软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2628−2641 [doi: 10.3724/SP.J.1001.2013.04470]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

属性加权的类属型数据非模聚类*

陈黎飞, 郭躬德

(福建师范大学 数学与计算机科学学院,福建 福州 350108)

通讯作者: 陈黎飞, E-mail: clfei@fjnu.edu.cn, http://math.fjnu.edu.cn/newmath/Article/ShowArticle.asp?ArticleID=741

摘 要: 类属型数据广泛分布于生物信息学等许多应用领域,其离散取值的特点使得类属数据聚类成为统计机器学习领域一项困难的任务.当前的主流方法依赖于类属属性的模进行聚类优化和相关属性的权重计算.提出一种非模的类属型数据统计聚类方法.首先,基于新定义的相异度度量,推导了属性加权的类属数据聚类目标函数.该函数以对象与簇之间的平均距离为基础,从而避免了现有方法以模为中心导致的问题.其次,定义了一种类属型数据的软子空间聚类算法.该算法在聚类过程中根据属性取值的总体分布,而不仅限于属性的模,赋予每个属性衡量其与簇类相关程度的权重,实现自动的特征选择.在合成数据和实际应用数据集上的实验结果表明,与现有的基于模的聚类算法和基于蒙特卡罗优化的其他非模算法相比,该算法有效地提高了聚类结果的质量.

关键词: 聚类;类属型数据;模;属性加权

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 陈黎飞,郭躬德.属性加权的类属型数据非模聚类.软件学报,2013,24(11):2628−2641. http://www.jos.org.cn/1000-9825/4470.htm

英文引用格式: Chen LF, Guo GD. Non-Mode clustering of categorical data with attributes weighting. Ruan Jian Xue Bao/ Journal of Software, 2013,24(11):2628−2641 (in Chinese). http://www.jos.org.cn/1000-9825/4470.htm

Non-Mode Clustering of Categorical Data with Attributes Weighting

CHEN Li-Fei, GUO Gong-De

(School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350108, China)

Corresponding author: CHEN Li-Fei, E-mail: clfei@fjnu.edu.cn, http://math.fjnu.edu.cn/newmath/Article/ShowArticle.asp?ArticleID=741

Abstract: While categorical data are widely used in many applications such as Bioinformatics, clustering categorical data is a difficult task in the filed of statistical machine learning due to the characteristic of the data which can only take discrete values. Typically, the mainstream methods are dependent on the mode of the categorical attributes in order to optimize the clusters and weight the relevant attributes. A non-mode approach is proposed for statistically clustering of categorical data in this paper. First, based on a newly defined dissimilarity measure, an objective function with attributes weighting is derived for categorical data clustering. The objective function is defined based on the average distance between the objects and the clusters, therefore overcomes the problems in the existing methods based on the mode category. Then, a soft-subspace clustering algorithm is proposed for clustering categorical data. In this algorithm, each attribute is assigned with weights measuring its degree of relevance to the clusters in terms of the overall distribution of categories instead of the mode category, enabling automatic feature selection during the clustering process. Experimental results carried out on some synthetic datasets and real-world datasets demonstrate that the proposed method significantly improves clustering quality.

Key words: clustering; categorical data; mode; attribute weighting

类属型数据(categorical data)普遍存在于许多实际应用领域,例如生物信息学领域的DNA序列数据,序列中的每个氨基酸以4种不同的符号A,T,G和C编码.迄今,研究者已提出多种聚类算法[1,2],其中的大多数研究集中

* 基金项目: 国家自然科学基金(61175123)
收稿时间: 2013-03-30; 修改时间: 2013-07-17; 定稿时间: 2013-08-27

于数值型数据(或称连续型数据)聚类.与数值型数据相比,类属型数据的属性值取自一个有限的符号集合,是离散的.这个特点使得类属型数据聚类成为统计机器学习领域富有挑战性的任务之一.其部分原因在于类属数据间相似度度量的定义困难,更重要地,是因为适用于类属型数据的统计方法和工具与已被广泛研究的数值型数据有实质性的区别.

与数值型数据不同,样本均值(mean)在类属型数据中没有意义.这意味着无法使用均值这个常用的统计工具为类属数据中的簇类定义几何可解释的簇中心,也使得包括 k-means[1]及其为数众多的变种[2−6]在内的许多成熟的聚类算法无法直接应用于类属型数据聚类.当前的主流方法[2,7]将类属属性的模视作类属型簇的中心,已开发出经典的 k-mode 算法[8]及其众多的改进算法[7,9−11].依据统计学的观点,这些方法仅考虑属性的模而忽略了其他符号的统计信息,是不完备的.近年来,业已提出若干非模(non-mode)方法,其中,k-representatives 算法[12]提出以所有符号的频度向量为簇中心,实质上,该方法并不直接处理类属型数据,而是在变换后的另一个大规模的二值型数据上进行聚类.一些层次聚类算法,如 ROCK[13]和 DHCC[14]等,可以进行非模的聚类,然而这些算法通常具有较高的时间复杂度.

自动属性加权是当前类属型数据聚类研究面临的另一个难题.自动属性加权技术已在连续型数据聚类领域得到广泛的研究和应用,可以在聚类过程中识别属性对簇类形成不同的贡献程度,从而达到自动特征选择和提高聚类有效性的目的[4,5].然而,这些技术同样无法直接应用于类属型数据聚类,因为在数值型数据聚类中常用的样本方差(variance)等概念在类属型数据中也没有意义[15].现有的方法,包括 WKM[9],MWKM[10]和 DHCC[14]等,严重依赖于类属属性的模,根据模的统计信息对属性进行加权,如前所述,这必然导致权重计算上的偏差.

本文提出一种非模的类属型数据统计聚类方法,该方法依据样本与整个簇类的相似性定义聚类模型,此相似性建立在样本与组成簇类的所有样本两两之间新提出的距离度量基础上,避免了传统方法对类属属性模概念的依赖.为了优化新的聚类模型,提出一种 EM 型的软子空间聚类算法,在线性时间复杂度内搜索最优的数据集聚类划分,并进行自动的类属型属性加权.新的属性加权方案根据类属值的总体分布,同样不是以模为基础,赋予每个属性表示其与簇类之间相关程度的权重.实际上,我们推导出的聚类优化目标函数是以基尼系数[16](Gini index)为基础的,在统计理论中,基尼系数可视作刻画类属数据分布方差的一个指标[15],从这个意义上说,上述新模型和属性加权方式与数值型数据聚类领域的有关结果相一致.

本文第 1 节主要介绍背景知识和相关的研究工作.第 2 节给出新的聚类模型.第 3 节描述具体的聚类算法并进行算法分析.第 4 节介绍实验环境和实验结果分析.最后在第 5 节对本文进行总结并指出进一步的研究方向.

1 背景知识与相关工作

本节讨论类属型数据聚类相关背景知识并介绍和分析若干相关工作.首先约定全文使用的记号.设待聚类的数据集为 $DB=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_i,\ldots,\boldsymbol{x}_N\}$,其中,$\boldsymbol{x}_i=\langle x_{i1},x_{i2},\ldots,x_{ij},\ldots,x_{iD}\rangle$是第 i 个(i=1,2,…,N)由 D 个类属属性值构成的数据对象,其中,N(N>1)表示待聚类对象的数目.符号 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 用于表示任意对象.记第 d 个(d=1,2,…,D)属性取值的集合为 O_d,并用 $o\in O_d$ 表示其中的任一符号(离散值),属性 d 离散取值的总数用$|O_d|$表示.对 DB 进行硬聚类就是将 DB 划分为 K 个子集的集合 $C=\{c_1,c_2,\ldots,c_K\}$,这里$\forall k\neq l,1\leqslant k,l\leqslant K,c_k\cap c_l=\varnothing,c_k$ 称为 DB 的第 k 个簇(k=1,2,…,K),K(K>1)是给定的簇数目.c_k 包含的数据对象数目记为$|c_k|$.表 1 给出类属型数据的一个例子.

