ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 3,May 2020,pp. 549－562
DOI:10. 16337/j. 1004-9037. 2020. 03. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于耦合度量的多尺度聚类挖掘方法

田真真[1,2,3]，赵书良[1,2,3]，李文斌[4]，张璐璐[1,2,3]，陈润资[5]

(1. 河北师范大学计算机与网络空间安全学院,石家庄,050024; 2. 河北师范大学河北省供应链大数据分析与数据安全工程研究中心,石家庄,050024; 3. 河北师范大学河北省网络与信息安全重点实验室,石家庄,050024; 4. 河北地质大学信息工程学院,石家庄,050031; 5. 河北师范大学数学科学学院,石家庄,050024)

摘　要：为了能够更好地对非独立同分布的多尺度分类型数据集进行研究,基于无监督耦合度量相似性方法,提出针对非独立同分布的分类属性型数据集的多尺度聚类挖掘算法。首先,对基准尺度数据集进行基于耦合度量的基准尺度聚类;其次,提出基于单链的尺度上推和基于Lanczos核的尺度下推尺度转换算法;最后,利用公用数据集以及H省真实数据集进行实验验证。将耦合度量相似性(Couple metric similarity, CMS)、逆发生频率(Inverse occurrence frequency, IOF)、汉明距离(Hamming distance, HM)等方法与谱聚类结合作为对比算法,结果表明,尺度上推算法与对比算法相比,NMI值平均提高13.1%,MSE值平均减小0.827,F-score值平均提高12.8%;尺度下推算法NMI值平均提高19.2%,MSE值平均减小0.028,F-score值平均提高15.5%。实验结果表明,所提出的算法具有有效性和可行性。

关键词：多尺度;聚类;分类数据;尺度转换;度量学习

中图分类号：TP391　　**文献标志码**:A

Multi-scale Clustering Mining Method Based on Coupled Metric Similarity

TIAN Zhenzhen[1,2,3], ZHAO Shuliang[1,2,3], LI Wenbin[4], ZHANG Lulu[1,2,3], CHEN Runzi[5]

(1. College of Computer and Cyber Security, Hebei Normal University, Shijiazhuang, 050024, China; 2. Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics & Data Security, Hebei Normal University , Shijiazhuang, 050024, China; 3. Key Laboratory of Network & Information Security, Hebei Normal University, Shijiazhuang, 050024, China; 4. College of Information Engineering, Hebei GEO University, Shijiazhuang, 050031, China; 5. School of Mathematical Sciences, Hebei Normal University, Shijiazhuang, 050024, China)

Abstract: To better study the non-independent and identically distributed multi-scale categorical data sets, based on the unsupervised coupling measure similarity method, a multi-scale clustering mining algorithm for non-independent and identically distributed classification attribute data sets is proposed. Firstly, the data set of benchmark scale is clustered based on coupled metric similarity method. Secondly, scale conversion algorithms upscaling based on single chain and downscaling based on Lanczos kernel are proposed for scale conversion. Finally, experiments are performed using the public data sets and the real data sets of the H province. In the experiment, couple metric similarity (CMS), inverse occurrence frequency (IOF), hamming distance (HM) and other similarity metric methods combined with spectral clustering algorithm

基金项目:国家社会科学基金重大(13&ZD091, 18ZdA200)资助项目。
收稿日期:2019-12-01;**修订日期**:2019-12-29

are compared and the experimental results demonstrate that the NMI value of the upscaling increases by 13.1%, the mean of MSE value reduces by 0.827, and the mean of F-score value increases by 12.8%. Compared with other comparison algorithms, the mean of NMI value of downscaling increases by 19.2%, the mean of MSE value reduces by 0.028, and the mean of F-score value increases by 15.5%. Experimental results and theoretical analysis show that the proposed algorithm is effective and feasible.

Key words: multi-scale; clustering; categorical data; scale conversion; coupled metric similarity

引　言

多尺度聚类是多尺度研究方法的一种，旨在根据不同的分辨率，从不同的尺度将一堆无标签的物理或抽象对象分成由相似对象组成的簇，在数据挖掘、机器学习[1-2]、计算机视觉、图像处理和信息检索任务中应用较为广泛。这些簇中的对象与本簇中的对象相似，与其他簇中的对象相异，国内外学者根据多尺度聚类的这一特性作了一系列研究。国内方面，文献[3]结合关联规则和概念分层理论，提出了一种基于关联规则的多尺度挖掘方法；文献[4]在文本丰富的多维数据集中，根据关键字搜索提出了一种基于随即投影和哈希索引结构的新方法ProMiSH，为许多新的应用程序和工具提供了便利；文献[5]将粒计算的概念引入到多尺度数据挖掘中，并利用斑块模型进行尺度转换，提出多尺度聚类尺度上推算法；文献[6]结合非参数密度估计方法，提出了一种基于多尺度信息融合的层次聚类算法，该算法能够有效解决具有流形结构的数据集中局部噪声问题；文献[7]将多尺度对象同像素联合起来，提出了利用谱聚类对高空间分辨率遥感影像进行分割的方法。国外方面，文献[8]通过对加拿大魁北克南部莱姆病发生区不同空间尺度上肩胛若虫分布的聚类分析，帮助人们了解风险变化并采取相应措施；文献[9]通过区域化算法，提出了一种基于多尺度自举重采样的平均联动层次聚类算法，对非平稳降水时间序列的均匀降水区进行识别；文献[10]提出了一种多尺度高斯核诱导模糊C均值算法，对病变进行分割以确定病变边缘。

从目前的研究情况来看，多尺度聚类已经在各个学科领域得到广泛研究；但从数据集的属性类型进行分析，大多数的研究只是针对数值型数据集，对数据进行定量的分析与预测，而对分类属性型数据集（简称为分类型数据集）进行定性分析研究的工作很少。分类型数据集大多用字符表示属性值，不具有数的大部分性质，即便使用数（整数）表示，也应当作符号，不能进行定量分析。对分类型数据集进行研究，不仅需要获取复杂的数据特征，还需要所提出的方法具有一定的灵活性。

