DOI:10.13546/j.cnki.tjyjc.2019.11.015

混合型数据聚类方法的比较

刘　超[1a,1b],姚清华[1a,1b],乐　然[2]

(1.北京航空航天大学a.数学与系统科学学院;b."数学、信息与行为"教育部重点实验室,北京 100083;
2.北京大学 前沿交叉学科研究院,北京 100871)

摘　要:为了科学使用真实世界数据,探索适用于日益常见的混合型数据的聚类方法,文章分析和比较了两种典型的混合型数据聚类方法K-prototypes与ClustMD,改进了聚类方法关键参数选择方法,并提出聚类稳定性指标。结果表明,两种聚类方法均具有很高的有效性和稳定性,各有优缺点。当数据相关性强、数据缺失严重或非连续变量较多时,建议使用K-prototypes。

关键词:混合型数据;聚类有效性;聚类稳定性

中图分类号:O212.4　　**文献标识码**:A　　**文章编号**:1002-6487(2019)11-0064-03

0　引言

近年来,丰富多样的诊疗、用药、检验等临床实践产生了大量的真实世界数据(RWD)。随着医疗实践的迫切需要,真实世界数据研究受到越来越多医学研究人员的关注。真实世界数据不仅仅是大数据,而且是多种数据来源的整合,真实世界的混合型数据正在大量涌现。研究者基于不同的研究思路提出了许多混合型数据聚类方法,其中最具代表性的两种方法是K-prototypes[1]和ClustMD[2]。但是,这两种方法应用于真实世界数据的效果、优劣以及适用范围等问题还有待研究。这些方法不仅在关键参数选择等理论上还有待完善[3],而且没有考虑真实世界数据常见的的指标或观测数据缺失对聚类结果的影响,这些问题都限制了这些方法在混合型真实世界数据上的应用效果。为有效解决以上难题,本文分析和比较了K-prototypes和ClustMD,改进了混合型数据聚类方法的关键参数选择方法,提出了聚类稳定性指标。这些改进的价值在于:一方面不仅使得聚类方法更适应混合型数据特点,而且降低了聚类方法计算时间;另一方面改进的聚类方法适用范围广,对有标签和无标签的混合型数据均适用,对混合型数据的分布也没有特定要求。这不仅为真实世界数据的分析和挖掘提供了新的思路,也为实现医疗资源优化、提高医疗服务水平提供了新的视角和解决方案。

1　两种典型的聚类方法

1.1　基本方法

设 $X=(X_1^{'},X_2^{'},\ldots,X_n^{'})^{'}$ 是m个变量n个样本构成的混合型数据矩阵,第i个样本 $X_i=(x_{i1},x_{i2},\ldots,x_{iC},x_{i(C+1)},\ldots,x_{i(C+O)},\ldots,x_{im})^{'}$,其中下标号1到C为连续型变量,C+1到C+O为顺序型变量,C+O+1到m为分类型变量。假设样本可以分为k类,第L类的中心是m维向量 V_L。

混合型数据聚类方法有两种典型的聚类思路。

第一种思路是将处理连续型数据的k-means和处理分类型数据的k-modes的分划思想相结合,代表性方法是Huang(1998)[1]提出的K-prototypes。该方法基于连续型与非连续型变量分别度量样本之间的相似性,再将二者加权量化样本之间的相似性,最终实现聚类。基本步骤如下:

(1)随机选取k个样本作为初始聚类中心。

(2)基于连续型与非连续型变量分别计算样本与聚类中心 V_L 之间的距离,选择使其最小的中心来确定其新的所属类别。具体来说,对连续型变量采用欧式距离 $d_1(X_i,V_L)=\sum_{j=1}^{C}\left|x_{ij}-V_{Lj}\right|^2$ 来计算样本之间的距离,而对非连续型变量计算海明威距离 $d_2(X_i,V_L)=\sum_{j=C+1}^{m}\delta(x_{ij},V_{Lj})$,其中 $\delta(x_{ij},V_{Lj})=\begin{cases}0,x_{ij}=V_{Lj}\\1,x_{ij}\neq V_{Lj}\end{cases}$。然后,引入权重系数 γ 对二者加权求和,得到 $d(X_i,V_L)=d_1(X_i,V_L)+\gamma^{*}d_2(X_i,V_L)$。

(3)计算各个类的中心坐标,其中连续型部分采用算术平均数,非连续型部分采用众数。

(4)反复进行步骤(2)与步骤(3)直至各个样本的类别不再发生变化,得到最终聚类结果。

第二种思路借鉴Gaussian Mixture Model的思想,引入服从多元正态分布且各个分量彼此独立的隐变量z,将其与观测变量对应,从而将非连续型数据转化为连续型数据,然后假设数据来源于混合正态分布,每个数据类对应一个子分布。代表性方法是McParland和Gormley(2016)[2]