Table 1 An example: A categorical cluster consisting of 5 data objects

表 1 一个例子:由 5 个数据对象组成的类属型簇

簇	数据对象	Attribute 1	Attribute 2	Attribute 3
c_1	$\boldsymbol{x}_1$	A	T	T
	$\boldsymbol{x}_2$	A	T	A
	$\boldsymbol{x}_3$	T	T	C
	$\boldsymbol{x}_4$	T	T	G
	$\boldsymbol{x}_5$	G	A	G

对于表 1 所列数据的 3 个类属属性,易知 O_1={A,T,G},O_2={A,T}和 O_3={A,T,G,C}.通常使用频度估计器(frequency estimator)估计类属属性的属性值分布,形式地,c_k 中属性值 $o\in O_d$ 的频度估计器定义为

$$f_k(o)=\frac{\#_k(o)}{|c_k|} \tag{1}$$

其中,$\#_k(o)$表示 c_k 中第 d 个属性取值为 o 的对象数目.簇 c_k 中属性 d 的模(mode)记为 m_{kd},特指出现频度最高的符号,也就是说,

$$m_{kd}=\arg\max_{o\in O_d} f_k(o) \tag{2}$$

以表 1 数据为例,3 个属性的模分别为 m_{11}=A(或 T),m_{12}=T 和 m_{13}=G.

依据是否为聚类过程定义显式的目标优化函数,现有的类属型数据聚类方法可以划分为两组.以层次聚类为代表的第 1 组方法,其目的是构造层次聚类树,因而没有必要显式地定义聚类目标函数,代表性算法包括凝聚型算法 ROCK[13]和分裂型算法 DHCC[14]等.此类算法具有较高的算法复杂度,通常达到 $O(N^2\log N)$.第 2 组方法首先基于分割熵[17]或频度统计信息[18]定义聚类目标函数,然后定义用于优化目标函数的聚类搜索算法.代表性算法包括基于熵的类属型数据聚类算法(entropy-based categorical clustering,简称 EBC)[17].该型算法通常使用蒙特卡罗抽样法[19]实现聚类优化,具体实现为:随机地选择一个簇中的对象,将其移动到另一个簇,使得移动之后的聚类划分具有更高的质量(优化目标函数值),重复这个过程直到目标函数值不再改变.因此,此类算法通常需要大量的迭代步骤来完成聚类.

在第 2 组方法中,另一种代表性算法借鉴了经典的 k-means 聚类[1](用于数值型数据聚类)思想,通过定义类 k-means 型聚类目标函数,使用 EM 型算法结构[20]进行类属型数据聚类.这种算法的优势在于几何可解释性(以簇中心代表簇)和较高的聚类效率.然而如前所述,对于类属型数据,目前尚缺乏一种有效的簇中心定义手段:以表 1 的 Attribute 2 为例,该属性的取值只能是 A 或 T,无法像处理数值型数据一样通过求平均值来估计它的数学期望值.以 k-modes[8]为代表的一类算法(还包括文献[7−11]等)将属性的模视作中心,以此为基础进行 k-means 型聚类.从统计角度看,这种方法必然是不完备的,因为此间其他非模符号的分布被忽略不计.以表 1 的数据为例,根据这种方法,Attribute 1 的中心可以是 A 也可以是 T,算法只能取其一而忽略了另一高频度的符号.近年来已提出了定义类属型簇中心的替代方案:k-representatives 算法[12]和 k-populations 算法[21],提出以一个由所有符号的频度构成的向量为簇中心(表 1 中的 Attribute 1 表示簇中心的向量为⟨0.4,0.4,0.2⟩).实质上,该方法是将类属型数据扩展为一个更大规模的 binary 型数据加以处理,并隐含假设不同属性上类属符号间的差异是相同的.

为了识别不同属性对簇类有差异的贡献程度,从而提高聚类结果的质量,数值型数据聚类领域已应用自动属性加权技术[3,6]:赋予每个属性一个类依赖(class-dependent)的权重并在聚类过程中给予优化,达到自动特征选择的目的.在类属型数据聚类中,现有方法可大致分为两种类型:基于距离的和基于模频度的属性加权方式.前者以 WKM 算法[9]为代表,根据簇内对象到模(视作簇“中心”)的平均距离赋予属性相应的权重;后者计算的权重以模的频度为依据,包括 DHCC 算法[14]等.新近出版的 MWKM[10]为每个属性计算两个权值,一个与模频度成反比,另一个与平均距离成反比,可以看作是 WKM 和 DHCC 加权方法的综合.还以表 1 数据为例,由于 Attribute 1 和 Attribute 3 的模具有相同的频度 0.4,在 MWKM 算法[10]中这两个属性将被赋予相同的权重.然而,这个结果与事实不符:Attribute 3 上属性值的分布与 Attribute 1 明显不同,二者与簇的相关程度理应存在区别.从上述分析可以看出,现有的属性加权方法均依赖于“模”,并没有考虑属性值的总体分布情况,这必然导致属性权重计算上的偏差,从而影响聚类结果的质量.

本文提出一种非模的类属型数据聚类模型,以弥补上述以模为中心的方法在聚类算法和属性加权方面存在的缺陷.我们定义了一个新的聚类目标优化函数,并称优化该目标函数的算法为 NMCC(non-mode clustering of categorical-data).新算法可划归上述的第 2 组类属型数据聚类方法.经过数学推导,给出了一个与上述两类现有方法完全不同的新属性加权方法,新方法依据属性类属值的总体分布情况(而不仅限于模)赋予属性权重.第 3.3 节将讨论不同加权方法之间的差别之处,下面首先提出一种非模的类属数据聚类模型.

2 属性加权的聚类模型

聚类与对象间的相似度度量密切相关.与数值型数据相比,定义类属属性之间的相似度较为困难.一种常用的度量是简单匹配系数(simple matching coefficient,简称 SMC).对于两个对象 x_i 和 x_j,其第 d 维属性的简单匹配系数为 $SMC(x_{id},x_{jd})=0$ 若 $x_{id}\neq x_{jd}$ 和 $SMC(x_{id},x_{jd})=1$ 若 $x_{id}=x_{jd}$.我们注意到,上述定义中的 1 可以用其他大于 0 的常数代替,这是因为对于同一个类属属性符号间的关系,只能区分出相同或不相同两种情形(因而用于衡量相同程度的数值只要大于 0 即可);而对于不同属性这种相同的程度是有区别的(因而可能有不同的数值).为此,为每个簇类 c_k 的第 d 维属性引入一个记号 w_{kd} 衡量该属性符号间的相同程度,并定义该属性上两个对象间的相异度(距离)为

$$dist_k(x_{id},x_{jd})=1-\begin{cases}w_{kd}^{-\beta}, & x_{id}=x_{jd}\\ 0, & x_{id}\neq x_{jd}\end{cases}=\begin{cases}1-w_{kd}^{-\beta}, & x_{id}=x_{jd}\\ 1, & x_{id}\neq x_{jd}\end{cases} \tag{3}$$

这里,$\beta(>0)$是一个预定义参数,$w_{kd}(k=1,2,\ldots,K;d=1,2,\ldots,D)$满足约束条件:

$$\begin{cases}\forall k,d: w_{kd}>0\\ \forall k: \sum_{d=1}^{D}\dfrac{1}{w_{kd}}=1\end{cases} \tag{4}$$

由公式(4)易知 $0<\dfrac{1}{w_{kd}}\leqslant 1$.当 $w_{kd}=1$ 时,$dist_k(x_{id},x_{jd})$与传统的简单匹配系数相对应,即 $dist_k(x_{id},x_{jd})=1-SMC(x_{id},x_{jd})$;当 $\dfrac{1}{w_{kd}}\to 0$ 时,$dist_k(x_{id},x_{jd})\approx 1$,意味着属性 d 不同符号间的差异被“平滑掉”.因此,w_{kd} 的引入实际上起到了类属属性平滑估计(smooth estimation)的作用.平滑估计是一种典型的类属数据概率估计方法[22].