针对存在的问题，本文的主要贡献有：(1)引入最新提出的无监督耦合度量相似性方法，提出基于耦合度量的多尺度聚类挖掘方法，对具有多尺度特性的分类型数据集进行基准尺度聚类，得到基准尺度聚类结果；(2)结合尺度转换理论以及凝聚层次聚类思想，提出基于单链的尺度上推算法，对基准尺度的聚类结果进行尺度转换，进而得到目标尺度聚类结果；(3)将尺度转换理论与兰索斯(Lanczos)插值思想相结合，并根据分裂层次聚类思想，提出基于Lanczos核的尺度下推算法，对非独立同分布的分类型数据集进行多尺度聚类尺度下推。

1　耦合度量相似性

耦合度量相似性（Couple metric similarity，CMS）是一种主要用于非独立同分布的无监督分类型数据集的相似性度量方法[11]。已经存在的度量分类型数据对象相似性的方法有汉明距离(Hamming distance，HM)、图像耦合分析[12]、发生频率(Occurrence frequency，OF)、逆发生频率(Inverse occurrence frequency，IOF)等，常用的算法有K-modes以及K-prototype算法[13]。其中汉明距离对应于基于匹配的相似性度量，使用0和1来表示不同和相同的分类值之间的相似性，发生频率和逆发生频率都是通过不同

属性值的发生频率来表示相似性,K-modes算法采用差异度来表示对象间的相似性,K-prototype则是对K-means和K-mdoes的结合,可用于同时存在数值型属性和分类型属性的数据集。文献[14]提出一种非监督耦合分类数据表示框架,用于捕获层次耦合关系;文献[15]利用概念格,提出一种新的动态加权模型来增强概念相似性测度。但这些方法都忽略了不同属性之间的关系。以表1中的数据为例,说明现有的用于分析分类型数据集相似性方法存在的挑战。人员工作统计表中,每个工作人员都由4个属性组成:性别、文化程度、职业和薪资水平。先前提出的一些相似性度量方法只考虑了对象之间的相似性,比如HM,使用HM衡量对象Staff1和Staff2之间的相似性为0.5,Staff2和Staff3之间的相似性也为0.5。但是很明显,同等教育程度和性别下,薪资水平跟职业有很大的关系。通过观察表1中的数据不难发现,文化程度在很大程度上会影响人们的职业和薪资水平,而由生活经验可知,性别对人们的工作性质也有一定的影响,因此同一属性下不同属性值和不同属性之间的关系对分类型数据集的相似性学习有很重要的参考价值。

表1 人员工作统计表

Table 1 Personnel work statistics

姓名	性别	文化程度	职业	薪资水平
Staff1	F	Master	Programmer	High
Staff2	F	Undergraduate	Programmer	High
Staff3	F	Undergraduate	Teacher	Medium
Staff4	F	High School	Repairman	Low

CMS在测量对象相似度之前,将基于频率的属性内相似度与基于共生的的属性间相似度结合起来。属性内相似性捕获属性值的频率分布和值之间的耦合,属性间相似度通过考虑不同属性属性值共现条件概率的交集来聚合不同属性值之间的属性依赖关系。CMS主要从属性内相似性、属性间相似性和耦合对象相似性来衡量两个对象之间的相似性。

定义1 属性内相似性 两个对象A和B关于属性j的属性内相似性定义为$S_{\mathrm{Ia}}(A_j,B_j)$,计算公式为

$$S_{\mathrm{Ia}}(A_j,B_j)=\begin{cases}1 & A_j=B_j\\ \dfrac{\log p\cdot\log q}{\log(p\cdot q)+\log p\cdot\log q} & \text{其他}\end{cases} \tag{1}$$

式中:$p=|N(A_j)+1|$;$q=|N(B_j)+1|$;A_j表示对象A在第j个属性上所对应的属性值;B_j表示对象B在第j个属性上所对应的属性值;$N(A_j)$表示所有在第j个属性取值为A_j的对象的集合,其中$N(A_j)+1$是为了避免分母取值为0;$|\cdot|$表示集合中元素的个数。如果属性值相同,则它们之间的属性内相似性为1;当属性值不一致时,它们的出现频率即表示它们的属性内相似性。

定义2 属性间相似性 在第j个属性中,两个属性值A_j和B_j关于除了属性j外其他属性的属性间相似性定义为

$$S_{\mathrm{Ie}}\left(A_j,B_j\right)=\sum_{k=1,k\neq j}^{d} r_{kj}S_{kj}(A_j,B_j) \tag{2}$$

式中:d表示数据集属性的个数;r_{kj}表示每个属性k到属性j的权重;$S_{\mathrm{Ie}}\left(A_j,B_j\right)$表示属性$j$中两个属性值$A_j$和$B_j$的属性间相似性;$S_{kj}(A_j,B_j)$表示属性值$A_j$和$B_j$关于属性$k$的属性间相似性,计算公式为

$$S_{kj}(A_j,B_j)=\begin{cases}1 & A_j=B_j\\ \dfrac{M}{2M-Q} & \text{其他}\end{cases} \tag{3}$$

式中:$M=\sum_{i=1}^{|W_k|}\max\left(\dfrac{|N(W_k^i,A_j)|}{|N(A_j)|},\dfrac{|N(W_k^i,B_j)|}{|N(B_j)|}\right)$;$Q=\sum_{i=1}^{|W_k|}\min\left(\dfrac{|N(W_k^i,A_j)|}{|N(A_j)|},\dfrac{|N(W_k^i,B_j)|}{|N(B_j)|}\right)$;$W_k=V_k^{N(A_j)}\cap V_k^{N(B_j)}$;$W_k$表示在第$j$个属性上取值为$A_j$的所有对象在第$k$个属性上取值的集合与在第$j$个属

性上取值为B_j的所有对象在属性k上所有可能取值的共现条件概率的交集。

定义3 耦合度量相似性 两个对象A和B之间的耦合度量相似性(CMS)定义为

$$S(A,B)=\sum_{j=1}^{d}\beta_j S_j(A_j,B_j) \tag{4}$$

式中:β_j表示属性j的耦合度量属性值相似性的权重;$S_j(A_j,B_j)$表示耦合度量属性值的相似性,即将属性值属性内的相似性与属性值属性间的相似性结合,其计算公式为

$$S_j(A_j,B_j)=\frac{1}{\alpha\cdot\frac{1}{S_{\mathrm{Ie}}(A_j,B_j)}+(1-\alpha)\cdot\frac{1}{S_{\mathrm{Ia}}(A_j,B_j)}} \tag{5}$$

