Journal of System Simulation

Volume 34 | Issue 10 Article 10

10-18-2022

A K-modes Clustering Method Based on Maximal Information Coefficient Data Preprocessing

Mingmei Li
1.Hangzhou Dianzi University, Hangzhou 310018, China;, 851628184@qq.com

Chenglin Wen
1.Hangzhou Dianzi University, Hangzhou 310018, China;2.Guangdong Institute of Petrochemical Technology, Maoming 525000, China;, wencl@hdu.edu.cn

Shaolin Hu
2.Guangdong Institute of Petrochemical Technology, Maoming 525000, China;

A K-modes Clustering Method Based on Maximal Information Coefficient Data Preprocessing

Abstract

Abstract: The existing k-modes clustering method ignores the weak correlation of variable attributes, which often results in poor clustering performance in practical applications. *A new k-modes clustering method that includes the weak correlation of attributes is proposed. Maximum information coefficient (MIC) is introduced to measure the correlation of variable attributes in the data set. The obtained MIC value is merged with the original distance to establish a new measurement method containing weak attribute correlation information to enhance the completeness of related information of variable attributes, and a more refined k-modes clustering method is established.* Three different data sets are used to compare the performance of the new method with the existing k-modes clustering and other improved k-modes clustering methods, the simulation results shows the effectness of the new method.

Keywords

clustering algorithm, k-modes, maximum information coefficient(MIC), distance metric, variable attribute

Recommended Citation

Mingmei Li, Chenglin Wen, Shaolin Hu. A K-modes Clustering Method Based on Maximal Information Coefficient Data Preprocessing[J]. Journal of System Simulation, 2022, 34(10): 2204-2212.

第 34 卷第 10 期 系统仿真学报© Vol. 34 No. 10
2022 年 10 月 Journal of System Simulation Oct. 2022

一种基于最大信息系数预处理的k-modes聚类方法

李明媚[1]，文成林[1,2*]，胡绍林[2]
(1. 杭州电子科技大学，浙江 杭州 310018；2. 广东石油化工学院，广东 茂名 525000)

摘要：为解决现有k-modes聚类方法因忽略了变量属性之间的弱相关性，常造成其在实际应用中聚类性能不佳的问题，*提出一种包含属性弱相关性的新k-modes聚类方法。引入最大信息系数(maximum information coefficient，MIC)度量数据集中变量属性之间的相关性；将得到的MIC值与原有距离进行融合，建立包含属性弱相关性信息的新度量方法，以增强变量属性间相关信息的完备性，建立更加精细的k-modes聚类方法；*调用3种不同的数据集，将新方法与原有的k-modes聚类方法和其他改进k-modes聚类方法的性能进行对比，并通过仿真结果表明了新方法的有效性。

关键词：聚类方法；k-modes；最大信息系数；距离度量；变量属性
中图分类号：TP311 文献标志码：A 文章编号：1004-731X(2022)10-2204-09
DOI: 10.16182/j.issn1004731x.joss.21-0484

A K-modes Clustering Method Based on Maximal Information Coefficient Data Preprocessing

Li Mingmei[1], *Wen Chenglin*[1,2*], *Hu Shaolin*[2]
(1. Hangzhou Dianzi University, Hangzhou 310018, China; 2. Guangdong Institute of Petrochemical Technology, Maoming 525000, China)

Abstract: The existing k-modes clustering method ignores the weak correlation of variable attributes, which often results in poor clustering performance in practical applications. *A new k-modes clustering method that includes the weak correlation of attributes is proposed. Maximum information coefficient (MIC) is introduced to measure the correlation of variable attributes in the data set. The obtained MIC value is merged with the original distance to establish a new measurement method containing weak attribute correlation information to enhance the completeness of related information of variable attributes, and a more refined k-modes clustering method is established.* Three different data sets are used to compare the performance of the new method with the existing k-modes clustering and other improved k-modes clustering methods, the simulation results shows the effectness of the new method.

Keywords: clustering algorithm; k-modes; maximum information coefficient(MIC); distance metric; variable attribute

引言

由于互联网的急速发展，数据量也随之激增，给数据挖掘带来更大的发展机遇。其中聚类分析是数据挖掘的重要功能之一，它是一种无监督的学习方法。市场营销、生物基因学、医学、人工智能等领域都是聚类分析的应用场景。早期的聚类方法都是在单一属性下进行聚类的，因此泛化能力弱，很难表现出属性基本特征。例如，k-means算法，基于数值属性间的欧氏距离度量进行聚类，不能准确地表达变量间的数据结构关系，

收稿日期：2021-05-26 修回日期：2021-08-06
基金项目：国家自然科学基金(61933013)
第一作者：李明媚(1997-)，女，硕士生，研究方向为数据挖掘。E-mail：851628184@qq.com
通讯作者：文成林(1963-)，男，博士，教授，研究方向为信息融合与故障诊断、深度学习。E-mail：wencl@hdu.edu.cn

