混合属性数据 k-prototypes 聚类算法①

余文利[1], 余建军[1], 方建文[2]

[1](衢州职业技术学院 信息工程学院, 衢州 324000)
[2](衢州学院 电气与信息工程学院, 衢州 324000)

摘 要: 在现实世界中经常遇到混合数值属性和分类属性的数据, k-prototypes 是聚类该类型数据的主要算法之一. 针对现有混合属性聚类算法的不足, 提出一种基于分布式质心和新差异测度的改进的 k-prototypes 算法. 在新算法中, 首先引入分布式质心来表示簇中的分类属性的簇中心, 然后结合均值和分布式质心来表示混合属性的簇中心, 并提出一种新的差异测度来计算数据对象与簇中心的距离, 新差异测度考虑了不同属性在聚类过程中的重要性. 在三个真实数据集上的仿真实验表明, 与传统的聚类算法相比, 本文算法的聚类精度要优于传统的聚类算法, 从而验证了本文算法的有效性.

关键词: 聚类; 分布式质心; 混合型数据; 新差异测度; 属性重要性

K-Prototypes Algorithm for Clustering of Data Mixed with Numeric and Categorical Attributes

YU Wen-Li[1], YU Jian-Jun[1], FANG Jian-Wen[2]

[1](College of Information Engineering, Quzhou College of Technology, Quzhou 32400, China)
[2](College of Electrical and Information Engineering, Quzhou University, Quzhou 32400, China)

Abstract: Data objects with mixed numeric and categorical attributes are commonly encountered in real world. The k-prototypes algorithm is one of the principals for clustering this type of data objects. An improved k-prototypes algorithm is proposed to cluster mixed data in this paper. In our method, the concept of the distribution centroid is introduced for representing the prototype of categorical attributes in a cluster. Then we combine both mean with distribution centroid to represent the prototype of the cluster with mixed attributes, and thus propose a new measure to calculate the dissimilarity between data objects and prototypes of clusters. This measure takes into account the significance of different attributes towards the clustering process. Finally, we present out algorithm for clustering mixed data, and the performance of our method is demonstrated by a series of experiments on three real-world datasets in comparison with that of traditional clustering algorithm.

Key words: clustering; distribution centroid; mixed data; new dissimilarity measure; attribute significance

聚类是数据挖掘中的一个重要分支, 聚类算法广泛应用于图像处理、客户细分、基因表达分析和文本分析等领域[1-4]. 聚类的目的是将一组数据对象划分到多个簇中, 使得一个对象与属于同一簇的其他对象彼此相似, 而与其他簇中的对象相异[5].

在现实世界中, 大多数数据集都同时具有数值属性和分类属性, 然而传统的聚类算法如k-means算法[6]、k-modes[7]算法、模糊 k-modes 算法[8]、TGCA 算法[9]和 COOLCAT 算法[10]等, 主要针对数值型数据或分类型数据. 当遇到混合属性数据时, 这些算法通常利用转换方法将某一类属性转换成另一类属性, 然后在其上应用传统的单类型聚类算法. 但是在大多数情况下, 转换方法通常会导致数据丢失, 得到非期望的聚类结果. 为克服传统聚类算法的局限性, Li 等人提出了基于

① 收稿时间:2014-09-24;收到修改稿时间:2014-11-14

相似度量方法的层次凝聚聚类算法 SBAC[11](Similarity-Based Agglomerative Clustering), SABC 算法采用文献[12]的相似度量方法来评估数据对象之间的相似度; Hsu 等人给出了基于方差和熵的混合数据聚类算法 CAVE[13], CAVE 算法需要为每一个分类属性构建一个距离层次结构, 而距离层次结构的确定需要领域的专业知识; 文献[14]使用改进的自组织映射来分析混合属性数据, 在该方法中, 距离层次结构通过分类属性值自动构建; 基于簇中提取的数据是高斯分布的假设, Chatzis 提出了 KL-FCM-GM 算法[15], 算法是专门为高斯多项式分布的数据设计的; Huang 提出了 k-prototypes 算法[16], 它结合 k-means 与 k-modes 算法对混合属性数据进行划分; 模糊 k-prototypes 算法[17]在 k-prototypes 算法的基础上考虑了数据对象的模糊特性; Zheng 等人将进化算法框架引入到 k-prototypes 中, 提出了 EKP 算法[18](Evolutionary K-Prototype), 从而具有全局搜索能力.

