利用划分方法进行混合数据聚类

梁 红
(68029 部队，甘肃 兰州 730020)

摘 要：目前常用的几种基于划分的聚类方法主要处理数值型数据，能有效处理实际应用领域中常用的包括数值和符号混合数据的聚类算法则较少。基于此问题，文章根据 k 均值、k 中心点和 k 众数等基于划分的聚类方法各自的特点，对其进行集成与改进，提出一种能够应用于混合类型数据的聚类分析方法，即将所有的混合类型变量转换到共同的标度区间[0.0,1.0]中，根据合并处理的相异度计算公式计算对象之间的相异度；对于聚类中心中的各变量采取适合于相应类型的最佳方法进行独立更新，从而实现混合类型变更量数据的有效聚类。

关键词：划分方法；聚类分析；混合类型数据；相异度
中图分类号：P237.3 **文献标志码**：B **文章编号**：1672-4623（2011）06-0018-03

聚类分析又称为数据分割，需要把一个数据对象分组，即分为多个子集。使得每个组内部对象之间的相关性比与其他组对象之间的相关性更加紧密[1]。聚类算法可分为：层次聚类方法（hierarchical methods）、基于密度的方法（density-based methods）、基于划分的方法（partitioning methods）、基于网格的方法（grid-based methods）和基于模型的方法（model-based methods）[2]。

基于划分的聚类方法（k 均值、k 中心点等）主要处理数值型数据，在实际应用中，人们经常需要处理如性别、形状、疾病类型等符号型（分类型）数据，或者是既有数值型又有符号型的混合数据，若将基于划分的聚类方法直接应用于混合数据的聚类分析中，获得的聚类结果往往会表现出明显的不合理和错误划分，从而大大限制了很多面向数值型数据的聚类算法的应用范围。在现有的聚类算法中，能有效处理符号型或者混合型数据的聚类算法较少。其中，Huang 提出了 k 众数算法[3]和模糊 k 众数算法[4]，k 众数算法用模（mode）来替换聚类中心，采用符号匹配的差异性计算方法来处理符号量，并利用基于频率方法对各聚类中心进行更新。K-prototypes（KP）算法结合了 k 均值和 k 众数算法的特征，能够对采用数值型和符号型混合数据描述的对象进行聚类分析。虽然这些算法解决了符号型数据的聚类问题，且算法的计算和存储复杂度并未增加，但其缺点也是明显的。对于一个类内的每一符号属性，通常无法用单一模来表示类内所有对象在该属性上的统计信息，k 众数算法是以丢失其他符号值的统计信息为代价的[5]。

文章结合 k 均值、k 中心点和 k 众数聚类方法各自的特点，对其进行集成与改进，提出一种能够应用于混合类型数据的聚类分析方法。

1 基于划分的聚类方法介绍

给定 n 个对象的数据集 D，以及要生成的簇的数目 k，划分算法将对象组织为 k 个划分（k≤n），每个划分代表一个簇。这些簇的形成旨在优化一个目标划分准则，如基于距离的相异度函数，使得根据数据集的属性，在同一个簇中的对象是“相似的”，而不同簇中的对象是“相异的”[2]。

最著名和最常用的划分方法是 k 均值、k 中心点以及它们的变种。其中 k 均值方法只能用于簇均值有定义的情况下，在某些应用中，例如当涉及具有分类属性的数据时，均值可能无定义；k 中心点方法不采用簇中对象的均值作为参照点，而是在每个簇中选出一个实际的对象来代表该簇，但是在聚类过程中，需要不断计算各次更新的总代价，而原 k 中心点方法的总代价计算并未考虑混合类型变量的情况，且该方法每次迭代的复杂度是 $O(k(n-k)^2)$（n 是聚类对象的总个数），当 n 和 k 的值较大时，计算代价相当高[6,7]。

2 混合类型数据的划分聚类方法研究

聚类分析处理变量的类型可能是区间标度的、对称二元的、非对称二元的、分类的、序数的或者比例标度的。在许多真实的数据库中，对象一般是混合类型的变量描述的。如何计算用混合类型变量描述的对象之间的相异度以及如何更新混合类型变量的聚类中心，是混合类型数据利用划分方法进行聚类所必须解决的。

2.1 混合类型变量的对象之间相异度计算

有两种方法计算常用的混合类型对象之间的相异

收稿日期：2011-04-27
项目来源：国家 863 计划资助项目（2009AA12Z228）。

度。① 将变量按类型进行分组，对每种类型的变量分别进行相异度计算；② 将所有类型的变量一起处理，将其转换到共同的标度区间。采用第一种方法进行聚类分析时，如果各数据类型的聚类结果能够兼容，则该方法是可行的。然而每种类型的变量的聚类分析有可能产生不兼容的结果。第二种方法是将不同类型的变量组合在单个相异矩阵中，并把所有有意义的变量转换到共同标度的区间［0.0,1.0］中，以下是对该方法的具体介绍 [2]。

