面向多尺度数据挖掘的数据尺度划分方法

张 昉 赵书良 武永亮

(河北师范大学数学与信息科学学院 石家庄 050024)

(河北师范大学河北省计算数学与应用重点实验室 石家庄 050024)

摘 要 多尺度挖掘在图形图像、地理信息、信号分析、数据挖掘等领域已有应用,多尺度数据挖掘在关联规则、聚类、分类挖掘领域也有相关研究与应用,但对如何对数据集进行普适性的多尺度划分以及如何构建多尺度数据集仍未展开研究,已有相关研究缺乏深度。文中从多尺度数据挖掘任务入手,定义了尺度概念,并给出了多尺度化数据集模型,以及基准尺度评分模型;依据概率密度估计的离散化方法提出了多尺度划分算法,扩展了可划分尺度的数据类型,划分结果更贴近数据的多尺度特性,且具有较低的时间复杂度;提出了多尺度化数据集方法、构建多尺度数据集算法和基准尺度选择算法,将多尺度熵与信息熵作为评价方法,在扩充多尺度化数据集方法的基础上,有效减弱了多尺度数据挖掘中因尺度推衍而产生的尺度效应,算法的时间复杂性也较为可控。利用H省真实人口数据集、UCI公用数据集和T10I4D100K数据集对所提算法和模型进行验证与实验分析,结果表明多尺度划分算法和多尺度化数据集方法是可行的,提出的多尺度化数据集方法和基准尺度评分模型是有效的,多尺度划分方法、构建多尺度数据集方法和基准尺度选择方法的应用平均提高了尺度推衍过程中1.6%的覆盖率、2.1%的F1-measure和3.7%的正确率,且具有较低的平均支持度误差。

关键词 多尺度数据挖掘,多尺度划分,离散化,构建多尺度数据集,基准尺度选择,多尺度熵,信息熵

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.009

Data Scaling Method for Multi-scale Data Mining

ZHANG Fang ZHAO Shu-liang WU Yong-liang

(College of Mathematics & Information Science, Hebei Normal University, Shijiazhuang 050024, China)

(Hebei Key Laboratory of Computational Mathematics & Applications, Hebei Normal University, Shijiazhuang, 050024, China)

Abstract Multi-scale mining has been applied in the fields of graphic images, geographic information, signal analysis, data mining, etc, and also has related research and application in the fields of association rules, clustering and classification mining. Nevertheless how to divide datasets into common scales and how to construct multi-scale datasets have not been studied in depth. Starting with the task of multi-scale data mining, this paper defined the concept of scale and gave a multi-scale dataset model and a benchmark scale scoring model. This paper proposed a multi-scale partition algorithm based on the discretization method of probability density estimation, which extends the data types of divisible scales, and its partition results are closer to the multi-scale characteristics of data with lower time complexity. This paper also proposed a multi-scale dataset method, a multi-scale data set algorithm and a benchmark scale selection algorithm. Multi-scale entropy and information entropy were used as evaluation methods. On the basis of expanding the multi-scale dataset method, the scale effect produced by the meso-scale derivation of multi-scale data mining can be effectively reduced, and the time complexity can be controlled. The proposed algorithm and model were validated and analyzed by using the real population dataset of H province, UCI common dataset and IBM dataset. The experimental results show that the proposed method is feasible and the proposed model is effective. The application of the proposed methods improves coverage by 1.6%, F1-measure by 2.1% and accuracy by 3.7% in scale deduction process, and has low average support error.

到稿日期:2018-09-04 返修日期:2018-12-25 本文受国家自然科学基金资助项目(71271067),国家社科基金重大项目(13&ZD091,18ZDA200),河北师范大学硕士基金资助项目(CXZZSS2017048)资助。

张 昉(1993-),女,硕士生,主要研究方向为数据挖掘、智能信息处理,E-mail:zhangfangapril@outlook.com;**赵书良**(1967-),男,博士,教授,博士生导师,CCF会员,主要研究方向为数据挖掘、智能信息处理,E-mail:zhaoshuliang@sina.com(通信作者);**武永亮**(1986-),男,博士生,CCF会员,主要研究方向为数据挖掘、智能信息处理。