针对现有混合属性聚类算法的不足, 本文提出一种基于分布式质心和新差异测度的改进的 k-prototypes 算法. 首先用分布式质心来表示簇中分类属性的簇中心, 然后结合分布式质心和均值来表示簇中混合属性的簇中心, 还提出了一种新的差异测度来评估数据对象与簇中心的差异, 新差异测度充分考虑了数据对象的每一属性在聚类过程中的影响力.

1 k-prototypes算法

k-prototypes 算法是由 Huang 等[8]提出的能够对混合数值属性和分类属性数据进行聚类的一种有效算法. 该算法将 k-means 与 k-modes 算法结合起来, 通过参数 μ_l 来控制数值属性与分类属性在聚类过程中的权重[19]. 令 $X=\{X_1,X_2,\ldots,X_n\}$表示 n 个数据对象集, 其中 $X_i(1\le i\le n)$是一个具有属性 $A_1,A_2,\ldots,A_m$ 的数据对象. 用 $Dom(A_j)$来表示属性 A_j的值域, 与混合数据相关联的属性的值域分别是数值型和分类型, 其中数值型值域用连续值表示, 分类型值域用有限的、无任何自然顺序的集合表示, 如性别、颜色等, 通常用 $Dom(A_j)=\{a_j^1,a_j^2,\ldots,a_j^t\}$表示, 其中 t 是属性 A_j的数目. 数据对象 X_i在逻辑上通常表示为属性值对的并集: $[A_1=x_{i1}]\wedge[A_2=x_{i2}]\wedge\ldots\wedge[A_j=x_{ij}]\wedge\ldots\wedge[A_m=x_{im}]$, 其中 $x_{ij}\in Dom(A_j)$, $1\le j\le m$. X_i 也通常表示为向量形式: $[x_{i1},x_{i2},\ldots,x_{im}]$.

设 k 是一个正整数, k-prototypes 算法的目的是将数据集 X 划分为 k 个簇. 通过最小化以下代价函数作为聚类准则

$$E(U,Q)=\sum_{l=1}^{k}\sum_{i=1}^{n}u_{il}d(x_i,Q_l) \tag{1}$$

式中, Q_l为簇 l 的中心, $u_{il}(0\le u_{il}\le 1)$是划分矩阵 $U_{n\times k}$中的元素, $d(x_i,Q_l)$是差异测度, 定义如下

$$d(x_i,Q_l)=\sum_{j=1}^{m}d(x_{ij},q_{lj}) \tag{2}$$

$$d(x_{ij},q_{lj})=\begin{cases}(x_{ij}-q_{lj})^2 & \text{如果第}l\text{个属性是数值属性}\\ \mu_l\delta(x_{ij},q_{lj}) & \text{如果第}l\text{个属性是分类属性}\end{cases} \tag{3}$$

式中, 当 $p=q$, 则 $\delta(p,q)=0$, 否则 $\delta(p,q)=1$; μ_l是簇 l 中分类属性的权值; 当 x_{ij} 是数值属性值时, q_{ij} 是簇 l 中第 j 个数值属性的均值; 当 x_{ij} 是分类属性值时, q_{ij} 是簇 l 中第 j 个分类属性的模式. k-prototypes 算法的流程如下所示

步骤 1. 从数据集 X 中随机选择 k 个数据对象作为初始的簇中心;

步骤 2. 对于数据集 X 中的每个数据对象, 根据式(2)计算其到各簇中心的距离, 将其划分到距离最近的簇中. 每一次划分结束后, 更新相应簇的簇中心.

步骤 3. 当数据集中的所有对象都分配到相应的簇后, 重新计算这些数据对象到当前簇中心的距离. 如果发现离某一数据对象最近的簇中心在其他簇中, 则将该数据对象重新分配到离其最近的簇中心所在的簇中, 然后更新发生数据对象变动的两个簇的簇中心.

步骤 4. 重复步骤 3, 直到经过新一轮计算之后没有改变簇的数据对象存在.