作者简介:(通讯作者)刘　超(1977—),男,湖北武汉人,博士,副教授,研究方向:复杂数据统计方法。
姚清华(1996—),女,河北保定人,硕士研究生,研究方向:应用统计学。
乐　然(1995—),男,辽宁营口人,硕士研究生,研究方向:自然语言处理与深度学习。

提出的ClustMD。基本步骤如下：

（1）根据不同类型变量的对应规则确定观测变量对应的隐变量。对应规则如下：①对于连续型变量 $x_{ij}(1\le j\le C)$，$x_{ij}=z_{ij}\sim N\left(\mu_{ij},\sigma_{ij}^2\right)$。②对于顺序型变量 $x_{ij}(C+1\le j\le C+O)$，若有 r_j 个取值，引入 (r_j+1) 维分划向量 γ_j，其分量由公式 $\gamma_{jr}=\Phi^{-1}(\delta_r)$ 确定，并且满足 $-\infty=\gamma_{j0}<\gamma_{j1}<\gamma_{j2}<\ldots<\gamma_{jr_j}=\infty$。初始情况下设定 $z_{ij}=\gamma_{jp}$，此外，若 $\gamma_{j(p-1)}<z_{ij}\le\gamma_{jp}$，则 $x_{ij}=p$。③对于分类型变量 $x_{ij}(C+O+1\le j\le m)$，若有 r_j 个取值，则引入 (r_j-1) 维隐变量 $z_{ij}=\left(z_{ij}^1,\ldots,z_{ij}^{r_j-1}\right)\sim MVN(\mu_{ij},\sum_{ij})$，并且 $x_{ik}=\begin{cases}1 & \text{若}\max_k\left\{z_{ij}^k\right\}<0\\ p & \text{若}z_{ij}^{p-1}=\max_k\left\{z_{ij}^k\right\}\text{且}z_{ij}^k>0\end{cases}(k=1,2,3,\ldots,r_j-1)$。然后，按照上述三个对应规则将各部分隐变量排成一个向量 z_i，其维数为 $C+O+\sum_{j=C+O+1}^{M}(r_j-1)$。

（2）对样本赋予初始类别标签，并根据标签对各个类别的占比 π_k、均值 μ_k 和协差阵 $\sum_k$ 三个参数求解极大似然估计，作为参数的初值。

（3）对转换后的连续型数据集通过蒙特卡洛EM算法迭代求解 π_k、μ_k 和 $\sum_k$，直到收敛。

（4）根据收敛时的参数估计值，确定各个样本所属类别。样本 X_i 的所属类别由 $\pi_k MVN(X_i|\mu_k,\sum_k)$ 中最大的k决定。

根据ClustMD的独立性假设，各类隐变量 z_k 的协方差矩阵 $\sum_k$ 满足 $\sum_k=\lambda_k A_k$，其中 λ_k 为行列式，A_k 为一对角矩阵。对各个类别的协方差矩阵分解结果进行不同程度的简化约束，得到ClustMD算法的6种模型（mode参数）不同模型中待估计的未知参数个数不同。

1.2 聚类效果评价及改进

1.2.1 聚类有效性

聚类有效性指标反映了聚类结果与真实情况的匹配程度[4]。

（1）外部度量指标（也称为监督性度量）。包括敏感性、特异性和准确率，这三个指标值越大，说明聚类效果越好。以病人数据为例，计算公式分别为：

$$\text{敏感性}=\frac{\text{被正确分类的患病人数}}{\text{患病人数}}$$

$$\text{特异性}=\frac{\text{被正确分类的不患病人数}}{\text{不患病人数}}$$

$$\text{准确率}=\frac{\text{被正确分类的人数}}{\text{总人数}}$$

（2）内部度量指标（也称为非监督性度量）。包括类内紧密度与类间分离度。其中，类内紧密度是第 j 个类内部的数据与其聚类中心的平均距离，该值越小，说明聚类效果越好。类间分离度是第 j、s 类间中心的距离，该值越大，说明聚类效果越好。

（3）相对度量指标。该指标综合考虑类内紧密度和类间分离度来评价聚类效果，本文主要选择了DB指标（Davies-Bouldin，1979）和XB指标（Xie和Beni，1991）。DB是各个类别和与其最为相似的类之间相似度的平均值；XB是样本与其相应的聚类中心距离的平均值与聚类中心最小间距的比值。DB和XB值都是越小，说明聚类效果越好。