并忽视了分类变量属性中的信息，在分类属型数据上聚类效果不好，因此如何对分类属型数据集进行更好的聚类已成为许多研究人员的重要课题。Huang 等[1]提出了 k-modes 算法，该算法是一种针对分类属型数据集的算法，是 k-means 聚类算法的延伸。该算法的优势是利用简单的 0-1 差异匹配法解决了分类属性变量之间的距离度量问题，可以有效处理大、小型数据集。但是 0-1 差异匹配法对变量之间的相关性进行二值分类，割裂了属性之间的基本联系，会削弱类内的相似性，不能很好地反映变量之间的最基本特征和更广泛的特征。

基于对 k-modes 聚类算法的优缺点分析，许多学者从 2 个方向对算法进行了改进。一种是基于众数作为类中心的思想，并在此基础上进行扩展的改进算法。另一种是基于对 k-modes 聚类算法中差异性度量函数的改进，通过改进距离度量的方法，得到更好的聚类结果。大部分学者都是基于第二种方法对 k-modes 聚类算法进行改进，通过数据预处理方法进行变量属性特征间相关性的度量，进而改进 k-modes 聚类中的距离度量方法。Ahmad 等[2]通过变量属性间的共现程度体现变量之间的距离远近；张小宇等[3]通过图形聚类理论，利用连接度方法衡量变量属性之间的相似度；吴润秀[4]通过样本间的互信息，描述变量属性间相互关系。虽然上述方法在一定程度上改善了 k-modes 聚类算法原有距离度量的影响，但是 Ahmad 等所定义的距离公式只考虑到同一个变量下不同属性值的差异；张小宇等所定义的基于图形聚类连接度的距离公式计算复杂度较高；吴润秀等所提出的基于互信息的距离公式，同样也面临着计算复杂度较高、互信息联合概率难求的问题。

本文引入最大信息系数(maximum information coefficient, MIC)，它是一种用于检测变量之间非线性相关性的新方法。最大信息系数 MIC 可以用来体现变量属性特征之间的相互关系，该方法的优势是既保持了强相关性变量属性信息，又补充了弱相关性变量属性间的关系，以 MIC 值作为相关性强弱的衡量指标更具有代表性，MIC 值越大代表相关性越强反之越弱。

1 数据预处理方法

在聚类分析中，数据预处理可以有效改善数据质量，进而提高聚类分析的准确度[5]。好的预处理方法在聚类过程中可以达到事半功倍的效果，因此数据预处理也是聚类分析中的重要一环。

常见的数据预处理方法主要有数据清洗、数据集成、数据变换和数据归约等。数据清洗主要解决数据错误、缺失、冗杂等问题[6]；数据集成主要解决多数据源统一数据形式的问题；数据变换主要解决数据表现形式转换的问题；数据归约主要指的是在保证原始数据代表性的前提下，尽可能地减少数据量。

本文所述数据预处理方法的主要目的是选取变量之间具有相关性关系的属性特征，方便在聚类过程中更好地度量变量之间的距离。

设 $X=\{x_1,x_2,\cdots,x_n\}$ 是 n 个变量组成的数据集，每一个变量有 m 个分类属性，其中变量 $x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ 由 m 个属性组成，$x_{ij}\in A$，$A=\{A_1,A_2,\cdots,A_m\}$ 是属性集合。计算变量 x_i 和 x_j 之间的相关度。

常见预处理方法有：

(1) 0-1 差异度量

0-1 差异度量指的是比较变量间属性特征的差异度，当两变量间某一属性相同，则记录为 0，反之为 1。则 2 个变量属性 x_{il} 和 x_{jl} 之间的差异性可以定义为

$$\delta(x_{il},x_{jl})=\begin{cases}0, x_{il}=x_{jl}\\ 1, x_{il}\neq x_{jl}\end{cases} \tag{1}$$

通过计算变量间所有属性特征的差异度之和来代表变量间的相关性，变量间差异度越小代表相关性越强，差异度越大代表相关性越弱。该方法的优点是计算简单，适用于很多领域；缺点是对变量之间的相关性二值分类，割裂了属性之间的基本联系，不能很好地反映对象变量之间的最

DOI: 10.16182/j.issn1004731x.joss.21-0484

基本的特征。

(2) 皮尔森相关系数

皮尔森相关系数[7]是一种线性相关系数，记为r，取值范围为[−1, 1]，可以体现变量之间线性相关性的强弱：

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{X})(x_j-\overline{X})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(x_j-\overline{X})^2}} \tag{2}$$

当$r\to 1$，表示变量间为正相关，r值越接近1正相关性越强，$r=1$时表示变量间完全正相关；当$r\to -1$，表示变量间为负相关，r值越接近−1负相关性越强，$r=-1$时表示变量间完全负相关；当$r\to 0$，表示变量间线性关系很弱，$r=0$时表示两变量独立不相关。该方法的优点是只计算r值便可代表相关度强弱，缺点为该系数只对变量间的线性函数敏感，对非线性函数不敏感。

