第 33 卷 第 8 期 计 算 机 工 程 2007 年 4 月
Vol.33 No.8 Computer Engineering April 2007
·软件技术与数据库· 文章编号：1000—3428(2007)08—0066—02 文献标识码：A 中图分类号：TP311

近似 k-median 分类属性数据聚类

赵 恒，张高煜
(西安电子科技大学电子工程学院，西安 710071)

摘 要：数据挖掘中解决分类属性数据聚类的算法有很多种，但大多数基于划分的方法得到的聚类中心一般不是数据集中的实际数据对象，缺乏实际的物理意义，有时会导致某一聚类为空。该文研究了近似 k-median 的求解算法，用数据的近似中值来代替模式进行聚类，提出了分类属性数据的近似 k-median 聚类算法，克服了一般基于划分的可分类属性数据聚类中所遇到的问题，仿真实验证明该算法有效。
关键词：数据挖掘；近似 k-median 聚类；分类属性数据

Approximate k-median Clustering for Categorical Data

ZHAO Heng, ZHANG Gaoyu
(School of Electronic Eng., Xidian University, Xi'an 710071)

【Abstract】 Based on the approximate k-median algorithm, an approximate k-median clustering algorithm for categorical data is developed. The algorithm replaces the modes in k-modes algorithm with the approximate medians of data set, and optimizes the center of cluster with the approximate k-median algorithm. The center of cluster is an actual sample of data set, which prevents the empty cluster. The experiments indicate the algorithm is effective.
【Key words】 Data mining; Approximate k-median clustering; Categorical data

数据挖掘中的数据类型多种多样，大多数实际的数据库和大的数据集不仅包括数值类型的数据而且包括大量的非数值类型数据，如二值(Binary)类型、符号(Symbolic)类型和分类(Categorical)类型数据等。分类属性数据的属性是有限和无序的，并且不可比较大小，如{“红”,“黄”,“蓝”,“绿”}，{“正方形”,“圆形”,“梯形”}等。k-modes 算法[1,2]是解决此类数据聚类的一种比较有效的方法，它用模式取代了k-means算法中的均值，聚类模式根据数据属性取值的频率进行更新，一个属性的高频率的取值作为模式的属性值，通过迭代，使得目标函数达到局部最小。然而，这样得到的聚类模式往往缺乏实际的物理意义，而且，算法产生的聚类中心一般不是数据集中的实际数据对象，这一点有时会导致某一聚类为空。本文采用各类中的实际样本点代表聚类，以各类的某个子集为搜索空间，寻找数据的近似中值(approximated median)来代替模式(modes)进行聚类，提出了分类属性数据的近似k-median聚类算法，实验表明，对于比较大的数据集，数据量的增长能够使搜索空间增大从而使得近似k-median与k-modes的聚类性能相差不大，而k-median选用实际的数据样本作为聚类的中心，能够避免某些情况下k-modes面临的空聚类问题。

1 近似中值及其求解

数据中值可以从一个限定的数据集 D 或是整个数据空间 Ω 得到。

定义 1 假设 X 是一个给定的数据集，它的广义类中值由式(1)得到：

$$q=\arg\min_{x\in\Omega}\sum_{x'\in X}d(x,x') \tag{1}$$

其中，$d(x,x')$ 是距离函数，寻找这样的中值是一个NP难问题[3]，一些算法被提出来得到它的近似解[4]。

定义 2 假设 X 是一个给定的数据集，它的限定数据类中值由下式得到：

$$q=\arg\min_{x\in X}\sum_{x'\in X}d(x,x') \tag{2}$$

即在数据集 X 中寻求到 X 中所有数据距离和最小的数据点，它的计算复杂度是 $O(|X|^2)$。

可以看到，无论是求解广义中值还是限定中值，都存在计算复杂度高的缺点。文献[5]提出了一个近似中值的求解算法，它对数据点的结构和距离函数没有特定的要求，并且具有线性的时间复杂度。实验表明，在合适的参数下，这种近似算法能够得到比较精确的类中值。