式中:α表示属性值属性内的相似性和属性值属性间的相似性的加权调和平均。α越大,表明属性间耦合在对象相似性中起的作用越重要,即属性j与其他属性属性间的耦合比属性j属性内耦合更重要。

基于无监督耦合度量相似性的多尺度聚类算法,可以针对多尺度数据集中的分类型数据集进行多尺度数据挖掘,不仅能够考虑属性内之间的相互影响,还可以考虑到属性间的影响,这是耦合度量相似性的精髓所在,也是多尺度聚类数据挖掘在尺度转换时提高目标尺度聚类性能的关键。

2 多尺度聚类挖掘

多尺度聚类数据挖掘是多尺度数据挖掘算法中的一种,主要针对无标签多尺度数据集进行数据挖掘,基于耦合度量的多尺度聚类数据挖掘算法则是多尺度聚类数据挖掘的一种,主要针对非独立同分布的分类型数据集进行数据挖掘,其包含基准尺度聚类、基于单链的尺度上推和基于Lanczos核的尺度下推。

2.1 基准尺度聚类算法

2.1.1 算法思想

本文基于耦合度量相似性方法,提出了多尺度数据挖掘基准尺度聚类算法(Local scale clustering algorithm,LSCA),其基本思想是:首先根据概率密度离散化方法,利用概率密度来对表征尺度的属性进行多尺度划分,其次根据每层尺度信息熵的衰减来确定最优尺度[16],在选择好的基准尺度数据集上应用数据挖掘算法LSCA得到基准尺度聚类结果。该算法思想的具体步骤如下:

算法1 基准尺度聚类算法LSCA

输入:具有多尺度特性的原始分类型数据集Dataset

输出:最优基准尺度聚类结果

步骤:

(1)Data preprocessing /*对原始的分类型数据集进行编码,将分类型的离散或字符数据集数值化*/

LableEncoder dataset;/*对数据集进行编码,如:原始数据集中用female代表女性,经过编码后用数值0代表女性,将数据集数值化*/

(2) Select the attribute for scaling;/*选择进行尺度划分的属性*/

Choose the basic scale of dataset :

(3) $LS(d_{LS}^1,d_{LS}^2,d_{LS}^3,\cdots,d_{LS}^{m-2},d_{LS}^{m-1},d_{LS}^m)$;/*选择合适的基准尺度*/

(4) /*计算每块数据集中样本间的属性值属性内相似性*/

Use Eq.(1) to calculate $S_{\mathrm{Ia}}(A_j,B_j)$

(5) /*计算每块数据集中样本间的属性值属性间相似性*/

Get Wfunc(data,k,instance1List,instance2List)

/* Wfunc函数为属性值共现条件概率的交集*/

Use Eq.(2) to calculate coupled metric attribute value similarity $S_{1e}(A_j,B_j)$

(6) /*计算每块数据集中样本间的耦合度量相似性*/

Use formula to calculate $S_j(A_j,B_j)$

(7) /*获得数据集的连接矩阵、度矩阵以及拉普拉斯矩阵*/

D = get$D(W)$

$Ln = D-W$

Dn = np.power(np.linalg.matrix_power(D,−1),0.5)

L=np.dot(np.dot(Dn,Ln),Dn)

eigval, eigvec = getEigVec(L,cluster_num)/*获得特征值及特征向量*/

(8) /*对基准尺度上的每块数据集分别进行谱聚类,并得到基准尺度聚类结果*/

centers, centerslabel = getCenters(data,C)

/*得到基准尺度聚类结果和基准尺度样本间的相似性矩阵,并保存到文件中,方便进行尺度上推和尺度下推操作*/

2.1.2 理论基础

多尺度聚类数据挖掘中,目标尺度结果主要由基准尺度的聚类结果经过尺度转换得到,因此基准尺度在尺度转换得到目标尺度结果中具有重要作用。

(1)多尺度数据集的划分与基准尺度的选择

多尺度数据集等价划分均具有传递性、自反性和对称性[17]。数据集按不同的属性类型可以划分为分类的(定性)和数值的(定量)。分类的包括标称和序数;数值的包括区间和比率。非独立同分布分类型数据集划分方法的主要思想是将数据预处理中的概念分层方法与无监督离散化方法相结合,找出一个点或几个点对具有多尺度特性的属性进行离散化,并划分整个属性区间,以此产生属性值的多分辨率划分。文献[18]提出根据概率密度函数对数据集进行离散化尺度划分的方法,并借助得分函数评价划分点的优劣,得分函数值越大,表明划分点选择越合适。该方法不仅可以根据数据集的真实情况离散化分,同时也削弱了函数和区间划分宽度对划分的影响。

多尺度划分数据集后,数据集得到泛化,容易丢失一部分细节信息,数据集的混乱程度也变大,对基准尺度的选择成为多尺度数据挖掘的重要步骤。文献[14]提出一种基于信息熵衰减选择基准尺度的方法,该方法基于数据集划分尺度后信息量的变化情况对基准尺度进行评分,将评分最小的尺度作为基准尺度,以保证进行尺度转换时信息损失最小,减小尺度转换效应。

(2)基准尺度聚类

基准尺度的选择对于多尺度聚类数据挖掘来说非常重要,同样,基准尺度的聚类结果也很重要,聚类结果的好坏,将对尺度转换结果有很大的影响。因此,基准尺度的聚类结果应尽可能地保留数据的原始特性,如数据的信息量、异质性以及耦合性。

①信息量

所谓信息量是一种衡量数据携带信息多少的量度。数据的信息量应该以一种概率函数的形式表示,通常用信息熵表示,即$H(U)=-\sum_{i=1}^{n}p_i\log_2 p_i$,其中$p_i$表示对象出现的概率,$i=1,2,\cdots,n$。