(3) 互信息

互信息[8-10](mutual information, MI)是信息理论中有用的信息衡量标准，用于度量变量之间的相关性。MI值的大小是两变量之间依赖程度的直接体现。两变量之间的互信息可以定义为

$$I(x_i,x_j)=\sum_{x_i\in X}\sum_{x_j\in X}p(x_i,\mathrm{x}_j)\,\mathrm{lb}\left(\frac{p(x_i,\mathrm{x}_j)}{p(x_i)p(\mathrm{x}_j)}\right) \tag{3}$$

式中：$p(x_i,\mathrm{x}_j)$为x_i和x_j之间的联合概率；$p(x_i)$、$p(x_j)$分别为x_i，x_j的边缘概率。由MI的定义可以得出：当$I(x_i,x_j)=0$时，表示两变量之间无关，相互独立；当$I(x_i,x_j)\to 0$时，表示两变量之间相关性弱；当$I(x_i,x_j)\to 1$时，表示两变量之间相关性强。MI方法的优点是通过MI值的大小可以表示变量之间的相关程度。缺点是MI不属于度量方式，无法进行归一化处理，因此无法比较不同的数据集，而且MI对连续变量值不敏感。

2 基于最大信息系数的数据预处理方法

现有的k-modes聚类方法中，距离公式的度量是基于0-1差异匹配方法计算的，变量属性特征之间的相关性度量非0即1，但在实际应用中，变量属性特征间的相关性是有强弱之分的，如何更好地度量变量属性特征之间的相关性在聚类过程中是十分重要的。因此，本文提出一种基于MIC预处理的k-modes聚类方法，相比于传统的数据预处理方法，以MIC值作为相关性强弱的衡量指标更具有代表性。

2.1 最大信息系数理论

David N.Reshef[11]提出了MIC，它可以检测可变对象之间的潜在关系。MIC是衡量变量之间相互依存关系的一个很好的测度，它具有广泛性和公平性2个重要属性。MIC的广泛性是指它在多样本情况下对于多种函数关系都敏感，可以检测出多种关系类型。例如，非函数关系和多种函数关系合成的超函数关系等。MIC的公平性是指当在不同的关系类型中加入相同的噪声时，它们之间的MIC值是相近的；反之，当计算出2个变量MIC值相似或者相等时，对于加入的噪声程度的值也相近。

MIC方法的核心理念是：对于2个变量，如果2个变量之间有一定的相关性，那么可以在二维平面中通过绘制网格栅(x,y)将两变量组成的散点图的数据点分割开，再通过穷举所有的网格栅的方法计算MIC。该方法基于MI的基础提出，不仅有着MI所有的优良特性，同时还克服了MI的相关缺陷，它寻找出了一种最优的离散化方式，可以把MI取值转换成一种度量方式，以MIC值直接代表相关性强弱，其取值范围为[0, 1]。

2.2 MIC预处理过程

给定有限有序的数据集$X=\{x_1,x_2,\cdots,x_n\}$，如果将x轴划分为x个格子，y轴划分为y个格子，那么就可以得到一个$x\times y$的网格划分G，其中x，y是正整数，将落入G点的数量占X数量的比例看作是其概率密度$X|G$，而根据不同的网格划分情况得

到的概率分布$X|G$也不同。在$X=\{x_1,x_2,\cdots,x_n\}$中，两变量x_i和x_j之间的互信息可以定义为

$$I(x_i,x_j)=\sum_{x_i\in X}\sum_{x_j\in X}p(x_i,\mathrm{x}_j)\mathrm{lb}\left(\frac{p(x_i,\mathrm{x}_j)}{p(x_i)p(\mathrm{x}_j)}\right) \tag{4}$$

在x，y给定的情况下，若改变x，y的值，得到的MI值也会发生变化，记录其中最大的MI值为$I(X,x_i,x_j)$。然后执行归一化以比较不同维数下的数据集，并且归一化后的值在$[0,1]$之间。通过更改x，y的值，可以获得变量之间归一化后的MI值特征矩阵。特征矩阵的最大值是2个变量之间的最大信息系数MIC值。

$$M(X)_{x_i,x_j}=\frac{I(X,x_i,x_j)}{\mathrm{lb}(\min\{x_i,x_j\})} \tag{5}$$

$X=\{x_1,x_2,\cdots,x_n\}$样本容量取值为n，网格化的分数取值小于$B(n)$。则最大信息系数可以定义为

$$MIC(X)=\max_{xy<B(n)}\{M(X)_{x,y}\} \tag{6}$$

式中：x,y是在x轴y轴方向上的划分格子的个数，也就是网格分布；$B(n)$是一个变量，其大小一般为数据n的0.6次方左右，即$B(n)\approx n^{0.6}$。