Keywords Multi-scale data mining, Multi-scale scaling, Discretization, Construction of multi-scale datasets, Reference scale selection, Multi-scale entropy, Information entropy

1 引言

多尺度科学是基于数学、地理等基础科学理论的跨学科交叉科学，是一门同时具有多学科性和基础科学性的研究课题。多尺度科学与数据挖掘的结合是一种跨学科创新课题，国内外学者对此做了一系列相关研究。文献[1]基于空间数据的多尺度特性，将多尺度数据挖掘框架引入空间数据挖掘领域；文献[2]基于概念分层理论将多尺度数据挖掘框架引入一般数据挖掘领域；文献[3]结合地学知识尺度推衍，基于克里格尺度推衍方法，将多尺度数据挖掘引入聚类数据挖掘；文献[4]提出了多尺度深度特征学习方法，对图像提取多尺度特征后训练深度卷积神经网络；文献[5]使用多尺度熵对生物医学信号进行多尺度分析，并将多尺度科学引入信号分析领域；文献[6]提出了一种新的多尺度曝光融合算法，该算法融合不同曝光的低动态范围(LDR)图像；文献[7]针对图像分类中图像规模较大的问题，提出了多尺度的识别的字典学习方法(ML-DDL)，提高了图像分类性能；文献[8]将多尺度思想用于改进特征选择，提出了一种多尺度特征选择方法；文献[9]通过在多尺度金字塔变换中加入基于边缘方向的插值，改进了图像修复问题；文献[10-11]提出了基于相似度的频繁项集处理方法，基于金字塔理论将多尺度数据挖掘引入关联规则数据挖掘；文献[12-13]基于分形理论，结合豪斯多夫距离(HD)的相似性度量方法，将多尺度数据挖掘引入分类数据挖掘。由此可见，将多尺度科学应用于信号分析、图像处理、聚类、分类和关联规则等数据挖掘领域具有可行性，并且它能在保证一定程度的损失下明显提升挖掘效率。

但在目前的研究成果中，多尺度数据挖掘大多集中在具有明显尺度的时间、空间和图像数据，基于一般数据集的多尺度数据挖掘并未对数据尺度划分展开深度的研究。目前多尺度数据挖掘的尺度划分方法有基于概念分层[2]的多尺度划分方法，将概念分层和模糊粗糙集理论结合的模糊概念分层方法[14]，基于形式概念分析提出的形式语境中的属性粒、对象粒和关系粒的多层概念格方法[15]，基于等距离散化形成的单尺度和多尺度表示方法[11,16]和基于粒计算等价类划分理论的多尺度划分方法[17]等。其中基于概念分层的尺度化分适用范围较小，对没有明显概念分层的数据无法划分多尺度；基于等距离散化的尺度划分方法没有考虑数据本身在数据集中的尺度特性；基于粒计算等价类划分理论的尺度划分仅说明了研究理论支撑，并未提出多尺度划分方法，且目前并没有对如何构建多尺度数据集和如何选择基准尺度进行深度研究。

本文结合离散化方法，依据数据本身的尺度特性，提出了划分尺度方法；参照等价划分模型多尺度划分数据集，为削弱多尺度数据挖掘中的尺度效应，提出了利用多尺度熵和信息熵模型来构建多尺度数据集和基准尺度选择的方法；最后采用 H 省(应用户保密性要求隐去省份真实名称，而称为 H 省)真实人口数据集、IBM T10I4D100K 数据集和 UCI 公用数据集，结合多尺度数据挖掘算法，对提出的算法和模型进行验证与分析，证明算法的可行性和效率。

2 尺度定义

数据挖掘领域中“尺度”的概念最早来源于地学，是指比例尺的大小[2]。地学研究数据多为空间数据和时间数据，其尺度主要指空间、时间的固有尺度。“尺度”在生态学、图像学、物理学等中均有定义，主要指研究对象的测量单位。对于以一般数据集为研究对象的数据挖掘而言，“尺度”亦为数据的一种测量单位。一般关系型数据集(以下简称数据集)由属性和对象构成，属性的取值一般可以采用多个测量方式，比如“地域”属性可以有省、市、县、乡 4 种测量方式或者更多，每一种测量方式都对应着数据集的一种划分或者数据属性值的改变。根据数据属性测量单位定义“尺度”。