假设数据集包含 p 个混合类型的变量，对象 i 和对象 j 之间的相异度 d（i，j）定义为

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}} \tag{1}$$

其中，若对象 i 和对象 j 在变量 f 上的属性值 x_{if} 和 x_{jf} 中某一个或者两者都缺失，或者 $x_{if}=x_{jf}=0$，且变量 f 是非对称二元变量，则指示项 $\delta_{ij}^{(f)}=0$；否则，指示项 $\delta_{ij}^{(f)}=1$。变量 f 对 i 和 j 之间相异度的贡献 $d_{ij}^{(f)}$ 根据它的类型计算：

1）如果 f 是区间标度变量：

$$d_{ij}^{(f)}=\frac{|x_{if}-x_{jf}|}{\max_h x_{hf}-\min_h x_{hf}} \tag{2}$$

其中 h 遍取变量 f 的所有非缺失对象。

2）如果 f 是二元或分类变量 如果 $x_{if}=x_{jf}$，$d_{ij}^{(f)}=0$；否则 $d_{ij}^{(f)}=1$。

3）如果 f 是序数变量：计算秩 r_{if} 和 $z_{if}=\frac{r_{if}-1}{M_f-1}$，并将 z_{if} 作为区间标度变量对待。其中，M_f 是序数变量 f 的状态个数，这些有序的状态定义了一个秩评定 1,...，M_f 用对应的秩 $r_{if}\in\{1,\cdots,M_f\}$ 代替 x_{if}。

4）如果 f 是比例标度变量：要么进行对数变换，并且把变换后的数据当作区间标度数据；要么把 f 当作连续的序数数据，进行离散化，计算 r_{if} 和 z_{if}，然后把 z_{if} 当作区间标度的数据处理。

以上混合类型变量的对象之间相异度计算中的各种数据类型的处理方法与各种单一变量类型的处理方法基本相同。唯一的不同就是基于区间的变量，其中规格化使得变量值映射到区间［0.0,1.0］。这样，即使描述对象的变量具有不同类型，对象之间的相异度也能够进行计算。

2.2　混合类型变量数据的聚类中心更新

混合类型变量数据的聚类中心更新采用个类型变量中心进行分别更新的方法，即对于聚类中心中的各变量采取独立更新。

1）如果变量类型为区间标度型，根据 2.1 节中 1）的方法将其转换到区间［0.0,1.0］内。为了避免求均值过程中对于离群点的敏感性（一个具有很大的极端值的对象可能显著的扭曲数据的分布）以及求中心点过程十分耗时的弊病，采用改进的中值更新算法，即获得各簇对象中该变量的最大最小值，若簇原聚类中心相应该变量的值在［最小，最大］区间内，则更新的该簇聚类中心的相应变量值保持不变；若大于最大值，则用最大值替换该变量值；若小于最小值，则用最小值替换。

2）如果变量类型是二元或分类型，采用众数方法对其进行更新，即采用基于频率的方法对其进行更新；

3）如果变量类型是序数型或比例标度型，根据 2.1 节中 3）、4）介绍的转换方法将其转换为相应的区间标度型，利用改进中值的方法其进行更新，具体方法如该节 1）中所述。

2.3　本文的混合类型变量的聚类方法流程

本文提出的混合类型变量的聚类方法具体流程为：

Step1：将混合类型变量中的区间标度型、序数型以及比例标度型数据根据 3.1 节中介绍的相应类型数据转换的方法将其转换到共同的标度区间［0.0,1.0］中；

Step2：确定聚类个数 k，在转换后的混合类型数据中选取初始聚类中心 m_i，i=1，…，k；

Step3：根据公式（1）计算各对象到簇中心点的相异度，

if $d(i,p)=\min d(i,j)\quad i=1,\ldots,n,j=1,\cdots,k$

then 对象 i∈以 p 为聚类中心的簇 C_p。

Step4：根据 3.2 节介绍的簇中心更新方法，更新各簇中心值；

Step5：if 不再发生变化

then 停止计算，输出各簇及其聚类中心值；

else 转 Step3。

3　实验分析

实验将以 UCI 中的 3 个含符号属性的实际数据集(描述见表 1)为聚类对象来验证本文混合类型数据聚类方法的有效性。

表 1　混合类型数据集

数据集名称	对象数	属性数据		类别数
		数值型	符号型	
Zoo	101	1	15	7
Crx	490	6	9	2
Vowel	990	10	3	11

分别采用 k 均值、k 中心点、k 众数方法以及本文方法对以上 4 个数据集进行聚类分析，将所得到的分

类结果与已知类别进行比较，根据公式

$$Entr(C_k)=-\frac{1}{\log(N)}\sum_{t\in l}\frac{N_{tk}}{N_k}\log(\frac{N_{tk}}{N_k}) \quad (3)$$