由于类属数据簇类尚缺乏一个直观定义簇中心的方法,k-modes[8]提出将类属属性的模为簇中心,在此基础上,定义簇为分散度(scatter)最小(或紧凑度最大)的对象集合,其中的分散度以对象到簇中心的距离来衡量.显然,该型方法忽略了簇中非模符号的统计信息.这里,我们基于样本间的平均距离衡量簇的分散度,分散度越低则簇类的质量越高**.形式地,c_k 在第 d 维属性上簇的分散度定义为

$$\begin{aligned}Scat(k,d)&=\frac{1}{|c_k|(|c_k|-1)}\sum_{x_i\in c_k}\sum_{y_j\in c_k,y_j\neq x_i}dist(x_{id},y_{jd})\\ &=\frac{1}{|c_k|(|c_k|-1)}\sum_{x_i\in c_k}[(\#_k(x_{id})-1)(1-w_{kd}^{-\beta})+(|c_k|-\#_k(x_{id}))\times 1]\\ &=1-\frac{|c_k|}{|c_k|-1}w_{kd}^{-\beta}\left(\sum_{o\in Od}[f_k(o)]^2-\frac{1}{|c_k|}\right)\end{aligned} \tag{5}$$

上式第 1 步~第 2 步的计算依据如下:$\forall \boldsymbol{x}_i\in c_k$,$c_k$ 中属性 d 取值与 x_{id} 相同的其他对象数目为$\#_k(x_{id})-1$,根据公式(3),x_{id} 与其中每个对象在属性 d 上的距离为 $1-w_{kd}^{-\beta}$;c_k 中余下的$|c_k|-\#_k(x_{id})$个对象该属性取值均不同于 x_{id},根据公式(3),x_{id} 与这些对象的距离总和为$(|c_k|-\#_k(x_{id}))\times 1$,二者相加得到公式(5)的第 2 行.忽略 $Comp(k,d)$中的常数,并对 $k=1,2,\ldots,K$ 和 $d=1,2,\ldots,D$ 累加,可推导出新的聚类模型:给定 DB,所要搜索的目标簇类 $c_1,c_2,\ldots,c_K$ 是以下优化问题的解,其中,$W=\{w_{kd}|k=1,2,\ldots,K;d=1,2,\ldots,D\}$:

$$\min J_0(C,W)=\sum_{k=1}^{K}\frac{|c_k|}{|c_k|-1}\sum_{d=1}^{D}w_{kd}^{-\beta}\left(\frac{1}{|c_k|}-\sum_{o\in O_d}[f_k(o)]^2\right)\ \text{s.t. Eq.(4)} \tag{6}$$

模型中的 w_{kd} 可以视作表示第 d 维类属属性与簇 c_k 相关性的特征权重.下面通过推导任一对象 $\boldsymbol{x}_i$ 与簇 c_k 之间的距离计算公式来分析这个结论:根据公式(3)和公式(5),$\boldsymbol{x}_i$ 与 c_k 所有对象间的平均距离为

** 根据定义,聚类质量应包括簇内紧凑度(在数值上可以用簇的分散度来衡量,分散度越低则簇越紧凑)和簇间分离度两个方面的度量指标.若基于簇内样本两两间的距离定义分散度,那么簇间分离度就可以用非同簇样本两两间的距离来衡量.给定一个数据集,这两种距离的总和是常数,从这个意义上说,最小化簇的分散度实际上意味着最大化簇间分离度.因此,这里只考虑簇的分散度.

$$\begin{aligned} Dist(\boldsymbol{x}_i,c_k) &= \frac{1}{|c_k|}\sum_{d=1}^{D}\sum_{y_j\in c_k} dist(x_{id},y_{jd}) \\ &= \sum_{d=1}^{D}[(1-w_{kd}^{-\beta})f_k(x_{id})+(1-f_k(x_{id}))\times 1] \\ &= D-\sum_{d=1}^{D} w_{kd}^{-\beta} f_k(x_{id}) \end{aligned} \tag{7}$$

从公式(7)可知,给定β,w_{kd}的大小反映了属性 d 对距离度量的贡献程度.w_{kd}的数值较大时,测试对象 $\boldsymbol{x}_i$ 属性 d 上取值的差异将被放大,这意味着该属性与 c_k 具有较强的相关性.根据这个观察,下面称 w_{kd} 为属性 d 相对于簇 c_k 的特征权重,并记 $\boldsymbol{w}_k=\{w_{k1},\ldots,w_{kd},\ldots,w_{kD}\}$ 为 c_k 的权重向量.从效果上看,w_{kd} 起到了特征选择的作用:将 c_k 中的对象投影到 $\boldsymbol{w}_k$ 定义的一个软子空间(soft subspace)[4]中.由此,公式(6)定义的聚类模型可以视作一种软子空间聚类模型.与现有的模型(如 W-k-Means[3]和 MPC[5])不同,新聚类模型为类属型数据(而不是这些算法针对的数值型数据)定义了投影子空间,从而将特征选择作为类属型数据聚类模型的一部分.

下面对公式(6)定义的优化目标做进一步分析.用 $1-\frac{|c_k|-1}{|c_k|}$ 替换公式(6)中的 $\frac{1}{|c_k|}$ 项,整理后,目标函数可改写成:

$$J_0(C,W)=\sum_{k=1}^{K}\frac{|c_k|}{|c_k|-1}\sum_{d=1}^{D} w_{kd}^{-\beta}\left(1-\sum_{o\in O_d}[f_k(o)]^2\right)-\sum_{k=1}^{K}\sum_{d=1}^{D} w_{kd}^{-\beta} \tag{8}$$

注意到,上式第 1 项中的因子 $1-\sum_{o\in O_d}[f_k(o)]^2$ 正是常用的基尼系数[15,16].根据文献[15]的分析,基尼系数可用于表示类属型数据分布的方差,因此,最小化 J_0 意味着最小化类内对象分布的加权方差,这与数值型数据聚类的目标是一致的(例如自动特征加权的类 k-means 型算法[3−6]).另一方面,在新模型中,最小化 J_0 还意味着最大化第 2 项 $\sum_{k=1}^{K}\sum_{d=1}^{D} w_{kd}^{-\beta}$.由下面的性质 1 可知,当$\beta$>1 时,该项的数值越小,意味着特征权重越接近于均匀分布.

性质 1. 令 $J_1(\boldsymbol{w}_k)=\sum_{d=1}^{D} w_{kd}^{-\beta}$,若$\beta$>1,$J_1(\boldsymbol{w}_k)\in[D^{1-\beta},1]$,则 $J_1(\boldsymbol{w}_k)$取得最小值当且仅当 $w_{k1}=w_{k2}=\ldots=w_{kD}=D$.

证明:求带约束条件 $\frac{1}{w_{k1}}+\frac{1}{w_{k2}}+\ldots+\frac{1}{w_{kD}}=1$ 的 $J_1(\boldsymbol{w}_k)$极值问题,得到性质 1 的结论. □

依据性质 1,以下约定β>1.在此设置下,公式(6)定义的聚类模型体现了类属型数据软子空间聚类的目标:通过特征选择为每个簇选取一个最优投影子空间,使得投影子空间中簇内对象的分布具有最高的紧凑性.前者要求(对一个簇类而言)不同属性的权重值差异较大,数值上对应于公式(8)的第 2 项取得较大的值;后者希望最小化簇内对象分布的加权方差,这要求公式(8)第 1 项取得较小的值.最佳聚类质量对应于令两项取值之差最小的一种数据集划分.下一节讨论一种实现这个目标的优化算法.

3 新的聚类算法

本节提出一种称为 NMCC 的算法用于优化公式(6)定义的目标函数.这是一个带约束的非线性优化问题,应用拉格朗日乘子法,优化目标函数可转换为

$$J(C,W)=\sum_{k=1}^{K}\frac{|c_k|}{|c_k|-1}\sum_{d=1}^{D} w_{kd}^{-\beta}\left(\frac{1}{|c_k|}-\sum_{o\in O_d}[f_k(o)]^2\right)+\sum_{k=1}^{K}\lambda_k\left(\sum_{d=1}^{D}\frac{1}{w_{kd}}-1\right) \tag{9}$$

其中,λ_k(k=1,2,…,K)是对应于约束条件公式(4)的拉格朗日乘子.下面开始讨论一种达成这个优化目标的 EM 型算法过程.