3 k-modes算法

在数据挖掘中，k-modes算法被认为是一种适用于进行分类属型数据聚类处理的聚类算法，可以用来处理例如花朵种类、颜色、形状等属性。k-modes算法是一种基于k-means算法的拓扑化延伸[12]。与k-means聚类算法不同的地方在于k-modes聚类采用0-1匹配差异性度量变量间的距离，并且使用modes更新类中心。在该距离度量方式下，变量和聚类中心之间的差异度即为它们之间不同属性的个数。相同的记录值为0，不同的记录值为1，最后计算它们之间的总和[13]。该总和即为变量到聚类中心的差异度。在k-modes算法中，差异度越小代表2个变量之间的距离越小。

设$X=\{x_1,x_2,\cdots,x_n\}$是n个变量组成的数据集，每一个变量有m个分类属性，其中变量$x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$由$m$个属性组成，$x_{ij}\in A$，$A=\{A_1,A_2,\cdots,A_m\}$是属性集合。则2个变量$x_i$，$x_j$的差异性被度量为

$$d(x_i,x_j)=\sum_{l=1}^{m}\delta(x_{il},x_{jl}) \tag{7}$$

其中，

$$\delta(x_{ij},x_{jr})=\begin{cases}0, & x_{ij}=x_{jr}\\ 1, & x_{ij}\neq x_{jr}\end{cases} \tag{8}$$

若将X中的n个变量分为k个类，第i个类的类中心为q_i，$i=1,2,\cdots,k$，第i个变量的数据属于第r类的距离度量$d(x_i,q_r)$，那么则定义所有变量到所有类中心的目标函数为

$$P(W,Q)=\sum_{r}^{k}\sum_{i}^{n}w_{ir}d(x_i,q_r) \tag{9}$$

式中：$q_r=\{q_{r1},q_{r2},\cdots,q_{rm}\}$，$q_{rm}$是第$r$个类的第$m$个属性的众数。并有：

$$\sum_{r=1}^{k}w_{ir}=1,\ \sum_{r=1}^{k}w_{ir}\in\{0,1\},0<\sum_{i}^{n}w_{ir}<n \tag{10}$$

式中：$1\leqslant i\leqslant n,1\leqslant r\leqslant k$。

k-modes算法流程如下：

step 1：给定含有n个变量的数据集X和聚类数k；

step 2：随机选择k个变量作为初始类中心，每一个类中心成为一类；

step 3：利用式(7)计算所有变量与上述k个初始类中心的距离度量$d(x_i,q_r)$，之后将每个变量分配到与其距离度量最小的那个类当中，得到新的k个类；

step 4：重新计算每个类的众数$q_{r1},q_{r2},\cdots,q_{rm}$，即在每个类别的各种属性中，选择一个类别中比例最大的一种属性值来作为此类中心的一个属性值，从而获得一个新的聚类中心q_i，$i=1,2,\cdots,k$；

step 5：返回到step 3，直到每个类的类中心不再发生改变为止。

4 基于MIC预处理的k-modes聚类方法步骤

定义$S=(X,A)$为一个信息系统，数据集X为

DOI: 10.16182/j.issn1004731x.joss.21-0484

n 个变量构成的非空有限集合 $X=\{x_1,x_2,\cdots,x_n\}$，属性集 A 由 m 个分类属型属性构成的非空有限集合 $A=\{A_1,A_2,\cdots,A_m\}$。其中变量 x_i 的 m 个变量属性可以表述为 $x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$。

首先通过MIC方法计算出属性集 A 中各变量属性特征之间的最大信息系数：

$$\sum_{r=1}^{m} MIC(A_r,A_l) \tag{11}$$

$$MIC(A_r,A_l)=\max_{xy<B(n)}\frac{I(A_r,A_l)}{\mathrm{lb}\{\min(x,y)\}} \tag{12}$$

在得到变量属性间的 MIC 值后，可以重新定义变量 x_i 和 x_j 的第 l 个属性特征之间的距离度量：

$$d(x_{il},x_{jl})=\begin{cases}0, x_{il}=x_{jl}\\ 1-\dfrac{1}{m}\displaystyle\sum_{r=1}^{m} MIC(A_r,A_l)f(x_{ir},x_{jr}), x_{il}\neq x_{jl}\end{cases} \tag{13}$$

函数 $f(x_{ir},x_{jr})$ 定义为

$$f(x_{ir},x_{jr})=\begin{cases}0, x_{ir}=x_{jr}\\ 1, x_{ir}\neq x_{jr}\end{cases} \tag{14}$$