1.2.2 聚类稳定性

考虑到被调查者失访、数据隐私、保护等情况，为了刻画指标或观测数据缺失发生变动后的聚类结果稳定程度，本文借鉴陈韡(2010)[5]和刘新涛(2013)[6]提出了聚类稳定性的概念。主要包括聚类的指标稳定性和观测值稳定性指标。

（1）指标稳定性，反映剔除指标后的数据集聚类结果与原聚类结果的相似程度或匹配率。计算公式为 $\theta=\frac{1}{mn}\sum_{i=1}^{m}n_{(-i)}^*$，其中，$n$ 为样本量，m为指标数，$n_{(-i)}^*$ 为剔除指标i后对数据集聚类得到与原聚类结果一致的观测数量。θ 值越大，说明聚类结果的指标稳定性越好。

（2）观测值稳定性，反映基于剔除一定比例样本后的数据集聚类结果与原聚类结果的相似程度或匹配率。计算公式为 $\tau_p=\frac{\overline{n^*}}{n(1-p)}$，其中 p 为给定的缺失比例，每次对随机去除数据集中 $100p\%$ 的样本后的数据集聚类，得到与原结果一致的样本数目 n^*，多次进行上述操作计算得到均值 $\overline{n^*}$。τ_p 值越大，说明聚类结果在样本缺失比例 $100p\%$ 的观测稳定性越好。

1.3 关键参数选择方法的改进

以往的文献对在类别数k确定的情况下，往往凭借人工经验确定K-prototypes的权重γ和ClustMD的协方差矩阵特征值分解模型，这难免会增大主观因素对聚类结果的影响，从而增加较大偏差出现的可能性。但是，如果对关键参数进行穷举，则会大大加剧运行时间，降低聚类效率。因此，结合混合型数据的特点，基于相对度量指标DB和XB对关键参数选择进行了改进，步骤如下：

（1）首先根据文献给出参数的合理取值集合，具体来说，将K-prototypes的γ从0.1取到6.0，每次递增0.1；而ClustMD遍历mode参数的6种可能情况：EII、VII、EEI、VEI、EVI和VVI。

（2）对K-prototypes和ClustMD分别遍历各自的备选参数集，计算各个参数聚类结果的评价指标。

（3）如果DB和XB最小值所对应的参数是一致的，则确定为关键参数的最优选择；如果基于DB和XB确定的参数不一致，那么在DB和XB的最小值确定的聚类模型基础上，再比较其他聚类评价指标（比如敏感性、特异性和准确率），从而确定最终的最优参数。

2 实例分析

2.1 研究对象

本文从美国加州大学尔湾分校（UCI）机器学习网站（http://archive.ics.uci.edu/ml/ datasets.html）上选取了心脏病、肝炎、超声心动三个混合型医疗数据集，每个数据集不仅有明确的患者分类标签，而且包括了患者的各项临床检

测指标。基本情况如表1所示。为了清晰和准确地展示聚类效果，本文使用的是二分类数据，并且可以很方便地推广到多分类的医疗数据集以及无分类标签的数据集。此外，三个数据集的连续、分类指标的比例存着明显不同，有助于了解不同类型的混合型指标结构对聚类方法效果的影响。

表1　　三个混合型临床医疗数据集的基本情况

数据集	连续指标数	分类指标数	顺序指标数	是否有缺失	病人类别	患病人数	观测数
心脏病	5	6	2	无	患病与否	120	269
肝炎	5	12	0	无	患病与否	90	109
超声心动	8	2	0	无	患病与否	43	60