②异质性

异质性就是一个群体里面,所有个体的特征差异程度。异质性越高,个体的特征分布越分散。一般大尺度下异质性会相对较低。

③耦合性

耦合性是程序结构中各个模块之间相互关系的度量。本文指的是属性内和属性间的相互关系,随着尺度的改变,属性间的相关性也会随之改变。

CMS相似性度量方法将属性内和属性间的相似性相结合,且该方法是基于度量的,满足度量空间的性质。度量空间在数学领域指的是一个集合,且集合中的各个元素之间的距离是可定义的,度量空间也称作距离空间,满足如下条件:

(a) 正定性:$\rho(x,y)\geqslant 0$且$\rho(x,y)=0$,当且仅当$x=y$

(b) 对称性:$\rho(x,y)=\rho(y,x)$

(c) 三角不等式:$\rho(x,y)\leqslant\rho(x,z)+\rho(y,z)$

度量空间有很多良好的性质,因此CMS具有作为度量的有效性。

谱聚类算法以谱图理论为基础,能够聚类任意分布的数据集,并且收敛于全局最优解,本质是将数据的聚类问题转换为图的最优划分问题,在数据聚类方面有很好的应用价值。该算法首先根据给定数据集生成一个描述样本间相似度的邻接矩阵,然后求出度矩阵,并根据邻接矩阵和度矩阵得出拉普拉斯矩阵,最后根据拉普拉斯矩阵得到特征值和特征向量,并构造分类器,根据特征向量完成对数据集的聚类。其中拉普拉斯矩阵具有对称性,它的所有特征值都是实数,且都大于等于0。

文献[12]分别将CMS与谱聚类和K-modes算法相结合,实验结果表明,将CMS与谱聚类结合比将CMS与K-mdoes算法结合具有更好的聚类结果。根据上述CMS的性质和谱聚类的优势,谱聚类算法与CMS相似性方法结合具有一定的理论基础,可以用于对具有多尺度特性的非独立同分布的分类型数据集进行多尺度聚类数据挖掘。

2.2 基于单链的尺度上推算法

2.2.1 算法思想

尺度转换是多尺度聚类数据挖掘的关键,是多尺度领域研究的重中之重,尺度上推算法是尺度转换的一种。本文借助凝聚层次聚类的思想,提出了多尺度数据挖掘尺度上推算法(Upscaling algorithm CMS, UACMS)。该算法的基本思想是:将每一个基准尺度聚类中心作为一个初始簇,根据相似性度量方法,将最相似的两个簇合并在一起,直到达到设定的簇的数目。具体实现步骤如下:

算法2 多尺度聚类尺度上推算法UACMS

输入:基准尺度聚类结果

输出:目标尺度聚类结果

步骤:

(1) /*将每个划分中的聚类中心作为一个输入样本,每个样本作为一个初始簇,计算每个簇之间的相似性*/

Get the similarity between each cluster

(2) /*根据单链的计算方式,计算两个簇之间的相似性,将最相似的两个簇合并*/

Merge the two most similar clusters

new_node=ClusterNode(vec,lef,right,distance)

(3) /*调取聚类结果,将所有节点都分类*/

for i, node in enumerate(self.nodes):

self.leaf_traversal(node, i)

(4) /*获得尺度上推目标尺度聚类中心*/

Get target scale clusters:Getcluster()

2.2.2 理论基础

尺度上推思想类似于层次聚类中的凝聚层次聚类。凝聚层次聚类是一种自底向上的方法,简单地说,其算法就是通过计算每个簇之间的相似性,并将相似性最高的两个簇进行合并,生成聚类树的过程。凝聚层次聚类各个簇之间相互合并的依据分为3种:单链(Single linkage)、全链(Complete linkage)和平均链(Average linkage)。

①单链

也称作最近邻(Nearest-neighbor),就是取两个簇当中相似性最大的两个样本的相似性作为这两个簇的相似性。这种合并方法容易造成一种链式(Chaining)效果,两个簇从整体来看离得相对较远,但是由于其中部分样本离得较近而合并,从而导致得到的合并簇较松散,进一步扩大了链式效应。所谓的链式效应,即前边产生的结果会对后边的结果产生一系列的影响。

②全链

全链就是将两个簇中相似性最小的两个样本间的相似性作为两个簇的相似性,效果刚好与单链相反,限制很大。

③平均链

平均链就是将两个簇中两两样本间的相似性求平均值作为两个簇之间的相似性。这种方法受异常点的影响相对较大,而且时间复杂度也比较高.

UACMS就是借助凝聚层次聚类思想中单链求两个簇之间相似性的方法,对基准尺度聚类结果进行簇合并,进而达到尺度上推的目的,其思想可用图1进行表示。图1中的一个虚线圈表示一个簇,一个实线圈表示尺度划分中的一个块。

2.3 基于Lanczos核的尺度下推算法

2.3.1 算法思想

尺度下推是尺度转换的另一种表现形式。本文借助Lanczos插值和分裂层次聚类的思想提出了多尺度聚类尺度下推算法(Downscaling based on Lanczos,DSAL)。该算法的思想是:首先得到基准尺度聚类结果,其次将基准尺度聚类结果作为已知样本,利用Lanczos核公式计算每个样本的权重,得到新的聚类中心,最后计算样本间的相似性,得到目标尺度聚类结果。尺度下推的核心相当于从宏观到微观,从展现整体的特征到显示个体特征的过程,这个过程中可以得到更多的细节信息。该过程类似于分裂层次聚类,即将一个簇不断地拆分为更多的簇。DSAL的思想如图2所示,具体实现步骤如下:

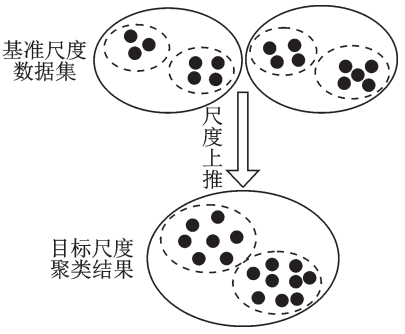

图1 尺度上推思想

Fig.1 Idea of upscaling

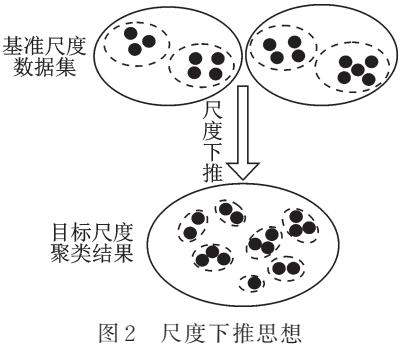

图2 尺度下推思想

Fig.2 Idea of downscaling

算法3 多尺度聚类尺度下推算法DSAL

输入:具有多尺度特性的原始分类型数据集Dataset

输出:目标尺度聚类

步骤:

