



文章编号:1001-9081(2017)04-1026-06

DOI:10.11772/j.issn.1001-9081.2017.04.1026

## 类属数据的贝叶斯聚类算法

朱杰<sup>1</sup>, 陈黎飞<sup>2\*</sup>

(1. 中国西南电子技术研究所, 成都 610036; 2. 福建师范大学 数学与计算机科学学院, 福州 350117)

(\*通信作者电子邮箱 clfei@fjnu.edu.cn)

**摘要:**针对类属型数据聚类中对象间距离函数定义的困难问题,提出一种基于贝叶斯概率估计的类属数据聚类算法。首先,提出一种属性加权的概率模型,在这个模型中每个类属属性被赋予一个反映其重要性的权重;其次,经过贝叶斯公式的变换,定义了基于最大似然估计的聚类优化目标函数,并提出了一种基于划分的聚类算法,该算法不再依赖于对象间的距离,而是根据对象与数据集划分间的加权似然进行聚类;第三,推导了计算属性权重的表达式,得出了类属型属性权重与其符号分布的信息熵成反比的结论。在实际数据和合成数据集上进行了实验,结果表明,与基于距离的现有聚类算法相比,所提算法提高了聚类精度,特别是在生物信息学数据上取得了5%~48%的提升幅度,并可以获得有实际意义的属性加权结果。

**关键词:**数据聚类;类属型属性;属性加权;贝叶斯聚类;概率模型

中图分类号:TP274.2 文献标志码:A

### Bayesian clustering algorithm for categorical data

ZHU Jie<sup>1</sup>, CHEN Lifei<sup>2\*</sup>

(1. Southwest China Institute of Electronic Technology, Chengdu Sichuan 610036, China;

2. School of Mathematics and Computer Science, Fujian Normal University, Fuzhou Fujian 350117, China)

**Abstract:** To address the difficulty of defining a meaningful distance measure for categorical data clustering, a new categorical data clustering algorithm was proposed based on Bayesian probability estimation. Firstly, a probability model with automatic attribute-weighting was proposed, in which each categorical attribute is assigned an individual weight to indicate its importance for clustering. Secondly, a clustering objective function was derived using maximum likelihood estimation and Bayesian transformation, then a partitioning algorithm was proposed to optimize the objective function which groups data according to the weighted likelihood between objects and clusters instead of the pairwise distances. Thirdly, an expression for estimating the attribute weights was derived, indicating that the weight should be inversely proportional to the entropy of category distribution. The experiments were conducted on some real datasets and a synthetic dataset. The results show that the proposed algorithm yields higher clustering accuracy than the existing distance-based algorithms, achieving 5% – 48% improvements on the Bioinformatics data with meaningful attribute-weighting results for the categorical attributes.

**Key words:** data clustering; categorical attribute; attribute weighting; Bayesian clustering; probability model

### 0 引言

在模式识别和数据挖掘领域,类属数据聚类(categorical data clustering)是一项重要但较为困难的任务:一方面,这是由于许多实际应用中的待聚类对象通常由该型数据或混合了数值型(numerical)及类属型的数据描述。例如,生物数据聚类任务中的DNA序列就是由A、T、G和C等代表不同氨基酸的符号构成的;在描述患者体征时,又可能使用诸如“心跳数”这样的数值型生理指标,而实际应用中人们通常将这些数值型数据转换为类属型数据加以处理<sup>[1]</sup>。另一方面,由于类属数据的属性值取自有限的符号集合,是离散的,定义有效的对象间距离度量较数值型数据显得困难<sup>[2]</sup>,这使得类属数据聚类成为一项富有挑战性的任务。

根据所生成聚类结构的差异,现有类属数据聚类算法大

致分为两类<sup>[3]</sup>:层次聚类和基于划分的聚类。前者的目的是构造层次聚类树,代表性算法包括凝聚型算法<sup>[4]</sup>和分裂型算法<sup>[5]</sup>等。本文着重于基于划分的聚类,主要原因是该型方法(与层次聚类算法相比)通常具有较低的时间复杂度且易于实现。实际上,以著名的K-means<sup>[6]</sup>为代表的基于划分的聚类算法已被广泛研究和应用,其基本原理是在给定聚类数K的前提下寻求数据集中簇内对象平方误差(squared error)最小的K个划分,这里的误差是依据对象与簇中心之间的距离定义的。

将K-means型算法运用于类属数据聚类需要处理两个主要问题:如何定义类属数据的簇中心以及如何衡量类属对象间的距离(或相似度)。对于第一个问题,已提出基于符号分布的“中心”<sup>[7-8]</sup>和基于“模”的定义<sup>[9-10]</sup>等,其中又以后者最为常见:使用模符号(mode category),即出现频率最高的那个

收稿日期:2016-09-12;修回日期:2016-12-23。

基金项目:国家自然科学基金资助项目(61175123);福建省自然科学基金资助项目(2015J01238)。

作者简介:朱杰(1971—),男,浙江余姚人,高级工程师,主要研究方向:模式识别、目标识别; 陈黎飞(1972—),男,福建长乐人,教授,博士,主要研究方向:统计机器学习、数据挖掘、模式识别。