则变量 x_i 和 x_j 的距离度量为

$$d(x_i,x_j)=\frac{1}{m}\sum_{l=1}^{m} d(x_{il},x_{jl})\in(0,1) \tag{15}$$

基于MIC预处理的k-modes聚类方法具体流程如下：

step 1：给定含有 n 个变量的数据集 X 和聚类数 k；

step 2：计算各变量属性之间的 MIC 值；

step 3：任意给出 X 的 k 个非空子集 $X_1,X_2,\cdots,X_k$，$(X_1\cup X_2\cup\cdots\cup X_k\in X)$ 作为初始的聚类划分；

step 4：计算当前 k 个非空子集 $X_1,X_2,\cdots,X_k$ 的类中心 $q_1,q_2,\cdots,q_k$，即在每个 $X_i(1\leqslant i\leqslant k)$ 类中，对每一个属性分别选取其在该类中占有最大比例的属性值，作为该类的类中心的属性值，从而得到每个类的类中心 $q_1,q_2,\cdots,q_k$；

step 5：对每一个变量 $x_i(1\leqslant i\leqslant n)$，将step 2中计算得到的 MIC 值代入式(13)中，可以得出第 i 个变量到第 r 类类中心的距离 $d(x_i,q_r)$，然后将变量 $x_i(1\leqslant i\leqslant n)$ 分配到与之距离最小的类中心所在的类，进而得到新划分后的 k 个非空子集 $X_{1'},X_{2'},\cdots,X_{k'}$；

step 6：返回到step 3，直到每个变量所属的类不再改变。

通过分析可知，底层算法中计算各变量属性之间的 MIC 值需要 $O(c^2x^3y)$ 步，计算变量之间距离需要 $O(mknt)$ 步，因此本文提出的方法需要 $O(c^2x^3y+dknt)$ 步。其中，c 为常数，取默认值15，x,y 为网格分布划分格子的个数，$xy\leqslant B(n)=n^{\alpha}, x\in[2,B(n)/2]$，$n$ 为数据集中变量个数，m 为变量属性的个数，k 为聚类数，t 为最大迭代次数。由于在数据集中 $n\gg m,k,t$，因此当 α 取默认值0.6时，MIC-k-modes方法的时间复杂度可近似为 $O(n^{2.4})$。

5 实验分析

为了验证基于MIC预处理的k-modes聚类方法效果，本文通过分类正确率(AC)、分类精度(PE)、召回率(RE) 3个指标对仿真结果进行评定：

$$AC=\frac{\sum_{i=1}^{k}a_i}{n},PE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+b_i}}{k},RE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+c_i}}{k} \tag{16}$$

式中：n 为数据库中的第 i 类聚类对象数；a_i 为第 i 类中正确的聚类对象数；b_i 为第 i 类中错误的聚类对象数；c_i 为应该分到第 i 类却被分到其他类的聚类对象数；k 为聚类类别的个数。AC 表示正确地分到一个类中的对象所占全部对象的比例。AC 值越大，代表分类划分的正确率就越高。PE 表示将每个类中正确分类的对象数与错误分类的对象数进行计算，得到 k 个类的平均分类精度。PE 值越大，代表错误的分类对象数量越少。RE 表示将每个类中正确的分类对象数与应该分到却没有分到的对象数进行计算，得到 k 个类中正确分类对象的平均数。RE 值越大，代

表应该分到却没分到类中的对象数越少。

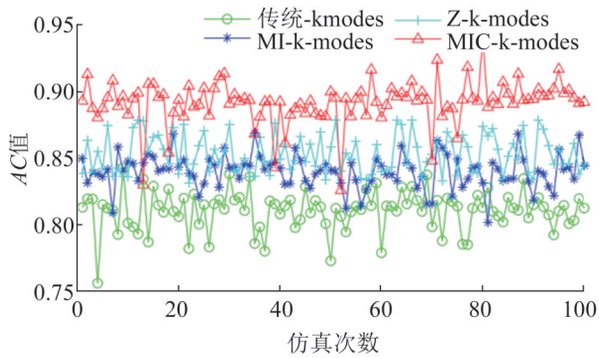
这里使用UCI数据集中的vote、cancer和mushroom数据集进行方法验证。表1为3个数据集的简单说明。

表1 数据集描述
Table 1 Dataset description

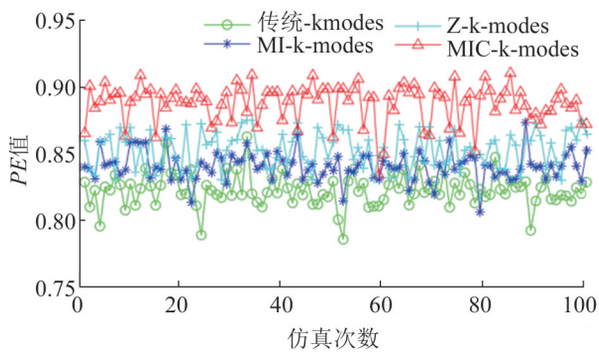
数据集	对象数n	属性数m
vote	435	16
cancer	699	10
mushroom	8 124	22