假设 X 是一个给定的数据集，近似中值由以下算法求得：

(1)从 X 中随机选择 n_r 个样本作为参考样本；

(2)计算 X 中每个点到 n_r 个参考点的距离和；

(3)选择距离和最小的 n_t 个样本点作为测试样本；

(4)计算每个测试样本到 X 中所有数据的距离和；

(5)距离和最小的测试样本就作为数据集的近似中值。

算法中 n_r 和 n_t 是两个输入参数，一般可以选择 $n_r=n_t$，而且实验表明，相对于整个数据集而言，很小的 n_r 和 n_t 就能够得到较精确的数据中值。

2 近似 k-median 聚类算法

定义 3 设 $X,Y\in D$，$d(X,Y)$ 是 X,Y 的差异度函数，

作者简介：赵 恒(1975－)，男，讲师，主研方向：数据挖掘，聚类分析，图像处理等；张高煜，博士生
收稿日期：2006-06-13 **E-mail：**yaheng2000@sina.com

定义：

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \tag{3}$$

其中：

$\delta(x_j,y_j)=0 \quad (x_j=y_j)$

$\delta(x_j,y_j)=1 \quad (x_j\neq y_j)$

设 $[C_1,C_2,...,C_k]$ 是数据集 $D=\{X_1,X_2,...,X_n\}$ 的一个划分，目标函数定义为

$$E_{am}=\sum_{l=1}^{k}\sum_{i=1}^{n}u_{il}d(X_i,Q_l)=\sum_{l=1}^{k}\sum_{C_l}d(X_i,Q_l) \tag{4}$$

这里，$Q_l=[q_{l1},q_{l2},...,q_{lm}]$ 是能够代表聚类 l 的数据中值，$u_{il}\in\{0,1\}$ 是划分矩阵 $U_{n\times k}$ 的一个元素，$u_{il}=1$ 表示数据对象 X_i 属于聚类 l，否则 X_i 不属于聚类 l。d 是差异度函数由式(3)定义。

算法的具体步骤如下：

(1)选择 k 个初始向量：$Q_1,Q_2,...,Q_k$，分别代表 k 个类；

(2)根据差异度函数的定义，将各数据对象分配到离它最近的初始中值所代表的类中；

(3)用上节的方法求取各个类的新的近似中值：$Q_1',Q_2',...,Q_k'$，并计算相应的目标函数 E_{am}；

(4)重复(2)(3)直到相邻两次的目标函数值的变化小于某一阈值 ξ 或迭代次数达到最大。

相对于 k-modes 聚类，近似 k-median 方法的聚类中心直观易于解释，聚类过程中不会出现空聚类的优点，算法复杂度是 $O(knn_r)$，大大低于 k-mediods 算法和 k-median 聚类算法。

3 实验结果分析

实验采用 Mushroom 数据，这是一组有关蘑菇性状分类的数据。包含 23 个蘑菇种类，总共 8 124 个数据对象，分为两大类，以它的毒性作为分类标记，“有毒”或“无毒”。其中有毒和无毒的蘑菇个数分别为 3 916 和 4 208 个。数据包括 22 个属性，全部可看作分类（categorical）属性，其中第 11 个属性（Stalk-root）有大量缺失，因此我们将该属性从数据集中去除。

一般评价聚类结果用到的“误分率”等统计方法是建立在聚类结果和输入样本的原始分类结构一一对应的基础上的，但聚类是无监督学习算法，其结果与输入样本原始分类结构并不一定有明显的对应关系，因此，我们用Folkes & Mallows（FM）指标[6~8]来评价其结果，它在 0 和 1 之间取值，其值越大表明划分 C' 和 C 越相似。