2.2　聚类结果分析

对三个混合型医疗数据集分别进行了聚类分析，结果如表2和图1所示。

表2　　K-prototypes和ClustMD聚类效果的比较

评价指标			聚类方法	肝炎	心脏病	超声心动
有效性指标	外部度量（%）	敏感性	K-prototypes	78.90	74.20	90.70
			ClustMD	78.90	78.30	95.30
		特异性	K-prototypes	58.90	82.60	100.00
			ClustMD	74.40	81.90	88.20
		准确率	K-prototypes	62.40	78.80	93.30
			ClustMD	75.20	80.30	86.70
	内部度量	类内紧密度	K-prototypes	1.4395	1.5648	0.6697
			ClustMD	1.5420	1.5766	0.6792
		类间分离度	K-prototypes	2.2732	2.4756	1.2150
			ClustMD	2.2521	2.4609	1.1780
	相对度量	XB指标	K-prototypes	0.7881	1.1894	0.8787
			ClustMD	1.2248	2.6592	2.3331
		DB指标	K-prototypes	1.5726	1.6722	1.2179
			ClustMD	1.4914	1.9816	2.0602
稳定性指标	指标稳定性	稳定系数	K-prototypes	0.8678	0.8781	0.8693
			ClustMD	0.8115	0.9016	0.8074
	观测稳定性	$\tau_{0.1}$	K-prototypes	0.8431	0.9325	0.9577
			ClustMD	0.8894	0.8166	0.9301
		$\tau_{0.5}$	K-prototypes	0.7644	0.8332	0.9089
			ClustMD	0.7560	0.6968	0.8452
		$\tau_{0.7}$	K-prototypes	0.7155	0.7719	0.8137
			ClustMD	0.6532	0.6283	0.7730
运行时间（秒）			K-prototypes	9	7	3
			ClustMD	138	409	7
连续型指标数占比（%）				22.2	38.5	80

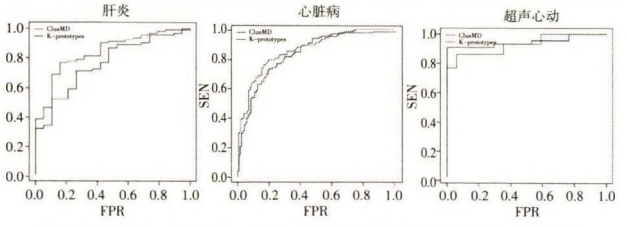

图1 K-prototypes和ClustMD对三个数据集聚类的ROC曲线比较

从表2和图1可以发现：

（1）K-prototypes和ClustMD均具有很高的有效性，但是K-prototypes优于ClustMD。各个数据集的两种聚类结果与真实情况匹配程度均高于60%，而且两种方法在外部度量（仅敏感性、特异性及ROC曲线）均大致接近，但是在内部度量与相对度量上，K-prototypes明显优于ClustMD。

（2）K-prototypes和ClustMD均具有很好的稳定性，但是K-prototypes优于ClustMD。在指标稳定性上，K-prototypes在肝炎以及超声心动数据集上整体优于ClustMD，而ClustMD在心脏病数据集上优于K-prototypes。在观测稳定性上，当数据剔除比例较大时，K-prototypes在三个数据集上均显著优于ClustMD；而在剔除比例较小时，K-prototypes在心脏病和超声心动数据集上均优于ClustMD，而ClustMD在肝炎数据集上优于K-prototypes。总的来说，K-prototypes对指标或观测数据缺失的容错能力更强，聚类结果稳定性更高。

（3）K-prototypes的计算时间显著低于ClustMD，而且当顺序型变量存在时，K-prototypes的优势更为明显。

（4）指标结构对K-prototypes和ClustMD的聚类结果有明显的影响。当连续型指标占全部指标的比例小于30%时，ClustMD的准确率高于K-prototypes；而当连续型指标所占比例较大时，K-prototypes在内部度量与相对度量上均优于ClustMD。

3　结论

本文分析和比较了两种典型的混合型数据聚类方法K-prototypes和ClustMD，而且基于混合型真实世界数据的特点，从关键参数选择与聚类结果评价两个角度对聚类方法进行了改进。结果发现：（1）基于相对度量的参数选择方法，克服了经验判断或随机选择确定关键参数对聚类结果的影响。（2）提出的聚类稳定性指标有效地度量了聚类方法对数据或指标缺失的容错能力，从而为聚类效果评价提供了新视角和新工具。（3）K-prototypes和ClustMD均符合可计算性、稳定性和可预测性的数据科学三原则，均可适用于混合型真实世界数据，但各有优缺点。当数据相关性强、数据缺失严重或非连续变量较多时，建议使用K-prototypes。

当前呈爆炸式增长的医疗知识已经远远超越了人脑的处理能力，有调查显示医生每天需要用87%的时间来阅读最新文献才能保持知识的同步更新。因此，探索和改进适用于混合型真实世界数据的聚类方法，科学合理利用真实世界数据，不仅可以帮助医生发展出临床相关性更强和成本效益更高的方法，实现个体化治疗，而且有助于医生知识的自我更新，降低医疗服务的社会总成本。

参考文献：

[1]Huang Z X. Extentions to the K-means Algorithm for Clustering Large Data Sets With Categorical Values[J].Data Mining and Knowledge Discovery,1998,(2).

[2]McParland D, Gormley I C. Model Based Clustering for Mixed Data: clustMD[J]. Advances in Data Analysis and Classification,2016,10(2).

[3]刘强，邓磊，贾振红等. 一种改进的加权K-prototypes算法[J].激光杂志，2014，35(1).