(1) /*根据基准尺度基于耦合度量相似性的聚类算法,得到基准尺度聚类结果*/

Get the basic scale clustering result:$LS(C_1,C_2,\cdots,C_d)$

(2) /*根据基准尺度聚类结果和Lanczos核得到样本权重*/

Get Lanczos Kernel:$L(x)=\mathrm{sinc}(x)\mathrm{sinc}(\frac{x}{a})$

Use $L(x)$ to obtain weight coefficients

(3) /*获得样本间的相似度值,并进行尺度转换*/

for i in range(n):

for j in range($i+1$,n):

GetSimCMS(C_i,C_j)

Convert scale:Conversion(DatasetCon)

(4) /*获得尺度下推目标尺度聚类结果*/

target scale clusters:Targetcluster()

2.3.2 理论基础

在层次聚类中,凝聚层次聚类与分裂层次聚类的原理相反。分裂层次聚类采用自上而下的策略,首先将所有样本都视为属于同一个簇,然后根据它们之间的相似性逐渐进行划分,得到越来越多更小的簇,直到满足终止条件。尺度下推算法的思想与分裂层次聚类类似,但又有所不同:它是将基准尺度数据集的聚类结果作为初始数据集,然后对其进行分裂,直到满足目标尺度的终止条件。

由于多尺度领域非常注重算法效率,因此需要严格把控尺度转换所需要的时间,这也是在尺度转换过程中用基准尺度的聚类结果得到目标尺度聚类结果的一个重要原因。本文提出的尺度下推算法DSAL利用的就是对基准尺度聚类结果进行插值操作,得到每个样本的权重,然后产生新的样本点。对于一维数据集,假设输入的点为x,则Lanczos对应位置的权重计算公式为

$$L(x)=\begin{cases}\mathrm{sinc}(x)\mathrm{sinc}(x/a) & -a<x<a\\ 0 & \text{其他}\end{cases} \tag{6}$$

通常a取值为2或3,当$a=2$时适用于缩小插值;$a=3$时,适用于放大插值。通常根据输入样本x的取值,就可以确定样本x所对应的权重,同理也可以得到其他样本的权重,然后对所有需要用到的样本的取值加权平均,就可以得到想要的插值结果,即

$$S(x)=\sum_{i=\lfloor x\rfloor-a+1}^{\lfloor x\rfloor+a} s_i L(x-i) \tag{7}$$

根据已有样本点之间的关系及其取值,可以得到新样本点的取值,进而可以得到目标尺度聚类结果。

3 实 验

3.1 数据集

本文使用H省全员人口数据集(简称renkou)、UCI和Kaggle公用数据集(Zoo, Soybeanlarge, Dermatology, BreastCancer, Titanic)验证算法的有效性和可行性。表2给出了数据集的名称、属性数、样本数、类别数和有无缺失值等方面信息。其中数据集Soybeanlarge,Dermatology和BreastCancer在实验结果中分别用Sol,Der和BrC表示。

表2 数据集相关信息

Table 2 Information of the dataset

数据集	属性数	样本数	类别数	有无缺失值
Zoo	16	101	7	无
Soybeanlarge	35	306	18	有
Dermatology	33	366	6	有
BreastCancer	9	699	2	无
Titanic	5	1 309	2	无
renkou	4	5 152	5	无

实验环境为Windows版本和Windows 10专业工作站版;处理器为Intel(R) Core(TM)i7-3770 CPU @ 3.40 GHz 3.40 GHz;已安装的内

存(RAM)8.00 GB;系统类型为64位操作系统、基于x64的处理器。

算法采用python语言具体实现,实验设计思路如下:首先利用谱聚类与CMS,HM[19],OF,IOF和Eskin相似性度量方法相结合,直接在目标尺度数据集上进行聚类挖掘;为了增加实验的对比性,使用最传统的分类型数据集聚类方法K-modes算法在目标尺度上进行数据挖掘;最后使用本文算法得到目标尺度聚类结果。

3.2 实验结果

使用MSE,归一化互信息(Normalized mutual information,NMI),F-score以及运行时间4个指标对文献[13]中的不同算法以及本文提出的尺度上推算法和尺度下推算法分别在目标尺度上进行了实验对比和分析。其中用到的对比算法均为根据公式或理论将算法还原后得到,其中CMS,HM和OF方法部分数据集的NMI值、F-score值取文献[12]中的最佳运行结果,其他数据集的运行结果均是还原算法后所得。为增加对比性,也与经典K-modes算法进行了对比,其中用到的相似性度量方法为python自带的匹配方法。下文实验结果中的黑体数据表示所有算法在数据集中的最优值。

3.2.1 尺度上推

尺度上推算法即UACMS算法,通过基准尺度聚类中心得到目标尺度聚类中心,从而可以得到目标尺度聚类结果。聚类结果表明,尺度上推UACMS算法在6个数据集中有4个在NMI上优于CMS,HM,OF,IOF,Eskin和K-modes方法。

不同算法的NMI值比较结果如表3所示。由表3可知,UACMS算法的NMI值相比其他算法平均提高了13.1%,其中OF方法的平均NMI值最小。NMI值越接近于1,表明数据集的预测类标签与真实类标签越接近。UACMS算法对于Titanic和BrC数据集的NMI值不如其他方法,主要是受这两个数据集中属性间关系的影响,控制两个属性间相互影响大小的参数需要不断调节,同时一个简单的数字很难描述出属性间的复杂关系,这将是未来工作的一个挑战。