符号作为簇的代表。典型算法包括著名的  $K$ -modes<sup>[9,11]</sup> 及其变种<sup>[10,12]</sup>。尽管已提出多种衡量类属对象间相似性的方法,例如信息熵度量<sup>[13]</sup>和频度度量<sup>[14]</sup>等,但应用于无监督聚类的并不多见。其中的简单匹配(simple matching)距离具有代表性,其原理是根据取值不同的属性数目计算对象距离。近年已提出多种基于属性加权的简单匹配距离度量,并以此为基础定义了多种  $K$ -modes 型子空间聚类算法<sup>[7,15-17]</sup>。由于这些算法所依赖的距离度量仅限于符号匹配,而忽略了符号的总体分布(一般而言,一个有效的度量应区分高频符号与低频符号等<sup>[2]</sup>),这是不完备的,聚类算法的有效性也将因此受到影响。

本文提出的新型聚类算法基于数据对象与数据集划分之间的似然(likelihood)衡量对象与簇之间的相似性,从而避免了由简单匹配距离和以模为簇中心带来的上述问题。该算法简称为 WBCC(Weighted Bayesian Clustering of Categories),是一种基于贝叶斯概率估计的类属数据聚类算法。WBCC 算法通过对新定义的属性加权概率模型的最大似然估计,实现类属数据的子空间聚类。在实际数据集上的实验结果验证了该算法的有效性。

## 1 相关工作

首先约定后文使用的记号。设  $\mathbf{x} = (x_1, x_2, \dots, x_d, \dots, x_D)$  或  $\mathbf{y} = (y_1, y_2, \dots, y_d, \dots, y_D)$  表示由  $D$  个类属属性值构成的数据对象,属性值  $x_d$  ( $d = 1, 2, \dots, D$ ) 的取值集合为  $X_d$ ,  $|X_d|$  表示集合中的元素数目,即  $x_d$  可取的符号数目。给定  $N$  个这样的类属对象和整数  $K$  ( $1 < K < N$ ),本文讨论的聚类算法的目的是将  $N$  个对象划分为  $K$  个簇的集合  $C = \{c_1, c_2, \dots, c_k, \dots, c_K\}$ ,  $c_k$  表示第  $k$  个簇,其包含的对象数记为  $|c_k|$ 。

对象间的相似性或距离度量是聚类分析的基础。在类属型数据聚类中,常用简单匹配函数  $dis(\mathbf{x}, \mathbf{y})$  衡量对象  $\mathbf{x}$  和  $\mathbf{y}$  之间的距离<sup>[1-2]</sup>,注意到此型度量仅考虑对象本身符号的匹配情况。

$$\begin{aligned} dis(\mathbf{x}, \mathbf{y}) &= \sum_{d=1}^D \delta(x_d, y_d) \\ \delta(x_d, y_d) &= \begin{cases} 1, & x_d \neq y_d \\ 0 & x_d = y_d \end{cases} \end{aligned} \quad (1)$$

Lin<sup>[13]</sup>根据(被匹配)符号的概率定义了基于熵理论的相似性度量,Goodall<sup>[14]</sup>的定义则使用了所有符号的概率(因此,该度量考虑了符号总体分布情况),但它们都未能用于基于划分的数据聚类。实际上,基于划分的聚类算法的优化目标与所采用对象间距离度量是密切相关的,例如,已广泛研究和应用的  $K$ -modes 系列算法<sup>[9-12]</sup>就基于式(1)所示的简单匹配距离定义其目标函数。

近年对基于划分的类属数据聚类算法研究主要集中在两个方面:提出替代“模”的簇表示方法以及改进距离度量。“非模”的簇表示方法包括符号分布法<sup>[7]</sup>和  $K$ -representatives<sup>[8]</sup>等,其实质是将一个类属型属性转换为若干个(等于该属性的符号数目)二元型属性加以处理。Chen 等<sup>[17-18]</sup>研究了非中心聚类法,在这种方法中,簇不再以“中心”为代表。本文提出的新算法 WBCC 也将是“非中心”的,直接使用簇内对象来表示簇。

对距离度量的改进多集中在属性加权的简单匹配距离

上,不同方法采用的属性加权方式有所差异。例如,WKP(W-K-Prototypes)算法<sup>[19]</sup>赋予每个属性  $d$  以一个全局权重  $\omega_d$ ,从而改进式(1)为  $\omega_d^\beta \times \delta(x_d, y_d)$ 。混合加权  $K$ -modes(Mixed Weighting  $K$ -modes, MWKM)<sup>[15]</sup>、互补熵加权  $K$ -modes(Complete-entropy Weighting  $K$ -modes, CWKM)<sup>[16]</sup>等新近提出的算法则使用了局部属性加权机制,赋予每个簇的每个属性一个权重,进而在投影子空间中进行聚类<sup>[20]</sup>。在这些算法中,权重计算方法大致归为两类:基于模频度<sup>[5,15,19]</sup>和基于符号分布的加权方式<sup>[13,16-18]</sup>,后者利用了所有符号(而不仅仅是模符号)的统计信息,通常可取得更好的聚类效果。本文算法中的属性加权方法也属于后者,根据符号分布的信息熵进行属性加权,所不同的是,其权重计算表达式是基于贝叶斯聚类模型推导而来的。