定义 1 尺度是一种具有含义的比例尺，包括范围和粒度两方面度量，其中范围尺度度量研究对象内容的大小，粒度尺度度量是在尺度范围下研究对象的最小测量单位，如图 1 和图 2 所示。

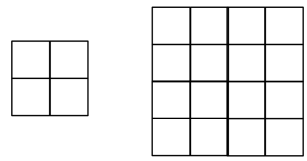

图 1 范围尺度

Fig. 1 Scope scale

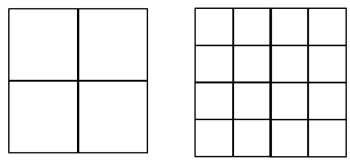

图 2 粒度尺度

Fig. 2 Granularity scale

定义 2 多尺度是指一组具有偏序结构的多个尺度 H_i 的集合[2]。

1)若 H_1 和 H_2 为相邻尺度，且 $H_1 \succ H_2$，则 H_1 为 H_2 的父尺度。

2)若 H_1 和 H_2 为相邻尺度，且 $H_1 \prec H_2$，则 H_1 为 H_2 的子尺度。

3)若 H_1 和 H_2 为不相邻尺度，且 $H_1 \succ H_2$，则 H_1 为 H_2 的祖先尺度。

4)若 H_1 和 H_2 为不相邻尺度，且 $H_1 \prec H_2$，则 H_1 为 H_2 的子孙尺度。

如 H 省真实人口数据集中“管理地名称”这一属性具有省、市、县、乡 4 个尺度，则其偏序结构为：省≻市≻县≻乡，“省≻市≻县≻乡”即为多尺度。

定义 3(多尺度数据集)　多尺度数据集是指使用多尺度表征的有偏序关系的数据集的集合。

表 1 给出了一个 2 个属性、5 个对象的多尺度数据集。从表 1 可以看出:地域属性的尺度 1 和尺度 2 的偏序关系为尺度 1≻尺度 2,学历属性的尺度 1 和尺度 2 的偏序关系为尺度 1≻尺度 2。假定地"属性为表征范围尺度的属性,学历属性为表征粒度尺度的属性,根据表 1 中多尺度数据集和对应尺度关系可构建如图 3 所示的多尺度层次结构。

表 1　多尺度数据集

Table 1　Multi-scale datasets

属性	地域		学历	
对象	尺度 1	尺度 2	尺度 1	尺度 2
v_1	河北	河北沧州	小学	小学肄业
v_2	河北	河北邢台	博士	博士结业
v_3	河南	河南漯河	中专	中专毕业
v_4	河南	河南郑州	硕士	硕士毕业
v_5	河北	河北石家庄	硕士	硕士结业

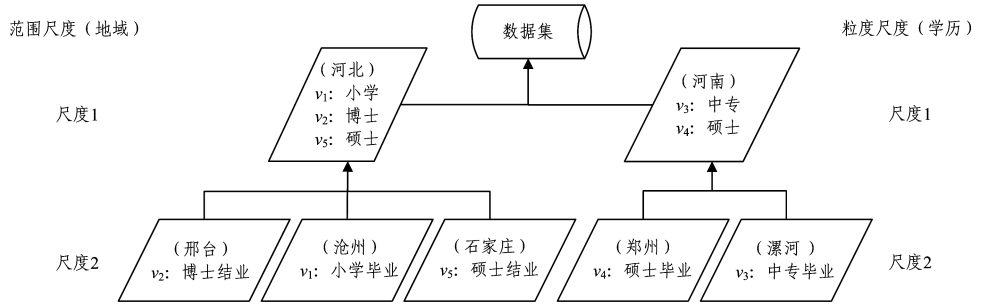

图 3　多尺度层次结构

Fig. 3　Multi-scale hierarchical structure

3　多尺度化数据集模型

根据尺度的定义推论,范围尺度是对数据集的划分,粒度尺度是对属性值的划分,划分均具有自反性、对称性和传递性,粒计算中等价划分同样具有自反性、对称性和传递性[9],参照粒计算中等价划分的模型定义多尺度化数据集模型。