表3 尺度上推:不同算法的NMI比较

Table 3 Upsacling: NMI comparsion of different algorithms

数据集	CMS	HM	OF	IOF	Eskin	k-modes	UACMS
Zoo	0.703	0.690	0.672	0.746	0.710	0.763	**0.779**
Soybeanlarge	0.697	0.673	0.632	0.719	0.737	0.641	**0.754**
Dermatology	0.844	0.748	0.491	0.798	0.712	0.628	**0.927**
BreastCancer	**0.806**	0.749	0.801	0.646	0.711	0.576	0.727
Titanic	0.321	0.229	0.230	0.091	**0.381**	**0.381**	0.365
renkou	0.662	0.306	0.306	0.298	0.619	0.266	**0.730**
Mean	0.672	0.566	0.522	0.550	0.645	0.543	**0.714**

不同算法MSE值的比较结果如表4所示。从表4可以看出,UACMS算法在6个数据集中有4个数据集比其他所有对比算法的MSE值都小,UACMS算法的MSE值相对于其他算法平均降低了0.827,其中OF算法的MSE次之。MSE值越小,表示簇内对象之间越紧密,尽管OF方法的NMI值最小,但是其生成的簇比较紧凑。

不同算法F-score值的比较结果如表5所示。由表中的数据可知,UACMS算法的F-score值整体上高于其他对比算法,UACMS算法的F-score值相对于其他算法平均提高了12.8%,其中较为经典的K-modes算法的F-score值最小,主要是因为K-modes算法的聚类中心具有较大的随机性,且其并没有考虑到属性间的相互影响。CMS方法在多尺度数据挖掘中的应用,提高了CMS的F-score值。

表4 尺度上推:不同算法的MSE比较

Table 4 Upsacling: MSE comparsion of different algorithms

数据集	CMS	HM	OF	IOF	Eskin	k-modes	UACMS
Zoo	1.784	2.228	1.813	1.680	1.874	**1.672**	2.211
Soybeanlarge	7.083	7.084	6.773	6.865	7.428	7.659	**6.545**
Dermatology	10.700	10.403	10.262	10.107	11.293	11.668	**9.957**
BreastCancer	28.648	30.597	27.605	**27.604**	28.805	34.468	31.961
Titanic	2.862	2.284	2.832	2.601	2.910	2.910	**1.750**
renkou	11.686	8.070	8.070	11.723	10.989	8.473	**4.530**
Mean	10.461	10.111	9.559	10.097	10.550	11.142	**9.492**

表5 尺度上推:不同算法的F-score比较

Table 5 Upsacling: F-score comparsion of different algorithms

数据集	CMS	HM	OF	IOF	Eskin	k-modes	UACMS
Zoo	0.525	0.518	0.888	0.792	0.825	0.495	**0.830**
Soybeanlarge	0.528	0.504	0.480	0.654	**0.733**	0.124	0.700
Dermatology	0.762	0.660	0.615	0.883	0.762	0.293	**0.816**
BreastCancer	**0.966**	0.921	0.966	0.938	0.939	0.908	0.946
Titanic	0.825	0.772	0.698	0.695	0.844	0.854	**0.855**
renkou	0.196	0.646	0.646	0.685	0.633	0.229	**0.687**
Mean	0.634	0.670	0.716	0.775	0.789	0.484	**0.806**

运行时间是评价算法好坏的重要指标。不同算法运行时间的比较结果如表6所示。从表6可以看出,UACMS算法在所有测试数据集上均快于其他对比算法,运行时间平均提高了11.32 min;其他对比算法的运行时间整体上随数据量的增大而逐渐增加,但UACMS算法的运行时间与数据集的大小没有明显关系。因为UACMS算法的运行时间与基准尺度的聚类数目和基准尺度的划分块数有关系,不受原始数据集大小的影响。由于运行时间受算法复杂度和运行环境的影响,有时运行速度也可以靠牺牲内存来提高,因此本文中运行时间的比较结果仅为参考数据,是所有对比算法在相同的实验环境下运行得到的,具有一定的相对性。

表6 尺度上推:不同算法的运行时间比较

Table 6 Upsacling: Running time comparsion of different algorithms s

数据集	CMS	HM	OF	IOF	Eskin	k-modes	UACMS
Zoo	44.029	0.208	0.838	0.905	0.811	0.266	**0.057**
Soybeanlarge	2 267.711	1.674	16.931	18.403	16.551	3.055	**0.231**
Dermatology	3 192.750	2.101	58.454	29.160	27.747	1.647	**0.078**
BreastCancer	1 784.752	2.642	44.205	43.379	52.121	0.650	**0.030**
Titanic	1 226.818	2.535	92.365	88.922	100.047	0.898	**0.035**
renkou	10 297.337	136.095	1 305.017	1 318.840	2 270.405	4.389	**0.050**
Mean	3 135.566	24.209	252.968	249.935	411.280	1.818	**0.080**

综上,通过对比实验证明了尺度上推算法UACMS的有效性和可行性。UACMS算法相对于其他算法而言,在NMI,MSE,F-score以及运行时间方面均得到很大改善,聚类质量显著提高。

3.2.2 尺度下推

DSAL算法以及其他对比算法在不同数据集上的NMI值比较结果如表7所示。从表7中的数据可知,DSAL算法整体上的NMI值高于其他对比算法。DSAL算法的NMI值相比其他算法平均提高了19.2%,其中K-modes算法的NMI值最低,主要有两个原因:一个是所选数据集属性之间有一定的影响,而K-modes算法并没有考虑不同属性间的影响;另一个原因为K-modes算法的聚类中心具有一定的随机性,聚类结果不稳定。对比实验结果表明DSAL算法的预测结果与真实结果更为相近,对于不同属性间具有相互影响的数据集比较有优势。

表7 尺度下推:不同算法的NMI值比较

Table 7 Downsacling: NMI comparsion of different algorithms

数据集	CMS	HM	OF	IOF	Eskin	k-modes	DSAL
Zoo	0.796	0.799	0.638	0.815	0.792	0.808	**0.937**
Soybeanlarge	0.770	0.757	0.736	0.714	0.805	0.653	**0.860**
Dermatology	0.853	0.888	0.889	0.828	0.808	0.650	**0.889**
BreastCancer	0.659	0.634	0.634	0.646	**0.667**	0.576	0.640
Titanic	0.158	0.096	0.096	0.074	0.265	0.278	**0.296**
renkou	0.255	0.155	0.155	0.298	0.479	0.160	**0.910**
Mean	0.582	0.555	0.525	0.563	0.636	0.521	**0.755**