本文分别运用k-modes、基于MI的k-modes(MI-k-modes)、基于加权连接度的k-modes(Z-k-modes)和基于MIC预处理的k-modes(MIC-k-modes)4种方法对表1中的数据集进行聚类仿真。由于聚类算法初始类中心的选择会对算法造成影响，即使是相同的聚类算法，当选择不同的初始类中心时，也可能会得到不同的聚类算法结果。因此为了使仿真结果更加真实有效，本文对每种聚类算法随机选取了100次初始类中心，并在相同的初始聚类中心下分别运行4种方法，通过取100次的平均聚类结果的*AC*、*PE*、*RE*值来验证方法的可行性和有效性。4种方法在数据集上运行100次所得的聚类评价标准如图1~3所示。为了更加直观地分析方法的有效性，算出这100次聚类结果的平均数，可以得出表2~4的结果，表5~7为MIC-k-modes对比其他方法的提高率。

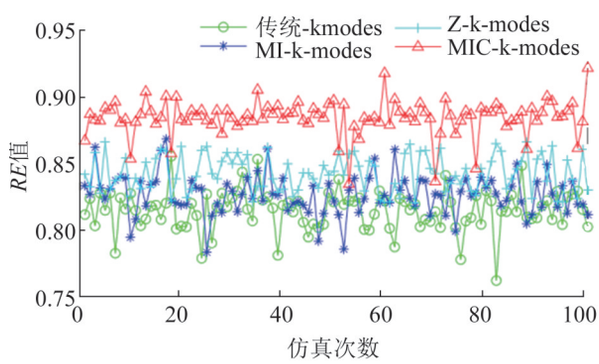
通过观察图1~3可以发现，由于初始聚类中心的选择不同，4种方法在vote、cancer和mushroom数据集的聚类结果都有一定程度的波动，但是基于MIC预处理的k-modes聚类方法的AC、PE、RE曲线整体趋势都在传统k-modes与MI-k-modes曲线之上，验证了该方法的可行性与有效性；通过表5~7的聚类结果形成表8，可以发现随着数据集变量以及变量属性个数的增加，该方法的提高率整体呈上升趋势，表现更加优异，可以得到更好的聚类精度。



(a) *AC*值分布图

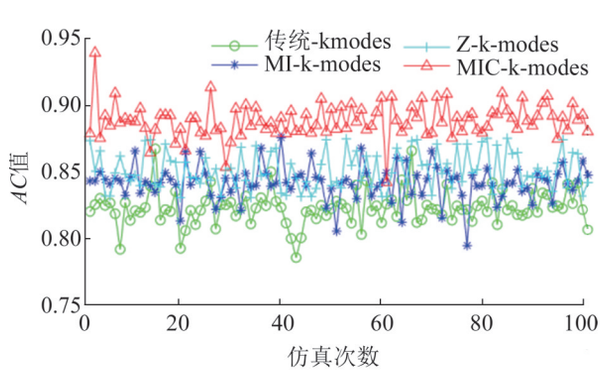


(b) *PE*值分布图



(c) *RE*值分布图

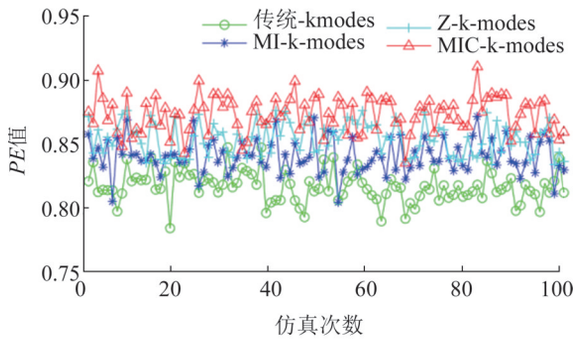
图1 100次评价标准在vote数据集上的分布情况
Fig. 1 Distribution of 100 rating criteria over vote data set



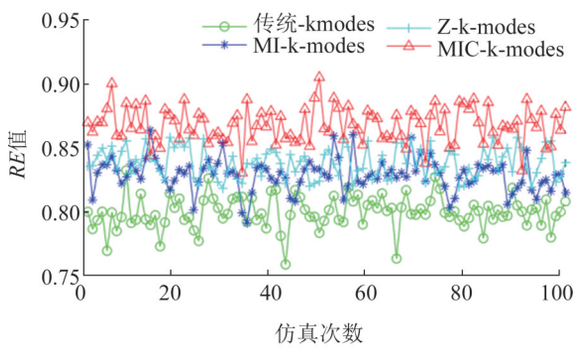
(a) *AC*值分布图

DOI: 10.16182/j.issn1004731x.joss.21-0484

第 34 卷第 10 期 系统仿真学报 Vol. 34 No. 10
2022 年 10 月 Journal of System Simulation Oct. 2022