## 2 贝叶斯聚类算法

### 2.1 贝叶斯聚类模型

给定对象集和聚类数  $K$ ,贝叶斯聚类算法的目的是生成  $K$  个簇的集合  $C$ ,以最大化每个对象(相对于其所在簇)的似然。该目标可以形式地表示为:

$$\max J_0(C) = \prod_{k=1}^K \prod_{x \in c_k} p(k | x)$$

其中:  $p(k | x)$  是对象  $x$  相对于簇  $c_k$  的后验概率。取以 2 为底的对数,并使用贝叶斯公式,优化目标变换为:

$$\begin{aligned} \max J_1(C) &= \sum_{k=1}^K \sum_{x \in c_k} [\ln p(k) + \ln p(x | k) - \ln p(x)] \\ &\sim \sum_{k=1}^K |c_k| \ln p(k) + \sum_{k=1}^K \sum_{x \in c_k} \ln p(x | k) \\ \text{s. t. } \sum_{k=1}^K p(k) &= 1 \end{aligned} \quad (2)$$

其中:  $p(x | k)$  表示  $x$  的先验概率,  $p(k)$  为簇  $c_k$  的概率;  $p(x)$  因与  $C$  无关被忽略。

基于相关研究中普遍采用的“朴素”假设<sup>[3,14-19]</sup>来估计  $p(x | k)$ :数据集的  $D$  个属性是统计独立的。在这个假设前提下,  $p(x | k)$  可以简单地通过  $p(x_1 | k) \times p(x_2 | k) \times \dots \times p(x_d | k) \times \dots \times p(x_D | k)$  来估计。在此基础上,为区分不同属性的贡献,引入局部属性加权机制,赋予属性  $d$  以权重  $w_{kd}$  衡量其对簇  $c_k$  的重要性,数值越大表明其越重要,且满足以下约束条件:

$$\begin{cases} \forall k, d: w_{kd} > 0 \\ \forall k: \prod_{d=1}^D w_{kd} = 1 \end{cases} \quad (3)$$

注意到式(3)与相关研究采用的“权重之和归一化”条件<sup>[3,15-19]</sup>不同,式(3)基于权重之积,这有助于放大属性权重的差异:例如,若某个属性被赋予很小的权重(接近 0,表明该属性不重要),则根据式(3)必有其他一些属性的权重远大于 1。

根据上述定义,用下式估计  $p(x | k)$ :

$$p(x | k) = \prod_{d=1}^D [p(x_d | k)]^{w_{kd}}$$

其中  $p(x_d | k)$  是符号  $x_d$  经 Laplace 校正的频度估计:

$$p(x_d | k) = \frac{\#_{kd}(x_d) + 1}{|c_k| + |X_d|}$$



$\#_{kd}(x_d)$  表示符号  $x_d$  出现在簇  $c_k$  属性  $d$  上的次数。综上,算法 WBCC 的聚类优化目标可以写作:

$$\begin{aligned} \max J_2(C, P, W) = & \sum_{k=1}^K |c_k| \ln p(k) + \\ & \sum_{k=1}^K \sum_{x \in c_k} \sum_{d=1}^D w_{kd} \ln p(x_d | k) \end{aligned} \quad (4)$$

s.t. Eqs. (2) and (3)

其中:  $P = \{p(k) | k = 1, 2, \dots, K\}$ ,  $W = \{w_{kd} | k = 1, 2, \dots, K; d = 1, 2, \dots, D\}$ 。

## 2.2 聚类算法

给定对象集和  $K$ , 聚类算法需求解式(4)所示带约束的非线性优化问题。应用拉格朗日乘子法引入式(2)和(3)定义的约束条件, 算法需优化(最大化)的目标函数转换为:

$$\begin{aligned} J(C, P, W) = J_2(C, P, W) + \\ \lambda \left( 1 - \sum_{k=1}^K p(k) \right) + \sum_{k=1}^K \xi_k \left( 1 - \prod_{d=1}^D w_{kd} \right) \end{aligned}$$

其中:  $\lambda$  和  $\xi_k (k = 1, 2, \dots, K)$  为拉格朗日乘子。

算法 WBCC 基于  $K$ -means 或  $K$ -modes 的算法结构<sup>[6,9,11]</sup>, 采用一个两步骤的迭代方案求取  $J(C, W)$  的局部最优解。在第一个迭代步骤中, 将  $W$  和  $P$  视为常数, 求解令函数  $J$  取得最大值的  $C$ , 这可以通过将每个对象  $x$  重新划分到与其最相似的簇来实现。对象  $x$  与簇  $c_k$  的相似度根据下面的对数似然函数计算:

$$Sim(x, k) = \ln p(k) + \sum_{d=1}^D w_{kd} \ln p(x_d | k) \quad (5)$$

第二个迭代步骤则将  $C$  视为常数, 求取最大化  $J$  的  $W$  和  $P$ 。为此, 令  $\frac{\partial J}{\partial \lambda} = 0$  和  $\forall k: \frac{\partial J}{\partial p(k)} = 0$ , 推导得给定  $C$  情况下最优的  $p(k)$  估计式:

$$p(k) = \frac{1}{N} |c_k| \quad (6)$$

同理, 令  $\forall k, d: \frac{\partial J}{\partial w_{kd}} = 0$  和  $\forall k: \frac{\partial J}{\partial \xi_k} = 0$  得到

$$w_{kd} = \tilde{w}_{kd} \left( \prod_{d'=1}^D \tilde{w}_{kd'} \right)^{-\frac{1}{D}} \quad (7)$$