整个熵 E（C）是所有簇集合所计算得到的 Entr（C_k）的均值，其中，N_k 是簇 C_k 的大小，N_{tk} 是该簇中标识为类 t 的数量[8,9]。计算结果如表 2 所示。

表 2 各聚类方法所得分类结果聚类熵

数据集名称	聚类熵 E(C)			
	k 均值	k 中心点	k 众数	本文方法
Zoo	0.517	0.503	0.161	0.083
Crx	0.421	0.354	0.224	0.126
Vowel	0.533	0.426	0.287	0.148
平均值	0.490	0.428	0.224	0.119

从利用传统的 k 均值、k 中心点、k 众数及本文提出的聚类方法对数据集 Zoo，Crx 和 Vowel 的聚类结果可以看出，本文提出的聚类方法在处理混合数据的能力上要优于传统聚类方法。

4 结 语

文章所提出的基于划分的混合类型数据聚类方法分别从混合类型变量聚类中心的更新以及相异度的计算两个问题着手，分析原有的 k 均值、k 中心点、k 众数聚类方法的特点和应用类型，并对其进行集成与改进，实现混合类型变量的有效聚类。实验表明本文提出的方法对于混合类型数据的聚类划分结果要优于原 k 均值、k 中心点和 k 众数聚类方法的划分结果，从而证明了本文方法的有效性。

参考文献

[1] Fraley C, Raftery A E. How Many Clusters? Which Clustering Method? - Answers Via Model Based Cluster Analysis[J]. The Computer Journal, 1998, 41 :578-588

[2] Han J W, Kamber M. 数据挖掘概念与技术[M]. 北京 :机械工业出版社, 2007

[3] Huang Z. Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining Knowledge Discovery, 1998,2(3) :283-304

[4] Huang Z. Ng M K. A Fuzzy K-modes Algorithm for Clustering Categorical Data[J]. IEEE Transaction on Fuzzy Systems, 1999, 7(4) :446-452

[5] 汪加才, 文巨峰, 陈奇, 等. 结构化模糊 K-prototypes 聚类算法[J]. 计算机科学, 2005, 32(5) :155-158

[6] 高峰. K-centers 聚类算法在教学评估中的应用[J]. 计算机工程与应用,2007, 43(12) :191-193

[7] 杨春成. 空间数据挖掘中聚类分析算法的研究[D]. 郑州 :解放军信息工程大学, 2004

[8] 陈四清, 杨春成. 基于聚类有效性函数的面状地理实体聚类[J]. 测绘科学技术学报, 2006, 23(1) :44-47

[9] 张惟皎, 刘春煌, 李芳玉. 聚类质量的评价方法[J]. 计算机工程. 2005, 31(20) :10-12

作者简介：梁红，高级工程师，主要研究方向为空间数据处理、GIS。

（上接第 17 页）以上的硬盘存储空间，PIV 1.0G MHZ 或更高处理器。

3 数据更新问题

在系统投入运行后，系统数据的实时更新是保证系统现势性的必要手段，包括动物防疫的基础信息库、动物防疫的业务信息库和基础地理信息数据库。动物防疫的基础信息和动物防疫的业务信息要结合当地村、乡、县的情况，尽可能短周期的进行汇总直报，以便提高数据的实时性和可利用性。基础地理信息数据的更新问题，在现有定位技术中，GPS 具有精度高、速度快、作用距离长、能全天候作业等特点，能对屠宰场分布、饲养场分布等进行采集，经数据处理将采集数据转换为数据库需要的格式导入到系统中即可完成实时更新。动物防疫的基础信息和动物防疫的业务信息可以在信息录入界面中进行更新。

4 结 语

基于 GIS 的动物防疫指挥系统是信息化管理的一部分，是数字化在畜牧业方面的具体实现和体现方式，是实现全面信息化的需要。随着地理信息科技的进步和畜牧业现代化管理的逐步实施，对动物免疫指挥系统的要求越来越迫切。系统的建立对动物防疫信息化管理和重大动物疫病控制指挥调度必将起到积极的作用。

参考文献

[1] 吴信才. 地理信息系统原理、方法及应用 [M]. 北京 :电子工业出版社, 2009

[2] 王家耀. 空间信息系统原理 [M]. 北京 :科技出版社, 2001

[3] 薛伟. MapObjects-地理信息系统程序设计 [M]. 北京 :国防工业出版社, 2004

[4] 袁博, 邵进达. 地理信息系统基础与实践 [M]. 北京 :国防工业出版社, 2006

[5] 毋河海. 关于 GIS 缓冲区的建立问题[J].武汉测绘科技大学学报, 1997, 22(4) :358-364

[6] GB/T17694-2009. 地理信息术语[S].

[7] 龚键雅.地理信息系统技术 [M]. 北京 :科学出版社, 2001

第一作者简介：池淑文，高级工程师，主要从事地理信息系统建设。