方面的研究.

$$d_{\text{mean}}(C_i,C_j)=d(m_i,m_j) \tag{1}$$

$$d_{ave}(C_i,C_j)=1/(|C_i||C_j|)\sum_{p\in C_i}\sum_{p'\in C_j}d(p,p') \tag{2}$$

$$d_{\max}(C_i,C_j)=\max_{p\in C_i,\,p'\in C_j}d(p,p') \tag{3}$$

$$d_{\min}(C_i,C_j)=\min_{p\in C_i,\,p'\in C_j}d(p,p') \tag{4}$$

式(2)～(4)分别对应group average，complete-link和single-link，由于这三种定义方式需要保留簇内所有的点，内存要求较高，且针对特殊的分布才有效，所以在海量数据处理中不采用，只采用式(1)的距离定义方式，如图1所示。从图1可以看出，簇C_i与簇C_j明显比簇C_j到簇C_m离得更远，用中心点作距离度量不能反映上述差别，所以必须定义一个新的簇间距离。

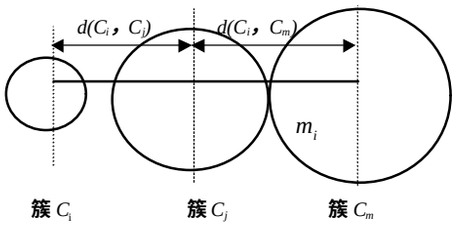

图 1 利用中心点定义簇距离的描述

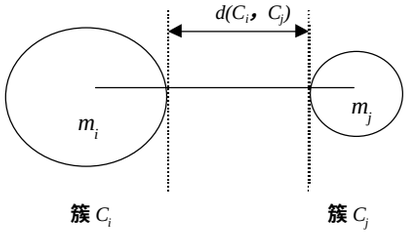

图 2 距离定义的描述

定义 1 设C_j是以m_j为中心的簇，C_i是以m_i为中心的簇，r为簇半径，簇间距离为

$$d(C_i,C_j)=d(m_i,m_j)-r(C_j)-r(C_i)$$

距离定义的描述如图2所示。

定义1能较准确确定簇与簇之间的相似度，另外，由于簇半径是通过平均的方法算得，所以该距离也有很强的抗噪声点的能力。为了处理大量数据以及进行增量聚类，算法只能保存簇的统计特征，称为聚类特征。

定义 2 簇$C=\{x_i,1\le i\le n\}$的聚类特征$C_F=(n，G，s)$，其中n为簇中数据点的数目，$G=\sum_i x_i$是所有数据点的线性和，$s=\sum_i x_i^2$是所有数据点的平方和。

由聚类特征可以容易地得出簇中心和半径，其中簇中心$m=G/n$，簇半径为

$$r(C)=\sqrt{\frac{\sum_{i=1}^{n}d^2(x_i,m)}{n}}$$

定理 1 对簇u，如果样本点数目v满足如下不等式[4]

$$v\ge fN+\frac{N}{|u|}\ln(1/d)+\frac{N}{|u|}\sqrt{[\ln(1/d)]^2+2f|u|\ln(1/d)}$$

则样本集含有属于簇u但数目少于$f|u|$的数据点的概率小于d，其中$0\le f\le 1$，N为数据总数。

2 ADA算法基本描述

ADA算法的详细流程如下：

算法1 Algorithm ADA(DataSet)

1) Initially select k sample points set SampleSet randomly;

2) BuildSubClusters(SampleSet，DataSet)，//建立k个子簇;

3) Build a heap，Q，with the distance of all combinations，//根据子簇间距离建立堆;

4) ClusterCombine(Q)，//递归合并子簇;

5) For each x_i;

6) LabelCluster(x_i，Q)，//标记数据点类别;

7) Until all points are passed by in the DataSet。

步骤1)随机选取k个数据点作样本集SampleSet，步骤2)利用样本集点建立k个子簇，并计算簇聚类特征。

算法2

$$G=G+x_i \qquad s=s+x_i^2 \qquad n=n+1$$

算法1中步骤3)、4)计算子簇间的距离，并建立优先堆数据结构，递归地合并k个子簇。如果簇与簇间的最小距离小于0，说明两个簇有交叉，合并这两个簇，直到没有小于0的簇间距离为止。

算法3　Algorithm ClusterCombine(Q)

1) Extract the nearest two cluster C_j，C_i　from the heap Q　which $d(C_j, C_i)<0$；

2) Merge the two subclusters which belong to into a new subcluster；

3) Add C_j，C_i　cluster feature to the new subcluster feature list，and build a new heap Q；

4) If there is C_{j*}，C_{i*}　in Q such that $d(C_{j*}, C_{i*})<0$；

5) ClusterCombine(Q)。

算法1中的步骤5)根据聚类特征链表标记数据集点的类别。

ADA算法在步骤1)的复杂度为$O(n)$，步骤2)的复杂度为$O(k\log n)$，步骤3)～5)的复杂度为$O(n)$，所以ADA算法复杂度最差情况为$O(kn)$。由于采用了堆等数据结构，所以算法速度较快。对于噪声点的处理如下：1) 在递归合并子簇之前，根据子簇数据点的数目判断，如果数据点很少，可以判断该子簇为噪声点组成的数据集；2) 对新生成的簇按上述办法判断，实验结果表明上述方法对处理噪声点非常有效。