其中  $\tilde{w}_{kd}$  为给定  $C$  情况下最优的属性权重计算表达式:

$$\tilde{w}_{kd} = \left( - \sum_{x \in c_k} \ln p(x_d | k) \right)^{-1}$$

基于上述优化策略的 WBCC 算法描述如下。

### 算法 1 类属数据贝叶斯聚类算法 WBCC。

输入:  $N$  个待聚类类属型数据对象及聚类数  $K$ 。

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

Begin

生成数据集初始划分  $C$ , 并初始化  $W$  中的每个属性权重为 1; 根据式(5)计算初始的  $P$ ;

Repeat

固定  $W$  和  $P$ , 根据式(5)计算每个对象  $x$  到当前每个簇  $c_k$

似然, 并将之划分至似然最大的簇, 生成新的  $C$ ;

固定  $C$ , 根据式(6)和(7)更新  $W$  和  $P$ ;

Until  $J(C, P, W)$  的变化小于  $10^{-6}$

End

算法从一个初始的聚类划分出发, 经过一系列迭代步骤, 直到目标函数不再发生变化(实际中, 当其值的变化很小, 比如小于  $10^{-6}$  时, 判断为算法收敛)。初始划分的生成借鉴了文献[17~18]的方法, 即首先随机选择  $K$  个对象为种子, 然后

根据式(1)所示的简单匹配距离, 将所有对象划分到最近的种子, 生成数据集的初始划分。与  $K$ -means<sup>[6]</sup>、 $K$ -modes<sup>[9,11]</sup> 等算法一样, 算法 1 输出的聚类结果对其初始状态有一定的依赖性; 同时, 鉴于初始状态的随机性, 算法通常只能输出所优化目标函数的局部优解。

## 2.3 算法分析

首先分析时间复杂度。从算法过程可以看出, 算法 WBCC 与传统  $K$ -means<sup>[6]</sup> 或  $K$ -modes<sup>[9,11]</sup> 具有相似的结构, 由于其间每个迭代步骤都使得目标函数值下降或保持不变, 且对于给定的对象集和  $K$ , 目标函数存在下界, 因此, 经过有限次迭代, 可以使得函数值不再下降, 此时算法收敛。设迭代次数为  $T$ , WBCC 的算法时间复杂度为  $O(NKDT)$ 。

其次, WBCC 算法没有使用簇“中心”概念, 也不是基于类属对象间距离的聚类算法, 它根据对象-簇间的似然进行聚类划分。在聚类过程中, 算法自动赋予每个属性  $K$  个权重, 进行子空间聚类。根据式(7)可知, WBCC 算法计算属性  $d$  相对于簇  $c_k$  的权重为:

$$w_{kd} \sim \left( - \sum_{x \in X_d} \ln p(x | k) \right)^{-1} \quad (8)$$

也就是说, WBCC 的属性权重与其符号分布的信息熵成反比。这与相关研究基于模符号进行属性加权的方式<sup>[5,15,19]</sup>不同, WBCC 进行属性加权的依据是符号的总体分布。

## 3 实验与分析

实验分析包括算法聚类结果对比和属性加权方式有效性验证两个方面, 并与若干相关工作相比较。

### 3.1 实验数据与实验设置

在 6 个数据集上检验 WBCC 算法的性能, 数据集的详细信息如表 1 所示, 它们都由类属型属性组成。表 1 所列前 5 个数据集均为 UCI 数据集, 其中的 Splice 和 Promoters 是 DNA 序列集, 其每条序列由 60 或 57 个氨基酸排列而成, 各氨基酸的位点已经过对齐处理, 故每个位点可以看作是一个类属型属性, 被给予顺序编码, 比如, 在 Splice 数据中, 这些位点(属性)命名为  $p-30 \sim p+30$ <sup>[17]</sup>。其余 UCI 数据集的属性多为序类型符号, 且可能包含有缺失值。例如, 在 Breastcancer(乳腺癌) 数据中, 其 9 个属性均为以 1~10 间整数表示的患者生理指标, 其中 2 个包含缺失数据。在实验中, 所有缺失数据看作一个特别的符号加以处理。如表 1 所示, 除 Promoters 外的 UCI 数据集的另一个特点是样本分布不均衡。

表 1 实验中使用的 UCI 数据集和合成数据集

Tab. 1 The UCI data and synthetic data sets used in the experiments

数据集	属性数 $D$	类数 $K$	样本数 $N$	各类样本数
Promoters	57	2	106	53:53
Dermatology	33	6	366	20:49:52:61:72:112
Breastcancer	9	2	699	241:458
Car	6	4	1728	65:69:384:1210
Splice	60	3	3190	767:768:1655
Synthetic	40	10	10000	各 1000