假设原始数据分类是 $C=\{C_1,C_2,...,C_k\}$，聚类算法所得到的结果是 $C'=\{C_1',C_2',...,C_k'\}$，$C_l$ 和 $C_l'(l=1,2,...,k)$ 分别表示类 l 中的数据集。建立结果表，以 C_l 和 $C_l'(l=1,2,...,k)$ 为表的纵向和横向单元，交叉元素表示 C_i 和 C_j' 中相同数据对象的个数。

(1)不同参数下的聚类结果

近似 k-median 需要指定参考样本和测试样本的参数，我们选择 $n_r=n_t$，并分别等于各类数据个数的 0.5%，1%，5%，10%进行实验。图 1 显示目标函数随迭代次数的变化曲线，参考样本和测试样本分别为总数据量的 0.5%，1%和 5%，目标函数都能够迅速收敛到一个局部极小。

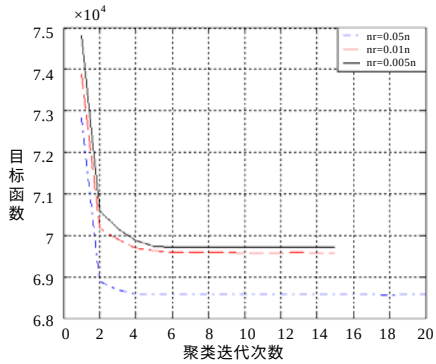

图 1 目标函数随迭代次数的变化曲线

表 1 为不同参数下近似 k-median 聚类的结果（n 为各类中的数据个数），显示出随着选用参考样本和测试样本的数量的不同，聚类效果也不同。由于 mushroom 的数据量较大，当参考样本和测试样本相对非常少时，占数据量的 0.5%，它们的样本选取的随机性影响就显现出来，因为参考样本的随机选取应该尽可能反映数据的分布特征，而数据量过少时，随机性变强，而统计性变差，近似 k-median 方法难以得到比较好的近似最优，因此也影响聚类的结果。随着参考样本数量的增加，聚类的误分率逐渐下降，精确度逐渐提高，但是，所用时间也随之增长。可以看到，聚类时间是随着所用的样本数近似于线性增长的，当 n_r 大于数据的 10%时，聚类性能已经不能明显提升，而运算时间则显著增长，因此合适的选择参考点和测试点的数量对聚类结果非常重要。

表 1 聚类的结果

	$n_r=0.005n$		$n_r=0.01n$		$n_r=0.05n$		$n_r=0.1n$	
	C_1'	C_2'	C_1'	C_2'	C_1'	C_2'	C_1'	C_2'
C_1（无毒）	3 474	734	3 597	611	3 811	397	3 785	423
C_2（有毒）	1 583	2 335	1 268	2 648	713	3 203	676	3240
误分率	28.52%		23.13%		13.66%		13.53%	
聚类精确度 FM	0.604 6		0.651 5		0.765 7		0.767 3	
运行时间（s）（迭代 10 次）	47		91		429		836	

(2)本算法与 k-modes 算法的比较

表 2 显示了本算法在参考样本取 $n_r=0.05n$ 时与 k-modes 聚类结果的比较。

表 2 本算法与 k-modes 算法的比较

	近似 k-median 算法 $n_r=0.05n$		k-modes 算法	
	C_1'	C_2'	C_1'	C_2'
C_1（无毒）	3 781	427	58	4 150
C_2（有毒）	603	3 313	3 088	828
误分率	12.68%		10.91%	
聚类精确度 FM	0.779 4		0.810 8	

可以看出，近似 k-mdian 算法和 k-modes 算法相比相差不大。k-modes 聚类是用基于频率的方法进行聚类过程中的模式的更新，模式各属性值取使得在各属性上数据对象频率最大的属性值，这个搜索空间要比近似 k-median 的搜索空间大，因而能够找到几乎最佳的属性组合，而近似 k-median 只在数据集上寻找最优解。对于比较大的数据集，数据量的增长能够使搜索空间增大从而使得近似 k-median 与 k-modes 的聚类性能相差不大，然而，k-median 选用实际的数据样本作

(下转第 70 页)