3.1 EM型优化过程

EM 算法[20]是求解如公式(9)问题局部最优解的一种常用方法:从一个初始状态 $\hat{C}$ 出发,采用迭代算法结构,每个迭代步骤局部优化部分模型参数.根据这个原理,每次迭代过程首先设定 $C=\hat{C}$ 以求解最小化 $J(\hat{C},W)$ 的 W,记为 $\hat{W}$;其次,在第 2 个迭代步骤中,设定 $W=\hat{W}$ 通过最小化 $J(C,\hat{W})$ 求解最优的 C,即 $\hat{C}$.后者可以通过将每个对象 $\boldsymbol{x}$ 划分到距离最近的簇来实现.形式地,算法根据如下规则将 $\boldsymbol{x}$ 划分到簇 c_k 中去:

$$k=\text{argmin}_{i=1,2,\ldots,K} Dist(\boldsymbol{x},c_l) \tag{10}$$

从而生成新的聚类划分 $\hat{C}$. 第 1 个问题的求解根据以下定理 1 进行:

定理 1. 设 $C=\hat{C}$, 目标函数值 $J(\hat{C},W)$ 最小化当且仅当(对于 k=1,2,…,K 和 d=1,2,…,D):

$$\hat{w}_{kd}=\left[\sum_{o\in O_d}[f_k(o)]^2-\frac{1}{|\hat{c}_k|}\right]^{\frac{1}{\beta-1}}\Bigg/\sum_{l=1}^{D}\left[\sum_{o\in O_l}[f_k(o)]^2-\frac{1}{|\hat{c}_k|}\right]^{\frac{1}{\beta-1}} \tag{11}$$

证明:设定 $C=\hat{C}$ 时,根据公式(9),可以定义 K 个独立的(分别对应于 k=1,2,…,K 的)子优化目标函数:

$$J_k(\boldsymbol{w}_k,\lambda_k)=\frac{|c_k|}{|c_k|-1}\sum_{d=1}^{D}w_{kd}^{-\beta}\left(\frac{1}{|c_k|}-\sum_{o\in O_d}[f_k(o)]^2\right)+\lambda_k\left(\sum_{d=1}^{D}\frac{1}{w_{kd}}-1\right).$$

设 $(\hat{\boldsymbol{w}}_k,\hat{\lambda}_k)$ 最小化 $J_k(\hat{\boldsymbol{w}}_k,\hat{\lambda}_k)$,有:

$$\frac{\partial J_k(\hat{\boldsymbol{w}}_k,\hat{\lambda}_k)}{\partial \hat{w}_{kd}}=\beta\frac{|\hat{c}_k|}{|\hat{c}_k|-1}\hat{w}_{kd}^{-\beta-1}\left(\sum_{o\in O_d}[f_k(o)]^2-\frac{1}{|\hat{c}_k|}\right)-\hat{\lambda}_k\frac{1}{\hat{w}_{kd}^2}=0,$$

$$\frac{\partial J_k(\hat{\boldsymbol{w}}_k,\hat{\lambda}_k)}{\partial \hat{\lambda}_k}=\sum_{d=1}^{D}\frac{1}{\hat{w}_{kd}}-1=0.$$

合并以上两个公式,公式(11)得证. □

3.2 聚类算法

新算法 NMCC 采用第 3.1 节描述的 EM 型优化过程寻求优化目标函数公式(9)的局部优解,算法描述如下:

算法 1. NMCC.

输入:类属型数据集 $DB=\{x_1,x_2,\ldots,x_N\}$及聚类数 K.

输出:聚类集合 $C=\{c_1,c_2,...,c_K\}$及属性权重集合 W.

Begin

令 t 表示算法迭代次数,t=0;

生成数据集初始划分,记为 $C^{(0)}$.

Repeat

1. 设定 $\hat{C}=C^{(t)}$, 使用公式(11)更新属性权重,得到 $W^{(t+1)}$;
2. 设定 $\hat{W}=W^{(t+1)}$, 根据公式(10)将每个数据对象划分到簇,生成新的聚类集合 $C^{(t+1)}$;
3. t=t+1.

Until 聚类集合不发生变化,也就是 $C^{(t)}$=$C^{(t-1)}$.

输出 $C^{(t)}$和 $W^{(t)}$.

End

和其他 EM 型算法一样,NMCC 算法对其初始状态有一定的依赖性.当前,为 EM 型算法设定最优的初始状态是一个开放性的难题[1,2],这里,我们借助 k-modes 算法[8]设定算法所需的初始数据集划分.首先,随机选择 K 个对象为种子;然后,以 SMC 为相似度度量,将所有数据对象划分到最相似的种子,得到一个数据集的划分,将这个划分结果作为 NMCC 算法的初始数据集划分.上述过程相当于 k-modes 算法的一次迭代.

NMCC 算法的时间复杂度为 $O(KNDT)$,其中,T 表示算法的迭代步数.定理 2 表明,T 是有限的,也就是说,NMCC 可以在有限的迭代步骤后收敛.

定理 2. 给定 DB 和 K,NMCC 算法的迭代步数是有限的.

证明:令 t>0 表示算法的一次迭代,$J(C^{(t)},W^{(t)})$是该次迭代结束后的目标函数值.

首先证明 NMCC 算法执行过程中目标函数值递减.算法循环中第 1 个步骤根据定理 1 更新参数 $W^{(t)}$为 $W^{(t+1)}$,因而 $J(C^{(t)},W^{(t)})\geqslant J(C^{(t)},W^{(t+1)})$;第 2 个步骤使用公式(10)按照对象与簇之间距离最近的的原则更新 $C^{(t)}$为 $C^{(t+1)}$,有 $J(C^{(t)},W^{(t+1)})\geqslant J(C^{(t+1)},W^{(t+1)})$,因此,$J(C^{(t)},W^{(t)})\geqslant J(C^{(t+1)},W^{(t+1)})$;另一方面,算法迭代终止条件是 $C^{(t)}$=$C^{(t+1)}$,

根据公式(11),若 $C^{(t)}=C^{(t+1)}$,则 $W^{(t)}=W^{(t+1)}$,这说明当 $J(C^{(t)},W^{(t)})=J(C^{(t+1)},W^{(t+1)})$时算法将终止.综上,在算法迭代中止前,$J(C^{(t)},W^{(t)})>J(C^{(t+1)},W^{(t+1)})$.

其次证明目标函数 $J(C,W)$有下界.令 $\hat{C}=C^{(t+1)}$ 及 $\hat{W}=W^{(t+1)}$, 根据上述结论并代入公式(11),有:

$$J(C^{(t)},W^{(t)})\geqslant\sum_{k=1}^{K}\frac{|\hat{c}_k|}{|\hat{c}_k|-1}\sum_{d=1}^{D}\hat{w}_{kd}^{-\beta}\left(\frac{1}{|\hat{c}_k|}-\sum_{o\in O_d}[f_k(o)]^2\right)=-\sum_{k=1}^{K}\frac{|\hat{c}_k|}{|\hat{c}_k|-1}\left[\sum_{d=1}^{D}\left(\sum_{o\in O_d}[f_k(o)]^2-\frac{1}{|\hat{c}_k|}\right)^{-\frac{1}{\beta-1}}\right]^{1-\beta}.$$

易知,$\frac{1}{|O_d|}\leqslant\sum_{o\in O_d}[f_k(o)]^2\leqslant 1$ 和 $1<|O_d|\leqslant|\hat{c}_k|$, 故 $0\leqslant\sum_{o\in O_d}[f_k(o)]^2-\frac{1}{|\hat{c}_k|}<1$.

又因为β>1,有 $\sum_{d=1}^{D}\left(\sum_{o\in O_d}[f_k(o)]^2-\frac{1}{|\hat{c}_k|}\right)^{-\frac{1}{\beta-1}}>1$, 那么上述不等式可进一步转换为

$$J(C^{(t)},W^{(t)})>-\sum_{k=1}^{K}\frac{|\hat{c}_k|}{|\hat{c}_k|-1}=-K-\sum_{k=1}^{K}\frac{1}{|\hat{c}_k|-1}\geqslant -2K.$$

由于 t 表示任意一次算法的迭代,上式表明,$J(C,W)$存在一个下界$-2K$.前面已经证明,算法迭代过程目标函数值递减,因此,NMCC 算法的迭代步数是有限的. □

3.3 关于属性加权方法的讨论

在 NMCC 聚类过程中,每个属性被自动地赋予一个衡量其重要性的权重.根据定理 1,簇 c_k 属性 d 的权重计算公式为

$$w_{kd}^{(\mathrm{NMCC})}\propto\left(\sum_{o\in O_d}[f_k(o)]^2-\frac{1}{|c_k|}\right)^{\frac{1}{\beta-1}}=\left(\frac{|c_k|-1}{|c_k|}-\left(1-\sum_{o\in O_d}[f_k(o)]^2\right)\right)^{\frac{1}{\beta-1}}.$$

上式中的$1-\sum_{o\in O_d}[f_k(o)]^2$ 是表示类属型数据分布离散程度的基尼系数,由此可知,在 NMCC 算法中,属性权重与数据分布的离散程度成反比.这与属性加权的数值型数据聚类方法的有关结论相一致,事实上,现有的软子空间聚类算法都基于这样的共同假设[3−5]:数据集中某个属性的取值越集中,其重要性就越高.