3　理论描述

ADA算法计算目标函数的最大值为

$$L(q_1,q_2,\mathrm{L}\ ,q_k;l_1,l_2,\mathrm{L}\ ,l_n \mid X)=\prod_{i=1}^{n} p(x_i \mid q_{l_i}) \tag{5}$$

式中　l_i 是每个数据的类标号，$l_i=j$，如果x_i属于类j，则$q_1,q_2,\mathrm{L}\ ,q_k$ 为分布模型的参数，n为数据集点的数目。

ADA算法是基于分布模型的层次聚类算法，从分类P～P'，式(5)值由L变化到L'，相对差距为

$$DL(P,P')=L(P')/L(P)$$

定理 2　如果k、e 使得概率密度估计误差很小，ADA算法合并子簇有 $\Delta L \geq 1$。

证明　设 $B(z_{l_i}, r_i)$ 是以样本点 z_{l_i} 为球心，r_i 为半径的超球，V表示体积，概率密度估计为

$$p(x_i \mid z_{l_i}, r_i)=\begin{cases}\dfrac{k/n}{V(B(z_{l_i}, r_i))} & x_i \in B(z_{l_i}, r_i)\\ 0 & x_i \notin B(z_{l_i}, r_i)\end{cases}$$

此时目标函数为

$$L(l_1,l_2,\mathrm{L}\ ,l_n, r)=\prod_{i=1}^{n} p(x_i \mid z_{l_i}, r_i)$$

如图3所示，如果簇C_1与C_2合并为C_3，即划分从P变到P'，在超球$B(z_1, r_1)$内距离C_1中心点最近点的数目为k_1'，在超球$B(z_2, r_2)$内距离C_2中心点最近点的数目为k_2'，即

$$L(l_1,l_2,\mathrm{L}\ ,l_n, r)=\left(\frac{k_1/n}{v_1}\right)^{k_1'}\left(\frac{k_2/n}{v_2}\right)^{k_2'}\mathrm{L}$$

式中　v_1、v_2分别为$B(z_1, r_1)$和$B(z_2, r_2)$的体积。 在合并时，密度高的子簇吸收密度低的子簇，设C_1密度高，图3所示阴影部分同属簇C_1和C_2，但合并为簇C_3后，阴影部分点的概率密度大于或等于簇C_2的概率密度，因此 $\Delta L \geq 1$。

图4所示为两个数据集，第一个数据集与文献[4]使用的数据集相同，第二个数据集是层次聚类算法的经典实例。

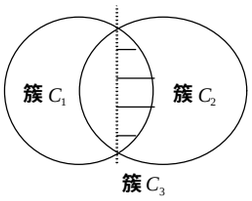

图 3 子簇合并描述

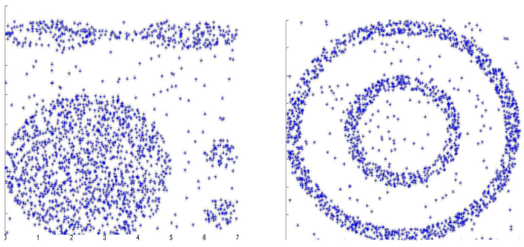

图 4 实验使用的数据集

ADA算法的实验结果如图5所示，其他算法的实验结果请参见文献[4]。图5中“ v ”与“ > ”点表示噪声点，其他不同形状的点表示不同的簇。

CURE算法的实验结果如图6所示，实验中设簇代表点数目为10。比较实验结果可以看出，ADA算法获得了更好的聚类结果。

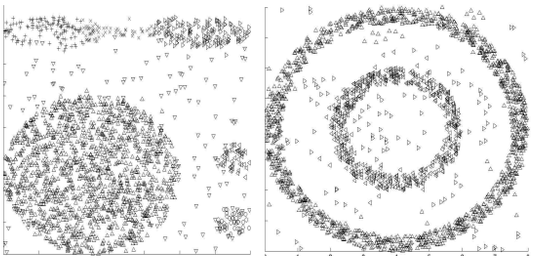

图5 ADA算法的聚类结果

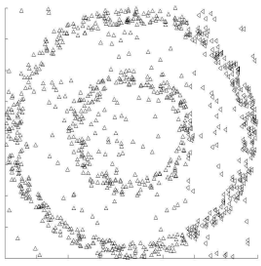

图6 CURE算法在数据集2的聚类结果

4 结 束 语

本文提出了一种聚类算法，该算法结合模型和层次聚类算法的优点，可以智能地确定簇的数目，并对噪声点有相当强的稳定性，其运算复杂度为$O(n)$且质量大大提高，实验验证了该算法的有效性和可行性。

本文研究工作得到了电子科技大学青年基金资助，在此表示感谢。

参 考 文 献

[1] George K, Han E H, Kumar V. Chameleon: a hierarchical clustering algorithm using dynamic modeling[J]. IEEE Computer, 1999, 32(8): 390-400

[2] Wang H X, Wang W, Yang J. Clustering by pattern similarity in large data sets[C]. SIGMOD Conference, 2002, 394-405

[3] Zhang T, Ramakrishman R, Livny M. BIRCH: an efficient data clustering method for very large databases[C]. Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data(SIGMOD’ 96), 1996, 103-114

[4] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases[C]. Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data(SIGMOD’ 98), 1998, 73-84

编 辑 徐培红