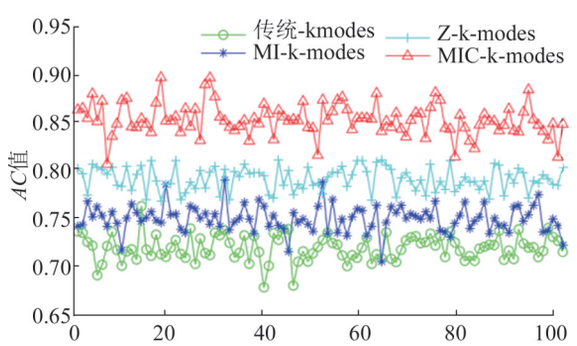
(b) *PE*值分布图



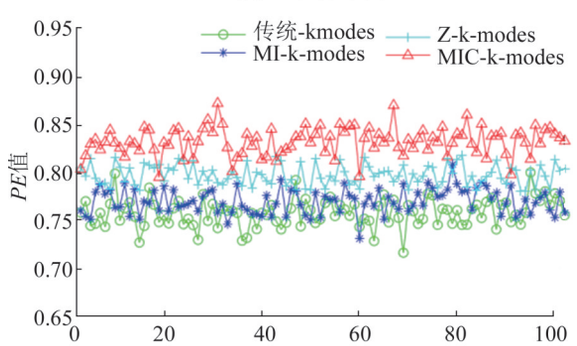
(c) *RE*值分布图

图2 100次评价标准在cancer数据集上的分布情况

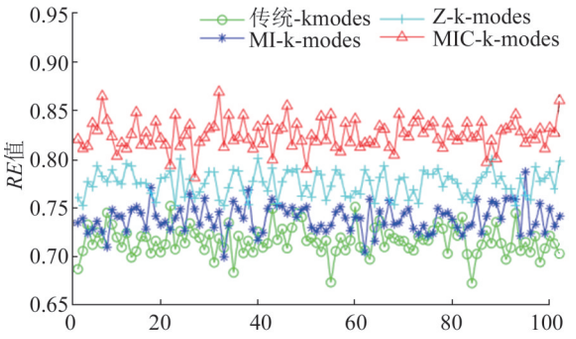
Fig. 2 Distribution of 100 rating criteria over cancer data set



(a) *AC*值分布图



(b) *PE*值分布图



(c) *RE*值分布图

图3 100次评价标准在mushroom数据集上的分布情况

Fig. 3 Distribution of 100 rating criteria over mushroom data set

表2 4种方法在vote数据集下的聚类结果

Table 2 Clustering results of 4 algorithms under the vote data set

指标	k-modes	MI-k-modes	Z-k-modes	MIC-k-modes
AC	0.815 7	0.844 9	0.866 1	0.909 3
PE	0.828 9	0.842 3	0.862 7	0.899 5
RE	0.817 2	0.831 4	0.851 3	0.889 7

表3 4种方法在cancer数据集下的聚类结果

Table 3 Clustering results of 4 algorithms under cancer data set

指标	k-modes	MI-k-modes	Z-k-modes	MIC-k-modes
AC	0.822 1	0.846 9	0.863 1	0.894 7
PE	0.818 3	0.840 0	0.861 1	0.882 3
RE	0.808 2	0.832 3	0.849 0	0.876 9

表4 4种方法在mushroom数据集下的聚类结果

Table 4 Clustering results of 4 algorithms under mushroom data set

指标	k-modes	MI-k-modes	Z-k-modes	MIC-k-modes
AC	0.723 9	0.752 1	0.795 3	0.856 2
PE	0.751 2	0.768 9	0.802 1	0.843 7
RE	0.718 9	0.745 1	0.782 4	0.833 4

表5 MIC-k-modes方法在vote数据集下的提高率

Table 5 Improvement rate of MIC-k-modes method under vote data set %

指标	对比k-modes	对比MI-k-modes	对比Z-k-modes
AC	11.475	7.622	4.987
PE	8.517	6.791	4.265
RE	8.872	7.012	4.510

表6 MIC-k-modes方法在cancer数据集下的提高率
Table 6 Improvement rate of MIC-k-modes method under the cancer data set %

指标	对比k-modes	对比MI-k-modes	对比Z-k-modes
AC	8.883	5.644	3.661
PE	7.821	5.036	2.462
RE	8.500	5.358	3.274

表7 MIC-k-modes方法在mushroom数据集下的提高率
Table 7 Improvement rate of MIC-k-modes method under mushroom data set %

指标	对比k-modes	对比MI-k-modes	对比Z-k-modes
AC	18.276	13.841	7.657
PE	12.314	9.728	5.186
RE	15.927	11.850	6.518

表8 新方法在3种数据集下对比k-modes的提高率
Table 8 The new method compares the improvement rate of k-modes under three data sets %

指标	数据集		
	vote	cancer	mushroom
对比k-modes *AC*提高率	11.475	8.883	18.276
对比k-modes *PE*提高率	8.517	7.821	12.314
对比k-modes *RE*提高率	8.872	8.500	15.927

6 结论

本文在现有k-modes聚类算法中，引入MIC的概念，对即将聚类的数据集进行了数据预处理，进而在聚类算法中定义了一种新的距离度量，该距离度量方法既保持了强相关性变量属性信息，又补充了弱相关性变量属性间的关系，通过计算它们之间的MIC值来体现变量之间的相关度，MIC值越大代表相关性越强反之越弱，使之更符合实际问题情况。通过在UCI数据集上的实验，证明了该方法的可行性和有效性。