需要指出的是,NMCC 根据属性值总体分布情况进行属性加权的方式是现有其他算法所不具备的,现有方法仅依据模进行属性加权.以 WKM[9]为代表的一类算法依据数据对象与模之间的平均距离的大小进行属性加权,其权重计算公式为

$$w_{kd}^{(\mathrm{WKM})}\propto\left(\sum_{x\in c_k}(1-SMC(x_d,m_{kd}))\right)^{-\frac{1}{\beta-1}}\propto\left(\frac{1}{1-f_k(m_{kd})}\right)^{\frac{1}{\beta-1}}.$$

MWKM[10]定义了另一种加权方法:赋予每个属性两种权重,计算公式分别为

$$w_{kd}^{(\mathrm{MWKM_1})}\propto\left(\frac{1}{1-f_k(m_{kd})+\frac{T_v}{|c_k|}}\right)^{\frac{1}{\beta-1}},\ w_{kd}^{(\mathrm{MWKM_2})}\propto\left(\frac{1}{f_k(m_{kd})+\frac{T_s}{|c_k|}}\right)^{\frac{1}{\beta-1}},$$

其中,T_v和 T_s是两个常数.

下面通过一个实例对比不同的加权方法.表 2 给出了上述 3 种方法对表 1 所列的 3 个属性计算权重的结果,在这个例子中,取 β=2,并设 MWKM[10]的两个常数 $T_v=T_s$=1.在表 2 中,使用 WKM[9]和 MWKM[10]的加权方法,Attribute 1 和 Attribute 3 的权重相同,这是因为这两个属性的模的频度相同:$f_1(m_{11})=f_1(m_{13})$=0.4,而这两种方法仅依据模来衡量属性的重要性.NMCC 的加权方法成功地区分出了 Attribute 1 和 Attribute 3 对 c_1 重要性的区别,表 1 中 Attribute 3 属性值的分布比 Attribute 1 更为分散,因而 NMCC 赋予 Attribute 1 更大的权重.对于 NMCC 而言,3 个属性的重要性排序为 Attribute 2>Attribute 1>Attribute 3,这与表 1 数据的实际情况相吻合.

Table 2 Attribute-Weighting results yielded by different methods for the data in Table 1

表 2 各种方法对表 1 数据的属性加权结果

属性加权方法	Attribute 1	Attribute 2	Attribute 3
WKM	0.20	0.60	0.20
$MWKM_1$	0.25	0.50	0.25
$MWKM_2$	0.38	0.24	0.38
NMCC	3.33	10	1.67

4 实验与分析

实验验证包括算法有效性和算法效率两个方面,并与若干相关工作相比较.实验设备为配置 2.4GHz CPU 和 3GB RAM 的计算机.

4.1 实验设置

实验选择 *k*-modes[8](简称 KM)、加权 *k*-modes[9](简称 WKM)、模加权的 *k*-modes[10](简称 MWKM)和基于熵的算法 EBC[17]这 4 种代表性算法为对比对象.KM[8]是一种经典的类属数据聚类算法.它较早地提出使用模表示类属数据中心的思路.WKM[9]在 KM 基础上引入基于距离的属性加权方法.MWKM 算法[10]根据模的频度进行属性加权.WKM 和 MWKM 代表了当前属性加权的类属数据聚类两种主流方法,但它们与 KM 一样都以模为中心对类属型数据进行类 *k*-means 聚类.在非模聚类算法中选择了 EBC[17]作对比,一方面,它代表了类属数据聚类领域的一类方法:以数据集的划分熵(类属数据分布的信息熵)为优化目标的方法;另一方面, EBC 采用了不同于其他 4 种算法(包括本文的 NMCC 算法)的 EM 型算法结构:蒙特卡罗抽样方法.选择这些算法使得我们可以在相同数据集上比较不同簇类表示方法、不同属性加权方法和不同优化过程的性能.

各种算法的性能将在合成数据集和若干实际应用数据集上进行测试.在合成数据集上进行实验对比的好处在于,便于控制数据集的簇结构,如簇的数目、相关维度的比例和类属属性的符号数目等,以分析不同算法对各样数据集的适应性以及算法性能与以上参数之间的关系.WKM[9]和 MWKM 算法[10]的参数β根据作者的建议值设置,即β=2 和 T_s=T_v=1.EBC 算法[17]基于蒙特卡罗抽样,实验设定算法在连续 z 次随机移动对象而无法继续提高聚类质量时终止.显然,z 越大,EBC 的聚类结果质量可能就越高,但这同时会带来聚类效率急剧下降的后果.为了给 EBC 在聚类效率和聚类质量之间取一个平衡点,在实验数据集(见第 4.2 节和第 4.3 节)上测试了 z 从 50 到 200 之间变化时的 EBC 算法表现,测试结果显示:在这些数据集上,当 z>100 时,聚类质量并没有得到明显的提高,却大幅度增加了算法执行时间,因此,以下设定 z=100.我们将在合成数据集上对 NMCC 算法的性能与参数β取值之间的关系进行实验分析(详见第 4.3 节).

实验采用两种指标衡量各种算法的聚类质量:CU 指标和 *F*-Score 指标.CU 指标是一种评价聚类质量的内部指标,定义[10]如下:

$$CU=\sum_{k=1}^{K}\frac{|c_k|}{N}\sum_{d=1}^{D}\sum_{o\in O_d}([f_k(o)]^2-[f(o)]^2).$$

其中,$f(o)$表示符号 o 在整个数据集上的频度.从上式可以看出,CU 指标衡量的是簇的各属性包含类属值的“纯度”,其数值越大,表明聚类质量越高.*F*-Score 指标通过比较聚类结果与对象实际类标号之间的差异来衡量算法的聚类精度,其定义[4]为

$$F\text{-}Score=\sum_{k=1}^{K}\frac{n_k}{N}\max_{1\leqslant l\leqslant K}\left(\frac{2\times Recall(class_k,c_l)\times Precision(class_k,c_l)}{Recall(class_k,c_l)+Precision(class_k,c_l)}\right).$$

这里,$Recall(class_k,c_l)$和 $Precision(class_k,c_l)$分别是数据集真实的类 $class_k$ 与聚类算法返回的第 l 个簇 c_l 相比较的召回率(计算公式为 n_{kl}/n_k)和精度(计算公式为 $n_{kl}/|c_l|$),n_k 表示 $class_k$ 包含的对象数目,n_{kl} 是同时出现在 $class_k$ 和 c_l 中的对象数目.*F*-Score 指标是一种评价聚类质量的外部指标(因此没有类别标号的数据无法使用该指标评价聚类结果的质量),越大的值对应于越好的聚类质量.

4.2 合成数据实验

实验采用的合成数据根据文献[5]提供的方法生成.该方法使用以下参数生成数据:N(对象数目)、D(属性数目)、K(聚类数目)、l(相关维度数目)和用于控制相关维度上属性值分布方差(文献[5]建议的方法只能合成数值型属性)的 r 和 s.在此基础上,再对合成得到的数值型属性进行等宽离散化处理,生成类属型数据,引入一个新的参数$|O|$表示离散化后每个维度包含的类属符号数目.根据以上方法,设定 K=5 和 r=s=2 合成了两组数据.第 1 组用于测试各种算法对类属符号数目变化的适应性以及 NMCC 算法对参数β取值的敏感性,为此,设定 N=5000,D=60,l=12 合成了 5 个数据集,数据集的参数$|O|$分别为 2,4,6,8 和 10;第 2 组数据用于检验各种算法相对于数据维度的可伸缩性,设定 N=5000,$|O|$=4 和 D 取 20,30,40,50,60 及 l=D×50%合成了 5 个数据集.这些合成数据集中的数据都包含有正确的类标,正确的聚类数目也是已知的,实验中,5 种算法所需的聚类数目参数 K 均设置为这些正确的聚类数目.