[4]刘燕驰，高学东，国宏伟等. 聚类有效性的组合评价方法[J].计算机工程与应用，2011，(19).

[5]陈黎，王雷，蒋子云.基于K-prototypes的混合属性数据聚类算法[J]. 计算机应用，2010，30(8).

[6]刘新涛，刘晓光，申琪等.合并与不合并：两个相似性聚类分析方法的比较[J].生态学报，2013，33(11).

（责任编辑/浩　天）

DOI:10.13546/j.cnki.tjyjc.2019.11.016

投资组合优化的新方法:Mean-CoVaR模型

张保帅[1],姜　婷[1],周孝华[2],段　俊[1]

(1.重庆师范大学 经济与管理学院,重庆 401331;2.重庆大学 经济与工商管理学院,重庆 400030)

摘　要:传统资产配置模型在资产组合优化过程中没有考虑系统性风险扩散,在面临金融风险尤其极端风险时,将导致资产组合遭受极大损失。为了解决这个问题,文章通过改进Markowitz的效率前沿,把引起个别标的资产收益率变动的因素纳入系统性风险考量,应用CoVaR模型衡量系统性风险扩散,构建新的基于Mean-CoVaR资产配置模型。结果表明,在考虑系统性风险冲击时,Mean-CoVaR投资组合遭受系统性风险扩散的影响显著低于传统的Mean-Variance投资组合,Mean-CoVaR模型对投资组合配置更有效率。

关键词:投资组合;优化;均值-系统性风险模型

中图分类号:F830.9　　**文献标识码:**A　　**文章编号:**1002-6487(2019)11-0067-04

0　引言

随着经济全球化进程的加快,近年来全球金融市场联系日益紧密,导致金融资产间关联程度更加密切;且随着科技进步,信息的传递越来越快,信息网络已将单个的市场连成一体,更易造成单一资产问题透过高度的市场联动而形成系统性风险。同时,在金融市场快速发展的环境下,产品不断推陈出新,相对地也容易产生极不稳定的风险情况,尤其像2008年次贷危机,造成全球投资环境恶化,个别机构或资产的巨大损失对其他机构或资产甚至是整体市场造成巨大的损失扩散。因此,在市场越发达、投资渠道越多元化的同时,风险管理更加重要。

以Markowitz为代表的传统资产组合风险管理理念只是注重自身的风险,没有考虑金融资产间风险传染效果,在极端市场风险条件下,这将导致投资组合面临极大亏损。如果能在投资组合模型中纳入风险传染的考量,将能提升分散风险的效果,进一步降低投资组合面临极端亏损的可能性,以增强投资组合的有效性,这对优化投资组合具有重要的现实意义。鉴于此,本文主要探讨将风险扩散的效果纳入到资产组合优化的统一分析框架中,检验其是否有助于降低投资组合损失的可能性。

1　模型构建及估计

随着金融市场的发展,衡量风险的方法也随之推陈出新,从一开始使用资产收益率的方差或标准差衡量风险,到之后发展出风险值(VaR)、条件风险值(CVaR)、CoVaR

基金项目:国家社会科学规划项目(18BJY09);重庆市教委科技项目(KJQN201800513);重庆市社会科学规划项目(2018PY74);重庆市教委人文社科项目(17SKG037);重庆市教委科学技术研究重点项目(KJ1600318);重庆师范大学基金资助项目(16XYY26)

作者简介:张保帅(1981—),男,河南泌阳人,博士,副教授,研究方向:金融风险管理。

姜　婷(1983—),女,湖北鄂州人,博士,副教授,研究方向:金融市场与证券投资。

周孝华(1965—),男,湖南武冈人,教授,博士生导师,研究方向:金融工程、金融市场及风险管理。

Comparison of Clustering Methods for Mixed Data

Liu Chao[1a,1b], Yao Qinghua[1a,1b], Le Ran[2]

(1.a.Mathematics and Systems Science Institute; b. LMIB of the Ministry of Education, Beijing University of Aeronautics and Astronautics, Beijing 100083, China; 2.Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China)

Abstract: In order to scientifically use real world data, this paper explores the clustering methods applicable to the increasingly common mixed medical data. The paper analyzes and compares the two typical clustering methods: K-prototypes and Clust-MD, improves the key parameter selection method, and also proposes the clustering stability index. Cases analysis results indicate that the two methods are highly effective and stable, each with advantages and disadvantages. When data correlation is strong, data missing is serious or there are relatively more non-continuous variables, K-prototypes is recommended for hybrid data.

Key words: mixed data; clustering validity; clustering stability