为检验算法在具有更多簇类的数据上的聚类性能, 采用文献[21]提供的方法人工合成了一个包含 10 个类的数据, 如表 1“Synthetic(合成)”所示。合成过程使用的其他参数如下: 每个类的平均相关属性数为 20(占全部 40 个属性的



50%),所有属性均等宽离散化为10个序型符号。

实验选择K-modes(简称KM)<sup>[11]</sup>、CWKM<sup>[16]</sup>、MWKM<sup>[15]</sup>以及两种混合型数据聚类算法WKP<sup>[19]</sup>和MKP(Modified K-Prototypes)<sup>[22]</sup>为对比算法。MWKM和WKP的参数分别设置为 $\beta=2$ <sup>[19]</sup>和 $\beta=9$ <sup>[15]</sup>。采用两种指标评价各种算法的聚类结果:CU(Category Utility)指标和聚类精度(Clustering Accuracy, CA)。其中CU是一种评价聚类质量的内部指标,其定义<sup>[18]</sup>为:

$$CU(C) = \sum_{k=1}^K \frac{|c_k|}{N} \sum_{d=1}^D \sum_{x \in c_k} \left[ \left( \frac{\#_{kd}(x)}{|c_k|} \right)^2 - \left( \frac{\#_d(x)}{N} \right)^2 \right]$$

其中: $\#_d(x)$ 表示符号 $x$ 出现整个数据集属性 $d$ 上的次数。CA是一种外部指标,是聚类结果中对象所在簇与其真实类别相匹配的对象比例,根据下式计算:

$$CA(C) = \frac{1}{N} \sum_{k=1}^K \sum_{x \in c_k} I(k = L(x))$$

其中: $I$ 是取值0或1的指示函数, $L(x)$ 是对象 $x$ 真实的类别标号。在计算CA之前,采用二部图(bipartite graph)最大权重匹配算法建立 $K$ 个簇标号与 $K$ 个真实类别标号的对应关系,其中二部图结点对间的权重为重合对象的数目。与CU一样,CA的值越大表示聚类结果质量越高。

### 3.2 聚类结果

由于各种算法的起点(初始中心或初始划分)都是随机选择的,为使得聚类结果具有可比性,对于每个数据集,每种算法均独立运行100次,然后从中选择20次具有最高精度的

结果作为实验对比的基础。表2列出了从每个数据集的20个最好结果中计算的平均性能,以“平均指标值±1个标准差”的格式呈现,每个数据集中最高的平均指标值以加粗方式标注。

如表2所示,WBCC算法在6个数据集上均取得了较高的聚类精度(CA),与其他算法相比,WBCC在这些数据集上都取得了明显的精度提升,尤其在属性数目较多的Splice和类数较多的Dermatology及合成数据上。注意到前两个数据集中各类样本分布很不均衡(见表1),这验证了WBCC算法根据对象-簇间似然进行贝叶斯聚类的有效性。对于样本数较少的类,模符号的代表性下降,进而降低了基于模的距离度量的有效性,这是5种对比算法在这些数据集上性能落后于WBCC算法的主要原因。

根据表2,CWKM算法在Dermatology上取得最高CU指标值。与MWKM和WKP仅根据模符号的频度进行属性加权不同,CWKM算法在计算属性权重时使用了所有符号的频度信息<sup>[16]</sup>,从这个意义上说,该算法与WBCC是比较接近的,因而取得了高质量的聚类结果。但是,WBCC基于似然估计而非对象-模间相似度计算,在Splice等其他5个数据上获得了明显优于CWKM的结果。表2的数据还说明,总体而言,WBCC算法的鲁棒性(体现在标准差上)略优于对比算法,这是由于WBCC的初始状态是数据集的 $K$ 个划分,而不是对比算法的 $K$ 个代表对象,一定程度上降低了算法对初始簇中心的敏感性。

表2 各算法平均CA和CU指标对比  
Tab. 2 Comparison of average CA and CU obtained by different algorithms