图 1 显示 NMCC 算法聚类第 1 组数据($|O|$=4)取得的 CU 和 F-Score 指标值与算法参数β取值之间的关系.图中的两个指标值是算法在该数据集上独立运行 10 次后求取的平均值.如图所示,当$\beta\geq 6$ 时,两个指标值都没有出现明显的变化,这说明当$\beta\geq 6$ 时,NMCC 的性能对算法参数是鲁棒的.根据这个结果,在本节下面的实验中,设定 NMCC 算法的参数β=6.

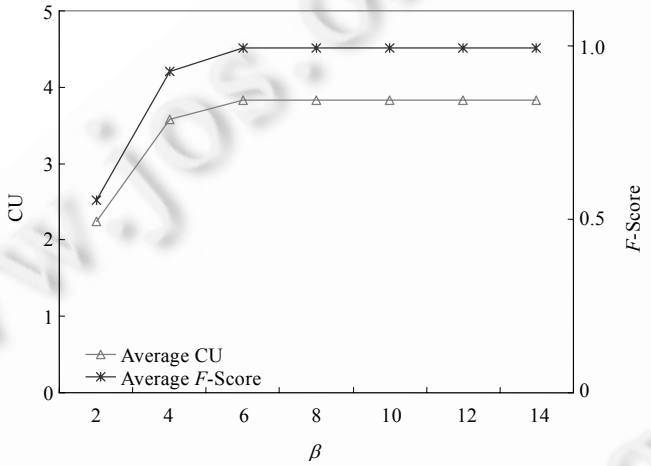

Fig.1 Change in clustering quality of NMCC with different values of the parameter β, on the synthetic datasets

图 1 合成数据集上 NMCC 聚类质量随参数值β变化情况

图 2 显示各种算法在第 1 组数据集上获得的聚类质量评价指标值.图中报告的数据同样是各种算法分别独立运行 10 次的平均结果.图 2 显示,NMCC 算法在类属值数目从 2~10 变化的所有数据集上都获得了最高的聚类质量,当符号数目较少($|O|\leq 8$)时,NMCC 的 CU 指标和 F-Score 指标值超出基于模的 3 种聚类算法(KM,WKM 和 MWKM)的幅度达 50%以上.各种算法的性能随$|O|$的变化大体呈下降趋势,其中,非模算法 EBC 聚类精度(F-Score 指标)随$|O|$变化显得较为稳定,其部分原因在于 EBC 采用蒙特卡罗抽样方法(而非其他 4 种算法使用的 EM 型结构)搜索聚类模型的局部优解,但是如图所示,多数情况下,NMCC 获得的聚类精度还是超出 EBC 算法 40%以上.NMCC 的性能优势源自两个方面:首先,NMCC 是一种非模聚类算法,从而避免了因基于模进行优化导致的算法易陷于局部优解的问题[1];其次,NMCC 在聚类过程中根据类属值总体分布进行软特征选择,与 WKM 和 MWKM 等算法依赖于模的属性加权方案相比,NMCC 的加权方法可以更准确地捕捉到属性对簇不同的重要性.本组数据中包含的近 80%的噪声属性(l=D×20%)是导致未进行自动特征选择的 EBC 算法聚类性能下降的另一个因素.

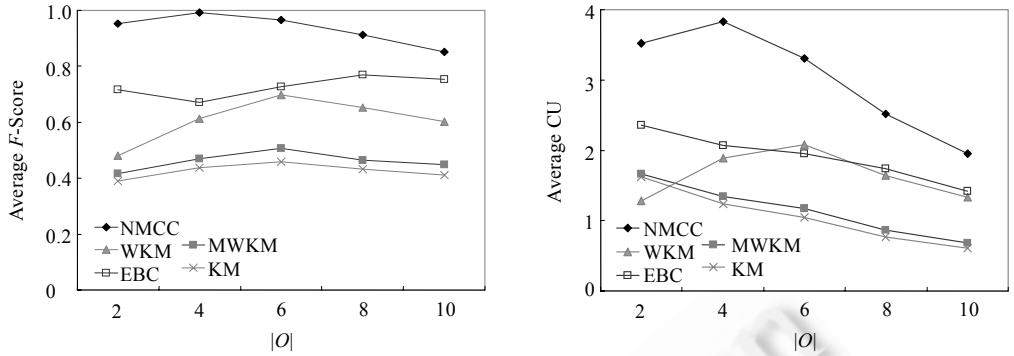

Fig.2 Change in clustering quality of different algorithms with various numbers of categorical values

图 2 各种算法的聚类质量随类属值数目变化情况

图 3 显示了各种算法平均聚类时间的对比结果.在这组实验中,第 1 组合成数据用于测试算法的平均运行时间与类属值数目之间的关系(左图),使用第 2 组数据测试算法相对于数据维度的可伸缩性(右图).从图 3 可知,KM,WKM 和 MWKM 这 3 种算法具有较高的聚类效率,这正是基于模的聚类算法的一个优势所在;EBC 算法使用蒙特卡罗抽样技术需要较多的时间,搜索最优解;NMCC 算法的时间性能介于二者之间,这是因为 NMCC 中没有模的概念,需要更多的时间,根据对象与簇类间的相似性进行数据集划分;NMCC 采用了 EM 型的算法结构,因而获得比 EBC 高的聚类效率.此外,如图所示,随着类属值数目及属性数目的增大,NMCC 算法花费的时间接近于线性增长,这说明 NMCC 具有良好的算法可伸缩性.

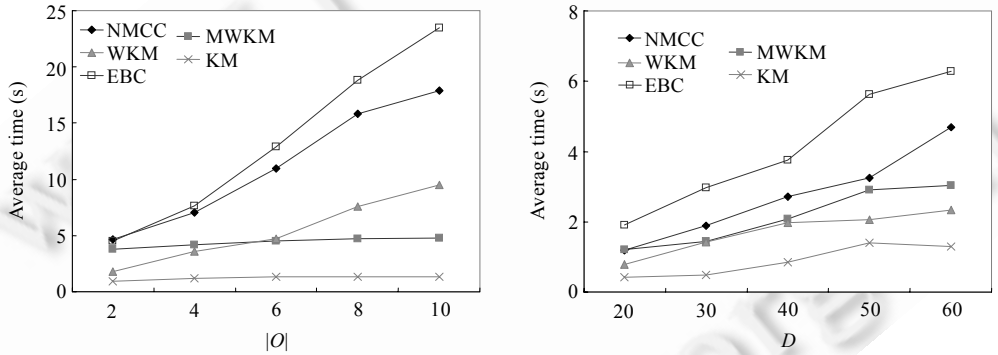

Fig.3 Relationships between runtime of different algorithms, and different numbers of categorical values and attributes

图 3 各种算法的运行时间与类属值数目及属性数目之间的关系

4.3 实际应用数据实验

本组实验的目的是在实际应用数据集上检验 NMCC 算法的聚类性能,并与其他 4 种算法 2 做对比.实验使用 7 个来自 UCI machine learning repository(ftp.ics.uci.edu:pub/machine-learning-databases)的常用数据集,数据集的有关信息参见表 3.

第 1 个实际数据集是 Breastcancer 是临床肺癌诊断数据,由 9 个类属型属性和良性(458 个样本)、恶性(241 个样本)两个类别组成,其中 2 个属性包含缺失数据.淋巴系造影术数据 Lymphography 包括 15 个类属型和 3 个数值型属性,实验仅使用其中的类属型属性,并移除其中两个较小的类(2 个样本的 normal find 类和 4 个样本的 fibrosis 类).数据集 Vote 来自美国国会投票记录,其 16 个类属型属性中有 15 个包含缺失数据.蘑菇数据集 Mushroom 包含的样本数较多,两个类别 edible 和 poisonous 各有 4 208 和 3 916 个对象,每个对象由 22 个类属

属性描述,其中的 veil-type 属性因取值唯一在实验中予以剔除.数据集 Dermatology 用于医疗领域皮肤病诊断,包括 6 个类别,每个类别的样本数呈现出分布不均衡的特点.剩下的两个数据集相对高维,其中,USCensus10k 从 US Census(1990)数据库中随机抽取 10 000 条记录而成,每条记录拥有 68 个类属属性;另一个数据集是保险公司客户数据,曾用于 CoIL Challenge 2000 竞赛(因此简称 Coil2000),每个客户由 86 个类属属性描述.这两个相对高维的数据集不包含类别标号.