定义 4(多尺度化数据集模型)　给定数据集 D,其中对象集合为 $V_i(i=1,2,\cdots)$,属性集合为 $A_j(j=1,2,\cdots)$,表征范围尺度的属性为 RS^{n_1},表征粒度尺度的属性为 GS^{n_2}。设 f 为对象集合 V_i 和属性集合 A_j 的划分关系,RS^{n_1} 对数据集的划分为 $f(RS_m^{n_1})$,GS 对属性值的划分为 $f(GS_m^{n_2})$,记 $RS_m^{n_1}$ 为数据集 D 的第 n_1 个范围尺度的第 m 层尺度,$GS_m^{n_2}$ 为数据集 D 第 n_2 粒度尺度的第 m 层尺度,$N(RS_m^{n_1})$ 为范围尺度 RS^{n_1} 的第 m 层尺度数据集划分块数,$N(GS_m^{n_2})$ 为粒度尺度 GS^{n_2} 的第 m 层尺度属性值划分的取值个数。

根据多尺度化数据集模型构建多尺度数据集模型需要解决两个问题:如何多尺度划分表征尺度的属性以及如何多尺度划分数据集。下面针对上述两个问题依次提出解决算法。

3.1　多尺度划分方法

根据定义 4,多尺度划分的对象为表征尺度的属性,多尺度划分的依据为表征尺度的属性取值。文献[11]依据统计学理论,将属性取值分为定类、定序、定比、定量 4 种数据类型。其中,定类数据指标被称为数据。定序数据指序数数据,定量和定比指数值型数据中的区间数据和比率数据,不同数据类型对应不同尺度的划分方法。对定类数据尺度划分可参照数据挖掘中数据预处理的概念分层方法[10],对定序数据和数值型数据的多尺度划分的本质为无监督离散化[11]。无监督离散化的方法主要分为等距、等频、基于概率密度的方法、基于聚类离散化的方法等,其中,基于概率密度的方法利用概率密度解释数据分布,根据数据本身的特性更加合理地离散化[18]。文献[18]提出了无监督离散化方法,使用得分函数评价离散划分点,优化离散结果,该得分函数的得分越高,说明划分点越合适。得分函数的计算公式如下:

$$Score(T)=\sum_{i=1}^{k}(p(x_i)-f(x_i))+\sum_{i=k+1}^{n}(p(x_i)-f(x_i)) \tag{1}$$

其中,T 代表划分点,$i=(1,2,\cdots,k)$ 代表位于 x_i 左边的点,$i=(k+1,k+2,\cdots,n)$ 代表位于 x_i 右边的点。其中 $f(x_i)$ 指使用直方图方法的非参数概率密度估计,计算公式如下:

$$f(x_i)=\frac{1}{hn}\# \tag{2}$$

其中,h 为划分宽度,n 为样本个数,# 代表落在 $[x_i-h/2, x_i+h/2]$ 之间的样本个数。其中,h 的选取会影响概率密度函数估计的结果。

$p(x_i)$ 指使用核密度函数来估算 x 点处的概率密度,计算公式如下:

$$p(x_i)=\frac{1}{hn}\sum_{j=1}^{n}K\left(\frac{x_i-x_j}{h}\right) \tag{3}$$

其中,n 代表样本个数,h 代表划分宽度,K 代表核函数。K 和 h 的选取对核概率密度估计的优劣有决定性作用,常用的核函数(窗函数)有方窗、高斯窗、超球窗。文献[18]提出使用得分函数评价离散化划分点的方法不仅可以削弱 K 和 h 对划分的影响,也可以根据数据的真实分布情况离散划分。

依据多尺度化数据集模型,以及范围尺度和粒度尺度均是对表征尺度属性的划分,将多尺度划分对象的范围尺度 RS 和粒度尺度 GS 均定义为尺度 S,利用文献[18]的离散化方法,迭代生成尺度 S 的多尺度划分方法。具体的代码如算法 1 所示。

算法 1　多尺度划分算法(MSA)