不同算法MSE值的比较结果如表8所示。从表8可以看出,DSAL算法在6个数据集中的3个上的MSE值比其他所有对比算法的MSE值都小,表明DSAL算法预测簇的紧密性与其他对比算法相比大体相近;DSAL算法的MSE值相对于其他算法平均降低了0.028,表明DSAL算法形成的簇整体比其他算法略加紧密,而IOF方法效果一般。

表8 尺度下推:不同算法的MSE值比较

Table 8 Downsacling: MSE comparsion of different algorithms

数据集	CMS	HM	OF	IOF	Eskin	k-modes	DSAL
Zoo	1.401	1.435	3.428	1.475	1.429	1.532	**1.327**
Soybeanlarge	6.586	6.630	7.221	7.621	7.217	7.582	**6.464**
Dermatology	9.974	10.262	10.288	10.725	10.545	10.367	**9.696**
BreastCancer	27.617	27.605	27.605	**27.604**	27.640	27.635	28.002
Titanic	2.619	2.549	**2.549**	2.601	2.739	2.633	2.961
renkou	10.879	4.853	**4.853**	11.723	11.656	10.890	10.042
Mean	9.846	**8.889**	9.324	10.292	10.204	10.106	9.749

不同算法F-score值的比较结果如表9所示。从表9中数据可知,DSAL算法的F-score值6个数据集中有5个都高于其他方法,而BreastCancer数据集效果不如其他方法的原因可能是其属性间的相互关系较复杂,无法用一个简单的参数表示其关系;DSAL算法的F-score值相对于其他算法平均提高了15.5%,OF方法得到的平均F-score值最小。

表9 尺度下推:不同算法的F-score值比较

Table 9 Downsacling: F-score comparsion of different algorithms

数据集	CMS	HM	OF	IOF	Eskin	k-modes	DSAL
Zoo	0.767	0.774	0.505	0.782	0.773	0.858	**0.971**
Soybeanlarge	0.702	0.722	0.634	0.639	0.718	0.632	**0.880**
Dermatology	0.918	0.812	0.848	0.864	0.762	0.831	**0.938**
BreastCancer	0.941	0.935	0.935	0.938	**0.943**	0.928	0.935
Titanic	0.738	0.698	0.698	0.678	0.804	0.714	**0.816**
renkou	0.654	0.646	0.646	0.685	0.733	0.623	**0.969**
Mean	0.787	0.765	0.711	0.765	0.789	0.764	**0.918**

运行时间是评价算法好坏的重要指标。不同算法运行时间的比较结果如表10所示。从表10可以看出,DSAL算法在所有测试数据集上均快于其他对比算法,运行时间平均提高了11.42 min。其中,由文献[12]可知,算法CMS的时间复杂度很高,其中求两个对象之间相似性的时间复杂度为$O(nm^3R^2)$,m表示属性个数,n表示对象个数,R表示不同属性值个数。将CMS方法应用到基准尺度,对原始数据集分块执行CMS方法,相当于减少了算法运行的对象个数,有时也可以减少不同属性值个数,从而减少运行时间。由表10还可知,DSAL算法的运行速度明显比原始的CMS方法快很多,CSM方法在多尺度数据挖掘中的应用也是对该方法的一种优化。其他对比算法的运行时间整体上随数据量的增大而逐渐增加,但DSAL算法的运行时间与数据集大小没有明显关系。因为DSAL算法与UACMS算法一样,运行时间与基准尺度的聚类数目和基准尺度的划分块数有关系,并不受原始数据集大小的影响。由于算法的运行时间受多种因素影响,比如运行算法的机器配置,代码优化程度等,因此本文中的运行时间仅为本实验环境下的运行时间,仅供参考。

表10 尺度下推:不同算法的运行时间比较

Table 10 Downsacling: Running time comparsion of different algorithms s

数据集	CMS	HM	OF	IOF	Eskin	k-modes	DSAL
Zoo	43.736	0.342	2.071	0.982	0.809	0.294	**0.015**
Soybeanlarge	2 466.067	8.092	17.349	18.343	16.583	3.331	**0.053**
Dermatology	3 303.631	3.187	58.247	59.937	28.071	1.998	**0.021**
BreastCancer	1 760.745	4.057	43.965	46.712	53.551	0.774	**0.078**
Titanic	1 034.446	11.473	89.457	97.620	98.447	0.966	**0.703**
renkou	10 028.442	277.075	1 316.525	1 380.572	2 398.618	5.051	**0.121**
Mean	3 106.178	50.704	254.602	267.361	432.680	2.069	**0.165**

综上,通过对比实验证明了尺度下推算法DSAL的有效性和可行性。DSAL算法相较其他算法而言,在NMI,MSE,F-score以及运行时间方面均得到很大改善,聚类质量得到显著提高。

4 结束语

现有的多尺度聚类算法主要是针对于数值属性型数据集进行定量的分析,对分类属性型数据集的研究相对较少,尽管有K-modes算法针对分类型数据集进行聚类,K-prototype算法可以对数值属性和分类属性混合的数据集进行聚类,但二者都没有考虑不同属性间的相互影响。而耦合度量相似性是一

种基于度量的相似性度量方法,并且考虑到了属性间和属性内的关系对样本间相似性的影响。因此,本文提出基于耦合度量相似性的多尺度聚类算法,针对非独立同分布的分类型多尺度数据集进行多尺度数据挖掘,并提出了基于基准尺度聚类结果进行尺度转换的尺度转换方法:基于单链的尺度上推算法UACMS以及基于Lanczos核的尺度下推算法,并且实验结果证明了所提算法的有效性和可行性。