2 改进混合属性数据k-prototypes聚类算法

2.1 分布式质心

文献[20]提出了适合于表示模糊场景下分类属性簇簇中心的模糊质心概念, 并且验证了在表示分类属性数据簇的簇中心上的有效性. 受该思想的启发, 本文定义了分布式质心来表示非模糊场景下分类属性簇的簇中心. 假设分类属性 j 的值域 $Dom(A_j)=\{a_j^1,a_j^2,\ldots,a_j^t\}$, 其中 t 是属性 j 在数据集 X 中取不同属性值的数目. 则簇 l 的分布式质心 C_l' 定义如下:

$$C_l'=\{c_{l1}',c_{l2}',\ldots c_{lj}',\ldots,c_{lm}'\} \tag{4}$$

式中, $c'_{lj}=\{\{a_j^1,\omega_{lj}^1\},\{a_j^2,\omega_{lj}^2\},\ldots\{a_j^k,\omega_{lj}^k\},\ldots,\{a_j^t,\omega_{lj}^t\}\}$.

上式中

$$\omega_{lj}^k=\sum_{i=1}^{n}\eta(x_{ij}) \tag{5}$$

这里 $\eta(x_{ij})=\begin{cases}u_{il}/\sum_{i=1}^{n}u_{il}, & \text{如果} x_{ij}=a_j^k\\ 0, & \text{如果} x_{ij}\neq a_j^k\end{cases}$; 如果数据对象 x_i 出现在簇 l 中, u_{il}=1, 否则, u_{il}=0.

在式(4) 中 ω_{ij}^k 满足如下条件: $\begin{cases}0\le\omega_{lj}^k\le 1,\ 1\le k\le t\\ \sum_{k=1}^{t}\omega_{lj}^k=2,\ 1\le j\le m\end{cases}$.

簇中每个分类属性的簇中心表示 $c'_{lj}\in C'_l$ 是值对 $\left\{a_j^k,\omega_{lj}^k\right\}$ 的集合, $1\le k\le t$, 这个值对的值是由分类属性 j 在簇 l 中的值的分布情况决定的. 从式(4)-式(5)可以看出, 分布式质心记录了簇中分类属性的每个值出现的频率. 因此, 使用分布式质心来表示簇中心可以充分体现簇中分类属性的特征.

2.2 属性重要性分配原则

属性重要性是指在聚类过程中属性的影响力. 在文献[21]中, Huang 等人提出一种计算属性重要性的方法, 在该方法中, 属性重要性值是通过最小化目标函数值得到. 给定一个数据集划分, 属性重要性的分配原则为: 对簇内距离(Within Cluster Distance, WCD)之和小的属性, 分配一个大的属性重要性值; 否则, 分配一个小的属性重要性值. 该原则由下式给出

$$s_j\propto\frac{1}{D_j} \tag{6}$$

式中, s_j 为属性 j 的重要性, $\propto$ 表示正比例符号, D_j 为属性 j 的簇内距离和.

2.3 改进的 k-prototypes 聚类算法描述

结合 3.1 节和 3.2 节的思想处理混合属性数据的聚类问题, 参照文献[21]中的 W-k-means 算法框架, 本文算法的目标函数为

$$E(U,Q,S)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}u_{il}s_j^{\lambda}d(x_{ij},q_{lj}) \tag{7}$$

式中, u_{il} 是划分矩阵 $U_{n\times k}$ 的元素, 满足 $u_{il}\in\{0,1\}$ 和 $\sum_{l=1}^{k}u_{il}=1$; s_j 是第 j 个属性的重要性值, 满足 $0\le s_j\le 1$ 和 $\sum_{j=1}^{m}s_j=1$; $\lambda>1$ 是重要性值 s_j 的指数; q_{lj} 是簇 l 中第 j 个属性的簇中心.

最小化式(7)所示的目标函数, 需要迭代解决以下三个子问题:

1) 子问题 1: 令 $Q=Q'$, $S=S'$为固定值, 最小化目标函数 $E(U,Q',S')$;

2) 子问题2: 令$U=U'$, $S=S'$为固定值, 最小化目标函数$E(U',Q,S')$;

3) 子问题 3: 令 $U=U'$, $Q=Q'$为固定值, 最小化目标函数 $E(U',Q',S)$.