输入:数据集 Dataset,尺度 S,底层尺度划分块数 numOfInternals

输出:尺度 S 对应的 m 层多尺度划分结果序列 S_m

步骤:

```
1. /* 初始化尺度 S 的尺度层数 */
   m=0
```

2./*初始化第 m 层划分块数*/
$N(S_m)=1$,简记 $N(S_m)=k_m$
3./*得到第 m 层第 k_m 块的划分序列*/
$(list)_{k_m}$=Getlist_data_preprocessing(Dataset)
4./*得到第 m 层第 k_m 块的划分区间*/
$(Interval)_{k_m}$=Getinterval_list_preprocessing$(list)_{k_m}$
5./*得到第 m 层第 k_m 块的候选划分点*/
$(candidateSplitPoints)_{k_m}$ = GenerateSplitPoints(numOfInternals, $((Interval)_{k_m})$)
6./*计算每个划分点的得分*/
Foreach $(CSP)_{k_m}$ in $(candidateSplitPoints)_{k_m}$ do begin:
$(scoreCSP)_{k_m}$ = ComputeScoreOfCSP($(CSP)_{k_m}$, numOfInternals)
End for
7./*得到得分最高的划分点*/
$SplitPoints_{k_m}$.Add($(CSP)_{k_m}$, $(scoreCSP)_{k_m}$);
$(BestSP)k_m$=SplitPointk_m.GetToppestScore();
$(BestSPs)k_m$.Add($(BestSP)k_m$)
8./*尺度增加一层*/
m=m+1
9./*根据划分点存储尺度 S 第 m 层第 k_m 块的划分结果*/
S_m=get_scale($(BestSP)k_m$)
10./*更新第 m 层尺度划分后的 k_m*/
k_m=update(k_m)
11./*根据 m 层划分尺度 S_m 更新候选划分区间序列*/
$(Interval)_{k_m}$=update($(Interval)_{k_m}$, S_m)
12./*更新尺度目标划分个数*/
numOfInternals=(numOfInternals−1)/2
13.对 m 层每块重复执行步骤 5−步骤 13,直到 k_m=numOfInternals 时停止循环
14./*循环输出每一层尺度划分结果*/
Foreach m do begin
15. Return S_m

分析多尺度划分算法,最终输出 m 层多尺度划分结果序列 S_m,根据 $Score(T)$的计算公式,可以得到计算 n 个尺度目标划分的时间复杂度为 $O(n)$[18],多尺度划分算法(MSA)中,划分过程至多执行了 $\log n$ 次。因此多尺度划分算法(MSA)的时间复杂度为 $O(n\log n)$。

3.2 多尺度划分数据集方法

依据定义 4,多尺度划分对象为范围尺度和粒度尺度,若划分尺度属于表征范围的尺度,则多尺度划分数据集指将数据集按多尺度划分结果对数据集分块;若该划分尺度属于表征粒度的尺度,则多尺度划分数据集指按多尺度划分结果更新属性值。

根据 3.1 节的尺度划分方法和多尺度化数据集模型,提出多尺度化数据集(MSDSA)方法,具体如算法 2 所示。

算法 2 多尺度划分数据集方法(MSDSA)

输入:原始数据集 Dataset,范围尺度 $RS^{n_1}_{m_1}$,粒度尺度 $GS^{n_2}_{m_2}$
输出:$f(RS^{n_1}_{m_1})$,$f(GS^{n_2}_{m_2})$
步骤:
1./*按照第 i 个 RS 尺度的第 j 层尺度的划分进行数据集切割*/
foreach $RS^i_{m_1}$ in $RS^{n_1}_{m_1}$ do begin:
foreach RS^i_j in $RS^i_{m_1}$ do begin:
$f(RS^i_j)$=cut_dataset(Dataset, RS^i_j)
End for
End for
2./*按照第 i 个 GS 尺度的第 j 层尺度的划分进行尺度值的更新*/
foreach $GS^i_{m_2}$ in $GS^{n_2}_{m_2}$ do begin:
foreach GS^i_j in $GS^i_{m_2}$ do begin:
$f(GS^i_j)$=update_value(Dataset, GS^i_j)
End for
End for