Table 3 Summary of the parameters for the real-world datasets

表 3 实际数据集参数汇总表

数据集	属性数目(D)	簇数目(K)	数据对象数目(N)	各类包含的对象数目
Breastcancer	9	2	699	458:241
Lymphography	15	2	142	61:81
Vote	16	2	435	168:267
Mushroom	21	2	8 124	3916:4208
Dermatology	33	6	366	61:112:72:52:49:20
USCensus10k	68	N/A	10 000	N/A
Coil2000	86	N/A	5 822	N/A

对于前 5 个类别标号已知的数据集,算法所需的聚类数目参数均设为表 3 所列的 K 值,并使用 F-Score 和 CU 两个指标分别评价各算法在这 5 个数据集上聚类结果的质量.数据具有的类标号仅仅用于计算 F-Score 指标值,在聚类过程中,类标号对各种算法都是未知的.对于不含类别标号的 USCensus10k 和 Coil2000,实验分别设定 K=2 和 K=3 调用各聚类算法,使用内部指标 CU 评价它们的聚类质量.实验数据集中的所有缺失数据都用一个特别的符号代替.例如,Mushroom 数据集中命名为 stalk-root 的属性取值范围为{‘b’,‘c’,‘u’,‘e’,‘z’,‘r’},共有 2 480 个样本的该属性值缺失,此时引入一个新的符号‘?’代替这些缺失数据,插入新符号之后的类属符号集合变为{‘b’,‘c’,‘u’,‘e’,‘z’,‘r’,‘?’},即认为这 2 480 个样本的属性 stalk-root 都取‘b’,‘c’,‘u’,‘e’,‘z’,‘r’之外的第 7 个离散值‘?’.

图 4 显示了 NMCC 算法在这些实际数据集上分别运行 10 次取得的平均聚类结果与算法参数β之间的关系.图示数据表明,NMCC 在实际数据集上的聚类结果相对于参数β的变化是比较鲁棒的,当$\beta\geqslant 6$ 时,随着β值的增加,聚类结果趋向稳定.根据这个结果,以下实验均设定β=6,这与第 4.2 节合成数据的实验设置相一致.

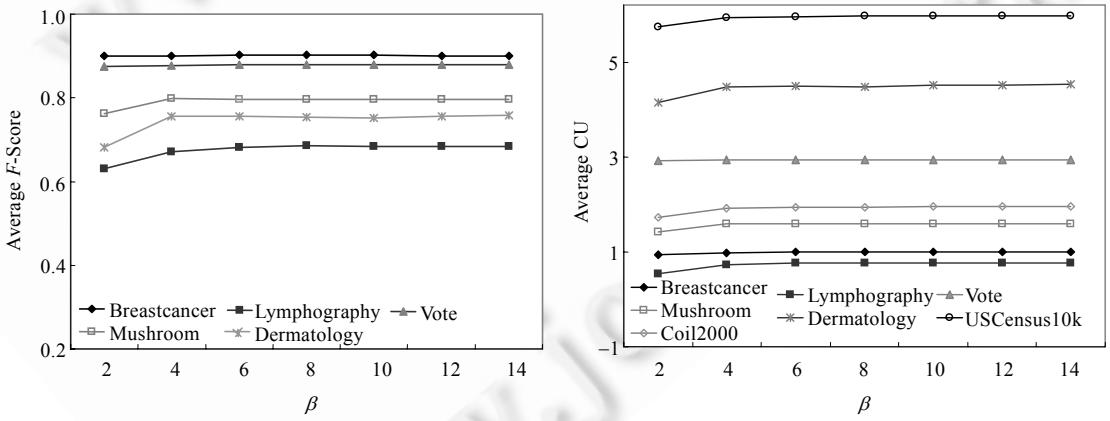

Fig.4 Change in clustering quality of NMCC with different values of the parameter β on the real-world datasets

图 4 实际数据集上 NMCC 聚类质量随参数值β变化情况

表 4 和表 5 所列为 5 种算法在实际数据集上获得的聚类结果,两个表分别报告 F-Score 指标和 CU 指标值.表中的“Max.”列为各种算法在对应数据集上独立运行 100 次间的最好结果,“Avg.”列为对应的平均结果,以“均值±1 个标准差”格式提供.由于 KM,WKM,MWKM 以及 NMCC 算法选择算法初始状态时具有一定的随机性

(KM,WKM 和 MWKM 随机选择各簇类的初始中心;NMCC 使用 KM 算法的一次迭代来确定数据集初始划分),EBC 使用的蒙特卡罗抽样方法本身就具有随机性,这些都导致算法每次执行结果的差异.表 4 和表 5 报告的均值反映了各种算法的总体性能,而所列的标准差可以作为判断各种算法稳定性的依据.对每个数据集,最高的“Avg.”指标值加粗标注在表中.

Table 4 Comparison of *F*-Score yielded by different algorithms, on the real-world datasets

表 4 实际数据集上不同算法的 *F*-Score 指标对比

数据集	NMCC		KM		WKM		MWKM		EBC	
	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.
Breastcancer	0.91	0.90±0.00	0.95	0.81±0.15	0.88	0.76±0.04	0.93	0.85±0.13	0.96	**0.94±0.02**
Lymphography	0.75	**0.68±0.03**	0.76	0.63±0.05	0.77	0.66±0.03	0.79	0.64±0.05	0.84	0.63±0.07
Vote	0.88	**0.88±0.00**	0.87	0.86±0.01	0.94	0.82±0.08	0.87	0.86±0.00	0.88	0.86±0.02
Mushroom	0.89	**0.78±0.15**	0.89	0.70±0.13	0.87	0.67±0.06	0.89	0.71±0.14	0.86	0.66±0.06
Dermatology	0.87	**0.72±0.07**	0.83	0.63±0.08	0.80	0.60±0.10	0.85	0.65±0.09	0.81	0.64±0.07

Table 5 Comparison of CU yielded by different algorithms, on the real-world datasets

表 5 实际数据集上不同算法的 CU 指标对比

数据集	NMCC		KM		WKM		MWKM		EBC	
	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.	Max.	Avg.
Breastcancer	1.01	0.99±0.01	1.14	0.74±0.42	0.96	0.56±0.13	1.04	0.90±0.35	1.21	**1.11±0.06**
Lymphography	0.80	0.75±0.08	0.81	0.70±0.13	0.79	0.35±0.17	0.83	0.72±0.13	0.85	**0.77±0.10**
Vote	2.93	**2.93±0.00**	2.90	2.90±0.01	2.79	2.45±0.54	2.91	2.90±0.00	2.92	2.73±0.13
Mushroom	1.73	**1.57±0.24**	1.73	1.36±0.30	1.58	0.71±0.47	1.73	1.44±0.22	1.50	0.95±0.27
Dermatology	4.65	**4.32±0.78**	4.63	3.90±0.60	3.96	2.52±0.79	4.66	3.96±0.57	4.47	3.92±0.29
USCensus10k	5.95	**5.95±0.00**	5.90	5.77±0.10	5.92	1.52±2.27	5.90	5.73±0.51	5.15	4.24±0.98
Coil2000	1.97	**1.93±0.05**	1.76	1.58±0.13	0.61	0.04±0.10	1.75	1.56±0.14	1.76	1.35±0.18

从表 4 和表 5 可以看出,NMCC 算法在 5 个数据集上均获得了较高的聚类质量,与对比算法相比,多数情况下都取得了明显的聚类质量提升,尤其在样本数较多的 Mushroom、类别数最多的 Dermatology 和相对高维的 USCensus10k 和 Coil2000 两个数据集上.这进一步验证了 NMCC 算法对特性各异的类属数据集的适应性.表中数据也显示 NMCC 具有比其他算法更稳定的性能(体现在平均精度的标准差上),特别是对于较高维的数据.一方面的原因在于 NMCC 是非模算法,避免了 *k*-modes 型其他算法在聚类过程仅考虑模而易陷于局部最优的问题;另一方面,NMCC 的算法初始状态是 K 个数据子集(每个子集由一组相似对象组成),而不是其他算法所依赖的对象个体(*k*-modes 型算法的初始簇中心是 K 个随机选择的对象;EBC 的每次迭代步骤随机地选择某个簇类中的对象),从而降低了算法对初始状态的敏感性,提高了算法的稳定性.