但本文提出的基于MIC预处理的k-modes方法也有不足之处：①该方法仍然采用随机选择初始化类中心，对于初始类中心的选择方法还有待研究；②该方法的时间复杂度为$O(n^{2.4})$，因此采用MIC值对变量属性进行相关性预处理会增加一定的计算成本，如何减少计算各变量属性间的MIC值的计算量是今后需要进一步研究的内容。

参考文献：

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

[2] Ahmad A, Dey L. A Method to Compute Distance between Two Cate-Gorical Values of Same Attributein Unsupervised Learning for Categori-Cal Data Set[J]. Pattern Recognition Letters(S0167-8655), 2007, 28(1): 110-118.

[3] 张小宇, 梁吉业, 曹付元, 等. 基于加权连接度的改进K-Modes聚类算法[J]. 广西师范大学学报(自然科学版), 2008, 26(3): 189-193.
Zhang Xiaoyu, Liang Jiye, Cao Fuyuan, et al. An Improved K-modes Clustering Algorithm Based on Weighted Connectivity [J]. Journal of Guangxi Normal University (Natural Science Edition), 2008, 26(3): 189-193.

[4] 吴润秀. 基于互信息量的改进K-Modes聚类方法[J]. 统计与决策, 2012, 354(6): 89-91.
Wu Runxiu. An Improved K-Modes Clustering Method Based on Mutual Information [J]. Statistics and Decision, 2012, 354(6): 89-91.

[5] Jin-Kyu Choi, Youngcheol Shin, Soonil Hong, et al. A Study on Real-time Data Preprocessing Technique for Small Millimeter Wave Radar[J]. The Journal of The Institute of Internet, Broadcasting and Communication (S2289-0238), 2019, 19(6): 79-85.

[6] Owe Axelsson, Maya Neytcheva, Anders Ström. An Efficient Preconditioning Method for State Box-Constrained Optimal Control Problems[J]. Journal of Numerical Mathematics(S1570-2820), 2018, 26(4): 185-207.

[7] 谢娟英, 吴肇中, 郑清泉. 基于信息增益与皮尔森相关系数的2D自适应特征选择算法[J]. 陕西师范大学学报(自然科学版), 2020, 48(6): 69-81.
Xie Juanying, Wu Zhaozhong, Zheng Qingquan. A 2D Adaptive Feature Selection Algorithm Based on Information Gain and Pearson Correlation Coefficient [J]. Journal of Shaanxi Normal University (Natural Science Edition), 2020, 48(6): 69-81.

[8] 彭竹清, 王艳, 纪志成. 基于改进熵值法的离散制造系统能效评价方法[J]. 系统仿真学报, 2017, 29(9): 2121-2127.
Peng Zhuqing, Wang Yan, Ji Zhicheng. Energy Efficiency Evaluation Method of Discrete Manufacturing System Based on Improved Entropy Method[J]. Journal of System Simulation, 2017, 29(9): 2121-2127.

DOI: 10.16182/j.issn1004731x.joss.21-0484

第 34 卷第 10 期 系统仿真学报 Vol. 34 No. 10
2022 年 10 月 Journal of System Simulation Oct. 2022

[9] 徐洪峰, 孙振强. 多标签学习中基于互信息的快速特征选择方法[J]. 计算机应用, 2019, 39(10): 2815-2821.
Xu Hongfeng, Sun Zhenqiang. Fast Feature Selection Method Based on Mutual Information in Multi-label Learning [J]. Journal of Computer Applications, 2019, 39 (10): 2815-2821.

[10] 叶继华, 兰清平, 刘长红, 等. 结合互信息量和Log-Gabor特征的嵌入式人脸识别[J]. 系统仿真学报, 2016, 28(9): 2214-2219.
Ye Jihua, Lan Qingping, Liu Changhong, et al. Embedded Face Recognition Based on Mutual Information and Log-Gabor Feature [J]. Journal of System Simulation, 2016, 28(9): 2214-2219.

[11] David N Reshef, Yakir A Reshef, Hilary K Finucane, et al. Detecting Novel Associations in Large Data Sets[J]. Science(S0036-8075), 2011, 334(6062): 1518-1524.

[12] 施振佺, 陈世平. 一种改进的k-modes聚类算法[J]. 运筹与管理, 2019, 28(12): 112-117.
Shi Zhenquan, Chen Shiping. An Improved K-modes Clustering Algorithm [J]. Operations Research and Management, 2019, 28(12): 112-117.

[13] Zhou Hongfang, Zhang Yihui, Liu Yibin. A Global-Relationship Dissimilarity Measure for the k-modes Clustering Algorithm[J]. Computational Intelligence and Neuroscience(S1687-5265), 2017: 3691316.