根据算法 2 可知,多尺度划分数据集方法存在嵌套循环,因此时间复杂度为 $O(n*m)$,其中 n 代表尺度个数,m 代表尺度层数,在实际应用中,n 和 m 的大小较为可控。

4 构建多尺度数据集

构建多尺度数据集是多尺度数据挖掘的基础,也是多尺度数据挖掘算法的关键,多尺度划分呈现了属性的多尺度特性,提供了属性的多尺度测量方法,将构建多尺度数据集转化为组合多尺度的问题。进行多尺度划分后将会使数据集的复杂程度发生变化,“熵”是系统复杂性和规则性的一种测度,各种演化“熵”算法在生理、医学、交通等领域得到了广泛应用[19]。利用“熵”作为评价方法,逐步寻优组合尺度可以有效地构建多尺度数据集。

多尺度划分与时间序列尺度划分相似,时间序列的尺度划分中多尺度熵(MSE)、样本熵和相似熵是分析时间序列复杂程度的方法。MSE 可以在不同时间尺度下分析时间序列的复杂度。MSE 将原始时间序列多尺度化,计算相应的样本熵,随着尺度增大,样本熵变大,说明该序列的复杂性随着尺度的增大而变大[20]。可见 MSE 的基础是样本熵,样本熵是在近似熵的基础上所做的改进,样本熵的值越小,序列自我相似性就越高;样本熵的值越大,样本序列就越复杂。文献[20]对样本熵的定义如下:一个长度为 N 的原始时间序列 $X=\{u(i),u(i+1),\cdots,u(N)\}$,$u(i)$表示第 i 时间位置的取值,对原始时间序列重组构造 m 维矢量 $X^m(i)$,则 $X^m(i)$的构造公式为 $X^m(i)=\{u(i),u(i+1),\cdots,u(i+m-1)\}(i=1,2,\cdots,N-m+1)$,其中,$X^m(i)$即为对时间序列的第 m 层尺度划分。

样本熵的定义如下:

$$SampEn(m,r,n)=-\ln[\phi^{(m+1)}(r)/\phi^m(r)] \tag{4}$$

相似熵的定义如下:

$$ApEn(m,r)=\lim_{N\to\infty}(\phi^m(r)-\phi^{(m+1)}(r)) \tag{5}$$

其中:

$$\phi^m(r)=\frac{1}{N-m+1}\sum_{i=1}^{N-m+1}\ln(C_i^m(r)) \tag{6}$$

$$C_i^m(r)=\frac{1}{N-m+1}num\{d[X(i)-X(j)]<r]\} \tag{7}$$

$$d[X(i)-X(j)]=\max_{k=N-m+1}|u(i+k)-u(j+k)| \tag{8}$$

利用多尺度熵的计算方法计算尺度复杂性,m 层尺度复杂性的计算公式如下:

$$ScaleComplexity_m=\phi^m(r)=-\frac{1}{N_m}\sum_{i=1}^{N_m}\ln(C_i^m(r)) \tag{9}$$

其中,

$$C_i^m(r)=\frac{1}{N_m}num\{d[X(i)-X(j)]<r_m]\},\quad j=1,2,\cdots,N_m \tag{10}$$

$$d[X(i)-X(j)]=\max_{k_{m_i},k_{m_j}}|u(k_{m_i})-u(k_{m_j})| \tag{11}$$

其中,N_m 表示 m 层划分块数,k_{m_i} 表示第 m 层第 i 个划分块中对象的个数,$u(k_{m_i})$ 表示第 m 层第 i 个划分块内的第 k_{m_i} 个对象取值。文献[20]中定义 $r=|0.1-0.25SD|$,SD 为第 m 层尺度的标准差。尺度越大,尺度复杂性就越低,因此在构建多尺度数据集时,应逐步选择尺度复杂性较小的尺度。