数据集	WBCC 算法		KM 算法		MWKM 算法		CWKM 算法		WKP 算法		MKP 算法	
	CA	CU	CA	CU	CA	CU	CA	CU	CA	CU	CA	CU
Promoters	<b>0.794 ± 1.153 ± 0.021 ± 0.045</b>		0.722 ± 1.012 ± 0.039 ± 0.090		0.733 ± 1.075 ± 0.025 ± 0.061		0.739 ± 1.075 ± 0.025 ± 0.052		0.751 ± 1.104 ± 0.031 ± 0.061		0.708 ± 0.942 ± 0.033 ± 0.090	
Dermatology	<b>0.803 ± 4.342 ± 0.056 ± 0.389</b>		0.684 ± 4.249 ± 0.054 ± 0.279		0.736 ± 4.464 ± 0.045 ± 0.104		0.798 ± 4.608 ± 0.057 ± 0.100		0.688 ± 4.125 ± 0.092 ± 0.455		0.689 ± 4.229 ± 0.059 ± 0.300	
Breastcancer	<b>0.970 ± 1.208 ± 0.000 ± 0.000</b>		0.943 ± 1.136 ± 0.004 ± 0.009		0.927 ± 1.043 ± 0.000 ± 0.000		0.922 ± 1.026 ± 0.000 ± 0.005		0.937 ± 1.120 ± 0.000 ± 0.000		0.944 ± 1.137 ± 0.004 ± 0.012	
Car	<b>0.493 ± 0.661 ± 0.028 ± 0.026</b>		0.430 ± 0.452 ± 0.028 ± 0.037		0.386 ± 0.542 ± 0.011 ± 0.030		0.383 ± 0.532 ± 0.011 ± 0.028		0.403 ± 0.501 ± 0.019 ± 0.040		0.418 ± 0.444 ± 0.022 ± 0.020	
Splice	<b>0.918 ± 1.178 ± 0.000 ± 0.000</b>		0.433 ± 0.847 ± 0.008 ± 0.126		0.447 ± 0.904 ± 0.010 ± 0.088		0.444 ± 0.899 ± 0.009 ± 0.138		0.433 ± 0.852 ± 0.009 ± 0.123		0.433 ± 0.721 ± 0.013 ± 0.089	
Synthetic	<b>0.941 ± 4.730 ± 0.054 ± 0.254</b>		0.714 ± 3.330 ± 0.064 ± 0.263		0.774 ± 3.706 ± 0.075 ± 0.334		0.823 ± 4.015 ± 0.087 ± 0.449		0.712 ± 3.328 ± 0.077 ± 0.335		0.690 ± 3.185 ± 0.074 ± 0.317	

图1对比了各种算法的聚类效率。所用数据为表1所列的合成数据集,为检验各种算法相对于样本数量的可伸缩性,从原数据集(含10 000个样本)上随机抽取了1 250、2 500、5 000个样本组成3个新的测试数据集。如图所示,随样本量增加,WBCC算法使用的平均CPU时间呈线性增加的趋势。此外,图1也显示WBCC的聚类效率介于MWKM、WKP和CWKM、KM、MKP之间;MKP算法未进行属性加权操作,具有最高的效率;而MWKM算法为每个属性计算多个权重,因而需要更多的运行时间。

### 3.3 属性加权结果

本节从属性加权方法的角度进一步分析WBCC算法的性能,并于CWKM、MWKM和WKP这三种同样基于属性加权的聚类算法作对比。图2~5显示了四种算法在Splice数据

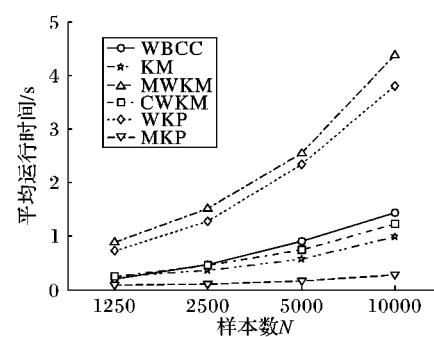


图1 合成数据上各种算法的平均运行时间对比  
Fig. 1 Comparison of average running time (in seconds) of different algorithms on synthetic data set



上计算的权重分布情况,权重数据采自各算法在 100 次运行中精度最高的聚类结果。为便于对比,图 2~5 中各算法生成的属性权值均规范化到区间 [0,1]。选用 Splice 的一个原因是该数据集拥有较多的属性(60 个),便于分析;另一个原因是该数据具有明确的生物学背景<sup>[17]</sup>,易于理解。Splice 数据包含三个类别:EI、IE 和 Neither,前 2 个类别在 p-2~p+2 氨基酸位点(类属型属性)上含有“移植体(donor)”或“受体(acceptor)”,而它们未出现在 Neither 类别的 DNA 序列上。

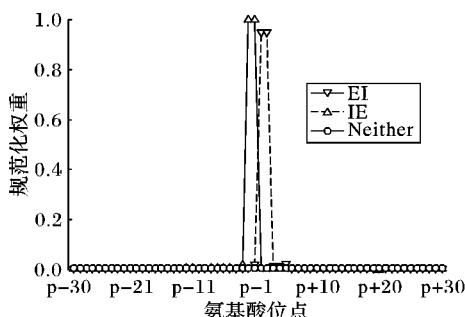


图 2 Splice 数据上 WBCC 算法计算的属性权重分布  
Fig. 2 Weight distributions yielded by WBCC on Splice

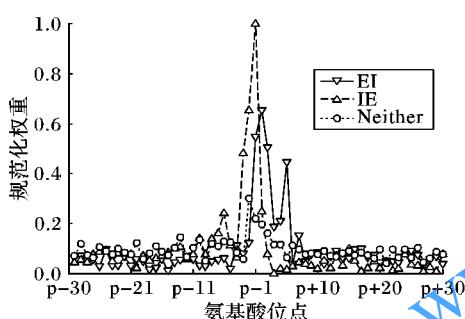


图 3 Splice 数据上 CWKM 算法计算的属性权重分布  
Fig. 3 Weight distributions yielded by CWKM on Splice

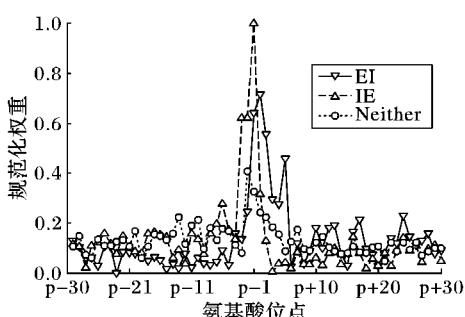


图 4 Splice 数据上 MWKM 算法计算的属性权重分布  
Fig. 4 Weight distributions yielded by MWKM on Splice