其中子问题 1 可以通过下式解决

$$\begin{cases}u_{il}=1 \text{ 如果}\sum_{j=1}^{m}s_j^{\lambda}d(x_{ij},q_{lj})\le\sum_{j=1}^{m}s_j^{\lambda}d(x_{ij},q_{ej}),\\ \qquad\qquad for\ 1\le e\le k\\ u_{ie}=0\ \ for\ e\neq l\end{cases} \tag{8}$$

对数值属性, 子问题 2 的解决公式如下

$$q_{lj}=\frac{\sum_{i=1}^{n}u_{il}\square x_{ij}}{\sum_{i=1}^{n}u_{il}} \tag{9}$$

而对于分类属性, $q_{lj}=c'_{lj}$, 其中 c'_{lj} 由式(4)给出.

在式(7)中, 新的差异测度 $d(x_{ij}, q_{lj})$定义如下

$$d(x_{ij},q_{lj})=\begin{cases}\left|x_{ij}-q_{lj}\right| & \text{如果第}l\text{个属性是数值属性}\\ \varphi(x_{ij},q_{lj}) & \text{如果第}l\text{个属性是分类属性}\end{cases} \tag{10}$$

式中, $\varphi(x_{ij},q_{lj})=\sum_{k=1}^{t}\vartheta(x_{ij},a_j^k)$, $\vartheta(x_{ij},a_j^k)=\begin{cases}0 & \text{如果} x_{ij}=a_j^k\\ \omega_{lj}^k & \text{如果} x_{ij}\neq a_j^k\end{cases}$, 其中 ω_{lj}^k 可以通过式(5)计算得到. 子问题 3 的由下述定理解决:

定理[20]. 令 $U=U'$,$Q=Q'$为固定值, 则 $E(U',Q',S)$ 取得最小值当且仅当

$$s'_j=\begin{cases}0 & \text{如果} D_j=0\\ 1\Big/\sum_{\tau=1}^{h}\left[D_j/D_\tau\right]^{1/(\lambda-1)} & \text{如果} D_j\neq 0\end{cases} \tag{11}$$

式中, $D_j=\sum_{l=1}^{k}\sum_{i=1}^{n}u'_{il}d(x_{ij},q'_{lj})$, h 为 $D_j\neq 0$ 的属性个数.

以上给出了所有相关变量的计算方法, 基于分布式质心和新差异测度的改进的 k-prototypes 算法的具体步骤如下:

输入: 最大迭代次数 *maxIte*, 簇数 k 和系数 λ;

输出: 聚类结果.

步骤 1. 从数据集中随机选择 k 个无缺失值的不同数据对象, 转化为初始的簇中心, 用 $Q^{(t)}=(Q_1,Q_2,\ldots,Q_k)$ 表示; 随机生成初始的属性重要性值, 用 $S^t=\{s_1^t,s_2^t,\ldots,s_m^t\}(\sum_{j=1}^{m}s_j^t=1)$ 表示, 并令 t=0.

步骤 2. 令 $Q'=Q^t$ 和 $S'=S^t$, 通过最小化目标函数 $E(U,Q',S')$得到 U^{t+1}.

步骤 3. 令 $U'=U^{t+1}$ 和 $S'=S^t$, 通过最小化目标函数 $E(U',Q,S')$得到 Q^{t+1}.

步骤 4. 令 $U'=U^{t+1}$ 和 $Q'=Q^{t+1}$, 通过最小化目标函数 $E(U',Q',S)$得到 S^{t+1}.

步骤 5. 如果目标函数 E 的值没有改变或最大迭代次数 $maxIte$ 等于 0, 则算法终止; 否则, 令 $t=t+1$, $maxIte=maxIte-1$, 返回步骤 2 继续执行.

算法中步骤 1 中的转换规则描述如下: 如果第 j 个属性是数值属性, 则该属性的簇中心 $q_{lj}\in Q_l$ 的值作为该属性的值; 如果第 j 个属性是分类属性, 则簇中心 $c'_{lj}\in Q_l$ 的值按以下方式初始化: 当 $x_{ij}=a_j^k$ 时, $\omega_{lj}^k=1$; 否则 $\omega_{lj}^k=0$.