本文中的对比算法CMS,HM,OF,IOF以及Eskin相似性度量方法都是根据参考文献[12]以及参考文献[13]中的公式经过代码还原得到,代码或许可经过牺牲内存缩短运行时间进行优化,因此实验中的运行时间仅为本实验环境下所消耗的时间,可通过优化算法节省时间,因此本文运行时间结果仅供参考。经过上面的实验可以发现所提出的算法与其他对比算法相比,在性能及效率上有了很大的提高,但是对于某些特殊数据集预测结果并未达到理想效果。在下一步工作中,首先将进一步完善用于分类型数据集的多尺度聚类算法以及对比算法,提高对比算法的运行效率,改善多尺度聚类算法,使其能在更多的数据集上得到更高的NMI值,提高预测结果的准确率以及所得簇的紧凑性,从理论和实践上寻找更适合的分类型多尺度数据集的相似性度量方法,提高聚类效率和性能;其次,属性间的相互关系并不能仅靠一个参数或者简单的加权来表示,还需要探索更好的方法来衡量对象之间的相似性,并不断完善所提出的算法。

参 考 文 献:

[1] ALABDULMOHSIN I, CISSE M, GAO X, et al. Large margin classification with indefinite similarities[J]. Machine Learning, 2016, 103(2): 215-237.

[2] GAO H, LIU X, PENG Y, et al. Sample-based extreme learning machine with missing data[J]. Mathematical Problems in Engineering, 2015(6): 1-11.

[3] 柳萌萌,赵书良,韩玉辉,等.多尺度数据挖掘方法[J].软件学报,2016,27(12):3030-3050.
LIU Mengmeng, ZHAO Shuliang, HAN Yuhui, et al. Research on multi-scale data mining method[J]. Journal of Software, 2016,27(12):3030-3050.

[4] SINGH V, ZONG B, SINGH A K, et al. Nearest keyword set search in multi-dimensional datasets[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(3): 741-755.

[5] 赵骏鹏,赵书良,李超,等.基于粒计算的多尺度聚类尺度上推算法[J].计算机应用研究,2018,35(2):362-366.
ZHAO Junpeng, ZHAO Shuliang, LI Chao, et al. A multi-scale clustering algorithm based on grain calculation [J].Application Research of Computers,2018,35(2):362-366.

[6] 李春忠,靖稳峰,徐健.基于多尺度信息融合的层次聚类算法[J].工程数学学报,2019,36(3):245-255.
LI Chunzhong, JING Wenfeng, XU Jian. Hierarchical clustering based on multi-scale information fusion [J]. Chinese Journal of Engineering Mathematics,2019,36(3): 245-255.

[7] 李军军,曹建农,程贝贝,等.联合像素与多尺度对象的高空间分辨率遥感影像谱聚类分割[J/OL].吉林大学学报(工学版):1-9[2019-06-19]. https://doi.org/10.13229/j.cnki.jdxbgxb20180537.
LI Junjun, CAO Jiannong, CHENG Beibei,et al. High spatial resolution remote sensing imagery segmentation based on combination of pixels and multi-scale objects using spectral clustering[J/OL]. Journal of JiLin university (Engineering): 1-9[2019-06-19].

[8] RIPOCHE M, LINDSAY L R, LUDWIG A, et al. Multi-scale clustering of lyme disease risk at the expanding leading edge of the range of ixodes scapularis in Canada[J]. International Journal of Environmental Research and Public Health, 2018, 15(4): 603.

[9] ZUN L C, NORISZURA I, WENDY L S, et al. The efficiency of average linkage hierarchical clustering algorithm associated multi-scale bootstrap resampling in identifying homogeneous precipitation catchments[J]. IOP Conference Series: Materials Science and Engineering,2018,342(1): 012070.

[10] PANIGRAHI L, VERMA K, SINGH B K, et al. Ultrasound image segmentation using a novel multi-scale Gaussian kernel fuzzy clustering and multi-scale vector field convolution[J]. Expert Systems With Applications, 2019, 115: 486-498.

[11] SHI Y, LI W, GAO Y, et al. Beyond IID: Learning to combine Non-IID metrics for vision tasks[C]//Proceedings of National Conference on Artificial Intelligence. San Francisco, California, USA: AAAI, 2017: 1524-1531.

[12] JIAN S, CAO L, LU K, et al. Unsupervised coupled metric similarity for non-IID categorical data[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(9): 1810-1823.

[13] BORIAH S, CHANDOLA V, KUMAR V, et al. Similarity measures for categorical data: A comparative evaluation[C]//Proceedings of Siam International Conference on Data Mining. Philadelphia, PA :SIAM, 2008: 243-254.

[14] JIAN S, PANG G, CAO L, et al. CURE: Flexible categorical data representation by hierarchical coupling learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(5): 853-866.

[15] ZHANG Q, SHI C, NIU Z, et al. HCBC: A hierarchical case-based classifier integrated with conceptual clustering[J]. IEEE Transactions on Knowledge & Data Engineering, 2019, 31(1): 152-165.

[16] 张昉,赵书良,武永亮.面向多尺度数据挖掘的数据尺度划分方法[J].计算机科学,2019,46(4): 57-65.
ZHANG Fang, ZHAO Shuliang, WU Yongliang. Data scaling method for multi-scale data mining[J]. Computer Science, 2019, 46(4): 57-65.

[17] LANGARI B, VASEGHI S, PROCHAZKA A, et al. Edge-guided image gap interpolation using multi-scale transformation [J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2016, 25(9): 4394-4405.

[18] BIBA M, ESPOSITO F, FERILLI S, et al. Unsupervised discretization using kernel density estimation[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India: [s.n.], 2007: 696-701.

[19] DIDAY E, BOCK H H. Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data[J]. Journal of Classification, 2000, 18(2): 291-294.

作者简介:

田真真(1994-),女,硕士研究生,研究方向:数据挖掘、智能信息处理,E-mail:sunshine_zhen@126.com。

赵书良(1967-),**通信作者**,男,教授,博士生导师,研究方向:数据挖掘、智能信息处理,E-mail:zhaoshuliang@sina.com。

李文斌(1974-),男,教授,研究方向:机器学习、演化计算、数据挖掘。

张璐璐(1993-),硕士研究生,研究方向:数据挖掘、智能信息处理,E-mail:sunshine_Zlulu@126.com。

陈润资(1981-),博士,研究方向:数据挖掘、智能信息处理,E-mail:498363439@qq.com。

(编辑:王静)