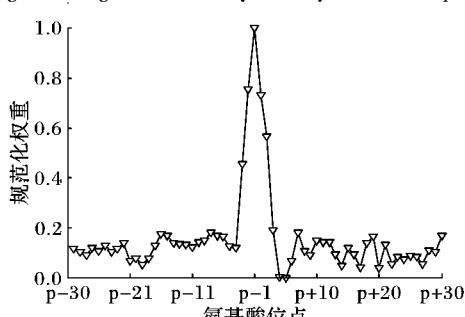


图 5 Splice 数据上 WKP 算法计算的属性权重分布  
Fig. 5 Weight distribution yielded by WKP on Splice

如图 2 所示,WBCC 算法成功地识别出了 p-2~p+2 位点(属性),对于 EI 和 IE 类,这些属性被赋予较高的权重,而其他属性的权重接近于 0;对于 Neither 类,各属性的权重都接近于 0,并没有明显区别。这些都与上述 Splice 数据的生物学应用背景相符。

反观对应于 CWKM 算法和 MWKM 算法的图 3 和图 4(MWKM 赋予属性两种权重,图上仅显示其中“与模频度成正比”的部分),其权重分布与 WBCC 的图 2 有显著区别。如图 3~4 所示,尽管在 p-2~p+2 位点的权重分布也呈现出了峰值,但其他属性也被赋予了较大的权重,尤其对于 Neither 类,其权重分布并不平滑。同样情况见于 WKP 算法输出的图 5(WKP 是全局属性加权算法<sup>[19]</sup>,因此只输出一组权重)。上述有差异的属性加权结果是算法不同加权方法和所采用的相似度(或距离)度量的体现。在 WBCC 中,属性权重根据符号分布的信息熵计算(见式(8)),且基于对象-簇似然进行高质量的数据集划分,因而可以获得具有实际应用意义的属性加权结果。

#### 4 结语

本文提出了一种新型的类属型数据聚类算法 WBCC,与当前多基于模的划分聚类算法不同,新算法基于贝叶斯概率框架,通过最大似然估计,而不是现有多数算法所采用的对象-模间简单符号匹配,进行数据集划分。在聚类过程中,WBCC 算法根据类属符号分布的信息熵自动赋予每个属性反映其重要性的权重,实现了类属型数据的子空间聚类。在多个实际应用数据集上进行了实验验证,结果表明新算法是有效的,与基于模和类属对象间距离的现有算法相比,新算法在实验数据上的聚类结果质量得到较为明显的改善,并输出了与实际应用需要相吻合的属性加权结果。

后续研究工作将着重于以下几个方面:将新算法推广到混合型数据,即在混合了数值型和类属型的数据上直接(不需要将数值型数据事先离散化成类属型)进行贝叶斯聚类;探讨建立数据集初始划分的方法,以提高算法的鲁棒性。鉴于当前的聚类模型评价准则多针对数值型数据且仅对全空间聚类结构进行质量评价,后续工作还将在本文给出的类属数据子空间聚类概率模型和传统的贝叶斯信息准则基础上,开展类属数据子空间聚类有效性指标的研究,提供类属数据集最佳聚类数目估计等问题的解决方案。

#### 参考文献(References)

- HUNT L, JORGENSEN M. Clustering mixed data[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, 1(4): 352~361.
- BORIAH S, CHANDOLA V, KUMAR V. Similarity measures for categorical data: a comparative evaluation[C]// Proceedings of the 8th SIAM International Conference on Data Mining. Philadelphia: SIAM, 2008: 243~254.
- CHEN L, WANG S. Central clustering of categorical data with automated feature weighting[C]// Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2013: 1260~1266.
- GUHA S, RASTOGI R, SHIM K. ROCK: a robust clustering algorithm for categorical attributes[J]. Information Systems, 2000, 25(5): 345~366.