2.4 算法复杂度分析

本文算法的时间复杂度主要包括每次迭代中簇中心的更新和划分矩阵的计算. 两者的计算代价分别为 $O(kmn)$和 $O(k(p+Nm-Np)n)$, 其中 k 是簇数, p 是数值属性数, m 是所有属性数, $N=max(t)$为所有分类属性中, 单个属性在数据集中拥有不同属性值的最大数目, n 是数据集中数据对象数. 因此总的时间复杂度为 $O(k(m+p+Nm-Np)nl)$, 其中 l 是本文算法收敛所需的迭代次数. k-prototypes 算法的时间复杂度为 $O((l+1)kn)$, 因此本文算法的时间复杂度稍大于传统 k-prototypes 算法. 而当 $n>>k,m,l$ 时, 两种算法均要快于时间复杂度为 $O(n^2)$的层次聚类算法. 至于空间复杂度, 本文算法存储簇中心和数据集 X 的空间复杂度为 $O(k(p+mN-pN)+mn)$, 存储划分矩阵的空间复杂度为 $O(kn)$, 因此本文算法的空间复杂度为 $O(k(p+n+mN-pN)+mn)$.

3 仿真实验与结果分析

为了评估本文算法的性能, 使用来自 UCI 机器学习库的三个真实数据集(描述如表 1 所示)作为聚类对象, 分别用本文算法、k-prototypes 算法[8]、SBAC 算法[11]和 KL-FCM-GM 算法[15]进行聚类分析. 在仿真实验中, 基于文献[21]的加权原则, 对于所有的数据集, 式(7)中的参数 λ 取值为 8; 传统 k-prototypes 算法从数据集中随机选择 k 个无缺失值的数据对象作为初始簇中心; 本文算法使用 3.3 节的转换规则将这 k 个数据对象转换成初始的簇中心.

表1 数据集描述

数据集名称	对象数	属性数		类别数
		数值	分类	
Iris	150	3	0	3
Syobean	47	0	35	4
Heart Disease	303	6	9	5或2

在聚类分析中, 使用文献[7]的聚类准确率(clustering accuracy)作为评估指标, 定义如下

$$r=\frac{\sum_{i=1}^{k}a_i}{n} \tag{12}$$

式中, a_i 是同时出现在第 i 个簇和第 i 个真实类中数据对象的数目, n 是数据集中数据对象的个数. 按照该评估指标, r 取值越大, 算法的聚类效果越好, 在聚类结果完全正确的情况下 r=1.0.

考虑到初始簇中心和属性重要性值的随机性, 本文算法与其他三种算法各运行 100 次, 然后取聚类准确率的平均值, 实验结果如表 2 所示. 为了更直观的显示各个算法的聚类结果, 各种算法的聚类结果的直方图如图 1 所示.

表2 聚类实验结果

数据集名称	算法	聚类准确率(r)
Iris	k-prototypes算法	0.819
	SBAC算法	0.373
	KL-FCM-GM算法	0.335(α=1.1)
	本文算法	0.822
Soybean	k-prototypes算法	0.856
	SBAC算法	0.617
	KL-FCM-GM算法	0.876(α=1.8)
	本文算法	0.908
Heart Disease	k-prototypes算法	0.546
	SBAC算法	0.425
	KL-FCM-GM算法	0.653(α=1.3)
	本文算法	0.842

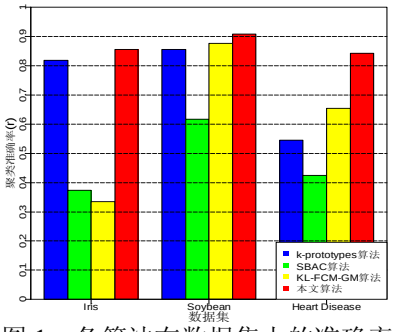

图 1 各算法在数据集上的准确率

从表 2 和图 1 可以看出, 与其他传统算法相比, 在

三个真实数据集上, 本文算法的聚类准确率都是最高的, 因此本文算法的性能更好. 另外以上结果还表明, 本文算法不但适合于处理混合属性数据, 还适合于处理数值属性数据和分类属性数据. 本文算法性能更好源于: 1) 簇中心同时包含了数值属性和分类属性在簇中的分布信息; 2) 差异测度考虑了属性在聚类过程中的重要性; 3) 通过使用文献[21]的加权策略, 本文算法能够自动计算不同属性在聚类过程中的作用.