正如文献[10]所指出的,WKM 算法采用的属性加权方法在某种程度上反而降低了重要属性上对象间的差异,这个缺陷导致 WKM 在多数数据集上的聚类精度反而低于 KM 算法.仔细分析 Breastcancer 数据可知,该数据集虽然维度较低(只有 9 个属性),但所有属性都有多达 10 个的类属型符号,这个特性直接影响了 KM,WKM 和 MWKM 这 3 种算法的性能,这是因为这 3 种算法基于模进行聚类优化,并根据模的频度计算属性与簇之间的相关性,当属性取值较多时,这种估计方法极易导致偏差.非模的 NMCC 算法根据类属值总体分布进行聚类优化和特征选择,因而可以取得更准确的聚类结果.EBC 算法虽然在该数据上获得了最高的平均 *F*-Score 和 CU 指标值,但在 Mushroom 数据上却表现最差.究其原因,Mushroom 数据 21 个属性的统计特性存在明显差别,NMCC 通过属性加权区分出 bruises(第 4 个属性)和 veil-color(第 16 个属性)等重要的属性,自动赋予它们较大的权重,从而降低其他属性对相似度计算的影响.EBC 算法虽然弥补了 KM,WKM 和 MWKM 等基于模的聚类方法的缺陷,但却未能区分属性的这种差异.从这个意义上说,NMCC 算法兼具了两种类型聚类算法的优点,在实际应用数据上可以获得高且较稳定的聚类质量.

5 总结及今后的工作方向

在类属型数据聚类领域,当前多数聚类方法以类属属性的模为基础,在这些方法中,非模属性被忽略,极易

导致聚类优化和属性重要性权重计算上的偏差.本文提出一种自动属性加权的聚类新方法,用于类属型数据无监督的统计学习.以簇类整体统计信息(不仅限于属性的模)为基础,推导了一个新的聚类目标函数,提出一种子空间聚类算法 NMCC 来优化目标函数.与层次聚类、蒙特卡罗聚类等非模聚类方法相比,NMCC 具有时间复杂度低的优点.在聚类过程中,算法根据类属属性值分布的离散程度自动赋予每个属性反映其重要性的权重,实现自动的软特征选择.新方法克服了现有主流方法仅依据模进行属性加权的缺陷,与数值型数据聚类领域得到的有关理论结果相一致.在多个合成数据和实际应用数据集上进行了实验验证,结果表明新方法是有效的,与以模为中心的现有方法相比,新方法在实验数据上的聚类质量得到较为明显的改善.

后续将开展以下几个方面的研究工作:探讨选择初始聚类划分和估计算法参数 β 取值的方法,以进一步提高聚类算法的鲁棒性;将新方法推广到混合型(混合了数值型和类属型属性的)数据聚类.当前对两种类型数据的聚类处理方法基本上是分立的,本文的一些结果似乎为合并处理两种类型数据提供了一条途径:类属数据的聚类目标可以定义为对方差(类属数据的 Gini index)的优化,类属型属性权重也可以依据数据分布的集中程度来计算(具有较小分布方差的属性获得较大的权重),这些都与数值型数据聚类的有关方法相一致(见 Huang 等人在文献[3]中的讨论).但是,如何平衡不同类型属性的权重数值仍是一个问题,这也将是下一步工作的重点.

References:

[1] Jain A, Murty M, Flynn P. Data clustering: A review. ACM Computing Survey, 1999,31(3):264−323. [doi: 10.1145/331499.331504]

[2] Sun J, Liu J, Zhao L. Clustering algorithms research. Ruan Jian Xue Bao/Journal of Software, 2008,19(1):48−61 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/48.htm [doi: 10.1360/jos190048]

[3] Huang JZ, Ng MK, Rong H, Li Z. Automated variable weighting in *k*-means type clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(5):657−668. [doi: 10.1109/TPAMI.2005.95]

[4] Jing L, Ng MK, Huang JZ. An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans. on Knowledge and Data Engineering, 2007,19(8):1−16. [doi: 10.1109/TKDE.2007.1048]

[5] Chen L, Jiang Q, Wang S. Model-Based method for projective clustering. IEEE Trans. on Knowledge and Data Engineering, 2012, 24(7):1291−1305. [doi: 10.1109/TKDE.2010.256]

[6] Gao J, Wang S. Fuzzy clustering algorithm with ranking features and identifying noise simultaneously. Acta Automatica Sinica, 2009,35(2):145−153 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2009.00145]

[7] Liang J, Bai L, Cao F. *K*-Modes clustering algorithm based on a new distance measure. Journal of Computer Research and Development, 2010,47(10):1749−1755 (in Chinese with English abstract).

[8] Huang JZ, Ng MK. A note on k-modes clustering. Journal of Classification, 2003,20(2):257−261. [doi: 10.1007/s00357-003-0014-4]

[9] Chan EY, Ching WK, Ng MK, Huang JZ. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recognition, 2004,37(5):943−952. [doi: 10.1016/j.patcog.2003.11.003]

[10] Bai L, Liang J, Dang C, Cao F. A novel attribute weighting algorithm for clustering high-dimensional categorical data. Pattern Recognition, 2011,44(12):2843−2861. [doi: 10.1016/j.patcog.2011.04.024]

[11] Lee M, Pedrycz W. The fuzzy *C*-means algorithm with fuzzy *P*-mode prototypes for clustering objects having mixed features. Fuzzy Sets and Systems, 2009,160(24):3590−3600. [doi: 10.1016/j.fss.2009.06.015]

[12] San OM, Huynh VN, Nakamori Y. An alternative extension of the *k*-means algorithm for clustering categorical data. Int'l Journal of Applied Mathematics and Computer Science, 2004,14(2):241−247.

[13] Guha S, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes. In: Kitsuregawa M, Maciaszek L, Papazoglou M, Pu C, eds. Proc. of the 15th Int'l Conf. on Data Engineering. Los Alamitos: IEEE Computer Society, 1999. 512−521. [doi: 10.1109/ICDE.1999.754967]

[14] Xiong T, Wang S, Mayers A, Monga E. DHCC: Divisive hierarchical clustering of categorical data. Data Mining and Knowledge Discovery, 2012,24(1):103−135. [doi: 10.1007/s10618-011-0221-2]

[15] Light RT, Marglin BH. An analysis of variance for categorical data. Journal of the American Statistical Association, 1971,66(335): 534−544.

[16] Sen PK. Gini diversity index, Hamming distance and curse of dimensionality. Metron—Int'l Journal of Statistics, 2005,LXIII(3): 329−349.

[17] Li T, Ma S, Ogihara M. Entropy-Based criterion in categorical clustering. In: Brodley CE, ed. Proc. of the 21st Int'l Conf. on Machine Learning. Banff: ACM Press, 2004. 536−543. [doi: 10.1145/1015330.1015404]

[18] Cesario E, Manco, Ortale R. Top-Down parameter-free clustering of high-dimensional categorical data. IEEE Trans. on Knowledge and Data Engineering, 2007,19(12):1607−1624. [doi: 10.1109/TKDE.2007.190649]

[19] Dickman BH, Gilman MJ. Monto Carlo optimization. Journal of Optimization Theory and Applications, 1989,60(1):149−157.

[20] Xu L, Jordan MI. On convergence properties of the EM algorithm for gaussian mixtures. Neural Computation, 1996,8(1):129−151. [doi: 10.1162/neco.1996.8.1.129]

[21] Kim DE, Lee K, Lee D, Lee KH. A *k*-populations algorithm for clustering categorical data. Pattern Recognition, 2005,38(7): 1131−1134. [doi: 10.1016/j.patcog.2004.11.017]

[22] Ouyang D, Li Q, Racine J. Cross-Validation and the estimation of probability distributions with categorical data. Nonparametric Statistics, 2006,18(1):69−100. [doi: 10.1080/10485250600569002]

附中文参考文献:

[2] 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1):48−61. http://www.jos.org.cn/1000-9825/19/48.htm [doi:10.1360/jos190048]

[6] 皋军,王士同.具有特征排序功能的鲁棒性模糊聚类方法.自动化学报,2009,35(2):145−153.

[7] 梁吉业,白亮,曹付元.基于新的距离度量的 *K*-Modes 聚类算法.计算机研究与发展,2010,47(10):1749−1755.

陈黎飞(1972－),男,福建长乐人,博士,副教授,主要研究领域为数据挖掘,机器学习.
E-mail: clfei@fjnu.edu.cn

郭躬德(1965－),男,博士,教授,博士生导师,主要研究领域为人工智能,数据挖掘.
E-mail: ggd@fjnu.edu.cn