- [5] XIONG T, WANG S, MAYER A, et al. DHCC: divisive hierarchical clustering of categorical data [J]. *Data Mining and Knowledge Discovery*, 2012, 24(1): 103–135.
- [6] MACQUEEN J. Some methods for classification and analysis of multivariate observation [C]// Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1967: 281–297.
- [7] JI J, BAI T, ZHOU C, et al. An improved  $k$ -prototypes clustering algorithm for mixed numeric and categorical data [J]. *Neurocomputing*, 2013, 120: 590–596.
- [8] SAN O, HUYNH V, NAKAMORI Y. An alternative extension of the  $k$ -means algorithm for clustering categorical data [J]. *International Journal of Applied Mathematics and Computer Science*, 2004, 14(2): 241–247.
- [9] HUANG Z, NG M. A note on  $k$ -modes clustering [J]. *Journal of Classification*, 2003, 20(2): 257–261.
- [10] 李仁侃, 叶东毅. 粗糙  $k$ -modes 聚类算法 [J]. *计算机应用*, 2011, 31(1): 97–100. (LI R K, YE D Y. Rough  $k$ -modes clustering algorithm [J]. *Journal of Computer Applications*, 2011, 31(1): 97–100.)
- [11] HUANG Z. Extensions to the  $k$ -means algorithm for clustering large data sets with categorical values [J]. *Data Mining and Knowledge Discovery*, 1998, 2(3): 283–304.
- [12] 梁吉业, 白亮, 曹付元. 基于新的距离度量的  $k$ -modes 聚类算法 [J]. *计算机研究与发展*, 2010, 47(10): 1749–1755. (LIANG J Y, BAI L, CAO F Y.  $k$ -modes clustering algorithm based on a new distance measure [J]. *Journal of Computer Research and Development*, 2010, 47(10): 1749–1755.)
- [13] LIN D. An information-theoretic definition of similarity [C]// Proceedings of the 15th International Conference on Machine Learning. San Francisco: Morgan Kaufmann, 1998: 296–304.
- [14] GOODALL D. A new similarity index based on probability [J]. *Biometrics*, 1966, 22(4): 882–907.
- [15] BAI L, LIANG J, DANG C, et al. A novel attribute weighting al-
- gorithm for clustering high-dimensional categorical data [J]. *Pattern Recognition*, 2011, 44(12): 2843–2861.
- [16] CAO F, LIANG J, LI D, et al. A weighting  $k$ -modes algorithm for subspace clustering of categorical data [J]. *Neurocomputing*, 2013, 108: 23–30.
- [17] CHEN L, WANG S, WANG K, et al. Soft subspace clustering of categorical data with probabilistic distance [J]. *Pattern Recognition*, 2016, 51: 322–332.
- [18] 陈黎飞, 郭躬德. 属性加权的类属型数据非模聚类 [J]. *软件学报*, 2013, 24(11): 2628–2641. (CHEN L F, GUO G D. Non-mode clustering of categorical data with attributes weighting [J]. *Journal of Software*, 2013, 24(11): 2628–2641.)
- [19] HUANG Z, NG M, RONG H, et al. Automated variable weighting in  $k$ -means type clustering [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(5): 657–668.
- [20] BOUGUESSA M. Clustering categorical data in projected spaces [J]. *Data Mining and Knowledge Discovery*, 2015, 29(1): 3–38.
- [21] CHEN L. A probabilistic framework for optimizing projected clusters with categorical attributes [J]. *Science China Information Sciences*, 2015, 58(7): 072104(15).
- [22] LIANG J, ZHAO X, LI D, et al. Determining the number of clusters using information entropy for mixed data [J]. *Pattern Recognition*, 2012, 45(6): 2251–2265.

This work is partially supported by the National Natural Science Foundation of China (61175123), the Natural Science Foundation of Fujian Province (2015J01238).

**ZHU Jie**, born in 1971, senior engineer. His research interests include pattern recognition, target identification.

**CHEN Lifei**, born in 1972, Ph. D., professor. His research interests include statistical machine learning, data mining, pattern recognition.

(上接第 1025 页)

- [6] CHENG W, PANG H H, TAN K. Authenticating multi-dimensional query results in data publishing [C]// DBSEC 2006: Proceedings of the 20th IFIP WG 11.3 Working Conference on Data and Applications Security. Berlin: Springer, 2006: 60–73.
- [7] LI F, HADJIELEFTHERIOU M, KOLLIOS G, et al. Dynamic authenticated index structures for outsourced databases [C]// SIGMOD 2006: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2006: 121–132.
- [8] YANG Y, PAPADOPOULOS S, PAPADIAS D, et al. Authenticated indexing for outsourced spatial databases [J]. *The VLDB Journal*, 2009, 18(3): 631–648.
- [9] YANG Y, PAPADOPOULOS S, PAPADIAS D, et al. Spatial outsourcing for location-based services [C]// ICDE 2008: Proceedings of the 2008 IEEE 24th International Conference on Data Engineering. Washington, DC: IEEE Computer Society, 2008: 1082–1091.
- [10] MOURATIDIS K, SACHARIDIS D, PANG H H. Partially materialized digest scheme: an efficient verification method for outsourced databases [J]. *The VLDB Journal*, 2009, 18(1): 363–381.

- [11] 谢晴晴, 王良民. 基于 PMD 的外包数据流范围查询验证方案 [J]. *计算机科学与探索*, 2015, 9(10): 1209–1218. (XIE Q Q, WANG L M. Data stream range query authentication scheme based on PMD in outsourced database [J]. *Journal of Frontiers of Computer Science and Technology*, 2015, 9(10): 1209–1218.)
- [12] HU H, XU J, CHEN G, et al. Authenticating location-based services without compromising location privacy [C]// SIGMOD 2012: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2012: 301–312.

This work is partially supported by the National Natural Science Foundation of China (61462056, 71363040), the Natural Science Foundation of Inner Mongolia (2016MS0609).

**HU Xiaoyan**, born in 1980, M. S., lecturer. Her research interests include cloud computing, database management.

**WANG Jingyu**, born in 1976, Ph. D., associate professor. His research interests include cloud computing, information security.

**LI Hairong**, born in 1976, M. S., associate professor. Her research interests include cloud storage.