4 结语

针对混合属性数据的聚类问题, 本文提出了一种新的基于分布式质心的改进的 k-prototypes 聚类算法. 对于混合属性数据簇的簇中心表示问题, 受模糊质心概念的启发, 提出分布式质心的概念来表示簇中分类属性的簇中心, 然后结合分布式质心和均值来表示混合属性簇的簇中心. 使用新的差异测度来计算数据对象与簇中心的距离, 该差异测度考虑了不同属性在聚类过程中的作用. 与其他传统算法相比, 本文算法具有三个优势: 新的簇中心的使用, 使得算法能够准确表示混合属性数据簇的特征; 算法考虑了不同属性在聚类过程中的作用; 加权策略的使用, 使得算法能够自动计算不同属性在聚类过程中的作用.

参考文献

1 郑伟,潘正勇.结合 FCM 和 RSF 模型的医学图像分割方法.计算机应用与软件,2014,31(2):198–200,204.

2 曾小青,徐泰,张丹,等.基于消费数据挖掘的多指标客户细分新方法.计算机应用研究,2013,30(10):2944–2947.

3 江雨燕,李平,王清.基于共享背景主题的 Labeled LDA 模型.电子学报,2013,41(9):1794–1799.

4 许涛,尚学群,杨密静,等.基于离散时序基因表达数据的双聚类算法.计算机应用研究,2013,30(12):3551–3556,3567.

5 海沫,张书云,马燕林.分布式环境中聚类问题算法研究综述.计算机应用研究,2013,30(9): 2561–2564.

6 Marques JP. Pattern Recognition: Concepts, Methods and Application. Springer, 2001.

7 Huang ZX. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2(3): 283–304.

8 Huang ZX, Michael KN. A fuzzy k-modes algorithm for clustering categorical data. IEEE Trans. on Fuzzy System, 1999, 7(4): 446–452.

9 He H, Tan Y. A two-stage genetic algorithm for automatic clustering. Neurocomputing, 2012, 81(1): 49–59.

10 Barbara D, Couto J, Li Y. COOLCATL: an entropy-based algorithm for categorical clustering. Proc. of the Eleventh International Conference on Information and Knowledge Management. 2002. 582–589.

11 Li C, Biswas G. Unsupervised learning with mixed numeric and nominal data. IEEE Trans. on Knowledge and Data Engineering, 2002, 14(4): 673–690.

12 Goodall DW. A new similarity index based on probability. Biometrics, 1966, 22(4): 882–907.

13 Hsu CC, Chen YC. Mining of mixed data with application to catalog marketing. Expert System and Application, 2007, 32(1): 12–23.

14 Hsu CC, Lin SH, Tai WS. Apply extended self-organizing map to cluster and classify mixed-type data. Neurocomputing, 2011, 74(18): 3832–3842.

15 Chatzis SP. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. Expert System with Application, 2011, 38(7): 8684–8689.

16 Huang ZX. Clustering large data sets with mixed numeric and categorical vlues. Proceedings of the 1st Pacific-Asia Knowledge Discovery and Data Mining Conference. Singapore, World Scientific. 1997. 21–34.

17 Bezdek JC, Keller J, Krisnapuram R, et al. Fuzzy Models and Algorithm for Pattern Recognition and Image Processing. Boston, Kluwer Academy Publishers, 1999.

18 Zheng Z, Gong MG, Ma JJ, et al. Unsupervised evolutionary clustering algorithm for mixed type data. Proc. of the IEEE Congress on Evolutionary Computation(CEC). 2010. 1–8.

19 刘吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1): 48–61.

20 Kim W, Lee KH, Lee D. Fuzzy clustering of categorical data using fuzzy centroids. Pattern Recognition Letters, 2004, 25(11): 1263–1271.

21 Huang ZX, Ng MK, et al. Automated variable weighting in k-means type clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657–668.