进行多尺度划分数据集后,数据集的混乱程度也会发生改变,信息熵是数据集混乱程度的一种度量,尺度划分越细,数据集的混乱程度越低[19]。在构建多尺度数据集的过程中不仅要考虑尺度复杂性,也要考虑数据集划分后信息熵的大小[19],根据信息熵的计算公式,定义第 m 层尺度的信息熵,如式(12)所示:

$$comentropy_m=\sum_{i=1}^{N_m}(-\sum_j p_{ij}\log p_{ij}) \tag{12}$$

其中,p_{ij} 指第 i 个划分块中第 j 个对象的出现概率;对数一般取 2 为底。在构建多尺度数据集时应逐步选择尺度复杂性低、信息熵高的尺度。目标函数为:

$$F(RS_m^n)=\max_m(comentropy_m-ScaleComplexity_m) \tag{13}$$

构建多尺度数据集算法的描述如算法 3 所示。

算法 3　构建多尺度数据集算法(BMDSA)

输入:尺度 $S_{m_n}^n$,原始数据集 Original Datasets

输出:多尺度数据集 Multiscale Datasets,组合尺度列表 Target_scale_list

1. /* 初始化尺度列表 */
   Target_scale_list={}
2. /* 初始化候选尺度列表 */
   Candidate_scale_list{$S_1^1,S_1^2,\cdots,S_1^n,S_2^1,S_2^2,\cdots,S_2^n,\cdots,S_{m_i}^i,\cdots$}
3. /* 对候选尺度列表前 n 尺度计算尺度复杂性和信息熵 */
   Foreach $S_{m_i}^i$ in Candidate_scale_list do begin
   Use formula (9) to calculate $ScaleComplexity_{m_i}$
   Use formula (12) to calculate $comentropy_{m_i}$
   End for
4. /* 根据式(13)选择目标尺度 */
   Use formula (13) to calculate $F(S_{m_i}^i)=m'$
5. /* 更新目标尺度列表 */
   Target_scale_list={$S_{m_i}^{i'}$}
6. /* 更新候选尺度列表 */
   Candidate_scale_list{$S_1^1,S_1^2,\cdots,S_1^n,S_2^1,S_2^2,\cdots,S_2^n,\cdots,S_{m_i}^i,\cdots$}−{$S_{m_i}^{i'}$}
7. 重复执行步骤 3−步骤 6,直到
   Candidate_scale_list=null
8. /* 按照 Target_scale_list 的组合 MSDSA($S_{m_i}^i$)形成 Multiscale Datasets */
   Foreach $S_{m_i}^i$ in Target_scale_list do begin
   MSDSA($S_{m_i}^i$)
   End for

构建多尺度数据集算法是将尺度 $S_{m_n}^n$ 按照目标函数更新为组合尺度列表 $Target\_scale\_list$,而计算尺度复杂性和信息熵的时间复杂度均为 $O(\sum_{i=1}^{N_m}\ln(N_m^{N_m}))=O(N_m\log N_m)$。根据步骤 3,对候选尺度列表前 n 尺度计算尺度复杂性和信息熵,因此 BMDSA 的整体时间复杂度为 $O(nN_{m_n}\log N_{m_n})$,其中 n 为尺度个数,N_{m_n} 为第 n 尺度 m 层划分的块数。

以表 1 两个尺度的数据集为例,构建多尺度数据。假设两个尺度均为表征范围的尺度 RS_i^1,RS_j^2,其中 $i=(1,2)$,$j=(1,2)$,RS_1^1,RS_2^1,RS_1^2,RS_2^2 分别表示地域属性尺度 1,地域属性尺度 2,学历属性尺度 1,学历属性尺度 2。

$$ScaleComplexity_1^1=-\frac{1}{2}(\ln\frac{1}{2}+\ln\frac{1}{2})=-\ln\frac{1}{2}$$

$$comentropy_1^1=-\frac{1}{2}(\log\frac{1}{2}+\log\frac{1}{2})=-\log\frac{1}{2}$$

$$ScaleComplexity_1^2=-\frac{1}{4}(\ln\frac{1}{4}+\ln\frac{1}{4}+\ln\frac{1}{4}+\ln\frac{1}{4})=-\ln\frac{1}{4}$$

$$comentropy_1^2=-\frac{1}{4}(0+0+0+0)=0$$

$(comentropy_1^1-ScaleComplexity_1^1)>(comentropy_1^2-ScaleComplexity_1^2)$

因此第一个尺度应选择 RS_1^1,经过计算:

$(comentropy_2^1-ScaleComplexity_2^1)=(comentropy_1^2-ScaleComplexity_1^2)=(comentropy_2^2-ScaleComplexity_2^2)=\ln\frac{1}{4}$

又因为 $RS_1^1>RS_2^1$,$RS_1^2>RS_2^2$,所以可组合 3 种尺度列表:

$Target\_scale\_list\_1=\{RS_1^1,RS_2^1,RS_1^2,RS_2^2\}$

$Target\_scale\_list\_2=\{RS_1^1,RS_1^2,RS_2^1,RS_2^2\}$

$Target\_scale\_list\_3=\{RS_1^1,RS_1^2,RS_2^2,RS_2^1\}$

5　基准尺度评分模型

对于构建好的多尺度数据集,如何选择基准尺度进行尺度推衍是多尺度数据挖掘的关键步骤,数据集在不同尺度下具有的信息量是不一样的,即每转换一层尺度,信息量都会改变,基于信息熵的衰减定义基准尺度评分模型。

定义 5(基准尺度评分模型)　设多尺度数据集 Multiscale Datasets 有 m 层尺度 $H_m(m=1,2,\cdots,i,\cdots,m)$,$1<i<m$,尺度 H_i 推衍到尺度 H_1 的信息熵衰减为:

$$EntropyDecay_{up}=\prod_i^2|comentropy_i-comentropy_{(i-1)}| \tag{14}$$

尺度 H_i 推衍到尺度 H_m 的信息熵衰减为:

$$EntropyDecay_{down}=\prod_i^{m-1}|comentropy_i-comentropy_{(i+1)}| \tag{15}$$

尺度 H_i 的基准尺度评分公式为:

$$Benchmark\_scale\_score_{H_i}=EntropyDecay_{up_{H_i}}+EntropyDecay_{down_{H_i}} \tag{16}$$

依据定义 5,基准尺度评分越小,说明信息熵衰减越少。依据尺度效应定义[4],将基准尺度评分最小的尺度选为基准尺度进行尺度推演时,信息损失最低,最能有效衰弱尺度效应。基于基准尺度评分模型提出基准尺度选择方法(BSSM)(见算法 4)。

算法 4 基准尺度选择算法(BSSM)

输入:多尺度数据集 Multiscale Datasets,尺度列表 scale_list

输出:基准尺度 H_k

步骤:

1. /* 对尺度列表中的每一层中间尺度,计算基准尺度评分 */

   Foreach H_i in scale_list

   Use formula (16) to calculate

   Benchmark_scale_score$_{H_i}$

   End for

2. /* 计算基准尺度评分最低的尺度 */

   $H_k=\min_{H_i}$(Benchmark_scale_score$_{H_i}$)

3. Return H_k

计算 $EntropyDecay_{up}$ 和 $EntropyDecay_{down}$ 的时间复杂度为 $O(mN_m\log N_m)$,计算 *Benchmark_scale_score* 的时间复杂度为 $O(mN_m\log N_m)$,基准尺度选择算法循环计算 *Benchmark_scale_score*,BSSM 的时间复杂度为 $O(m^2N_m\log N_m)$。

6 实验

文献[16]提出的多尺度化方法和 MSCSUA[12] 使用的多尺度化方法与本文提出的多尺度划分算法(MSA)均是将数据类型分为定类、定序、定比、定量 4 种。对于定类数据,其层次尺度明显,划分方法主要采用已有的概念分层方法,而对于定比、定量、定序数据,MSCSUA[12] 的多尺度化方法按照等距划分尺度,如第 m 层尺度第 i 块划分的计算公式为$[\frac{i-1}{n_m}(V_{max}-V_{min}),\frac{i}{n_m}(V_{max}-V_{min})]$,其中 n_m 表示第 m 层尺度下划分的块数,V_{max} 和 V_{min} 分别表示待划分属性的属性值最大值和最小值。以 H 省真实人口数据集中部分数据为例(见表 2)进行实验。

表 2 H 省真实人口数据集中的部分数据

Table 2 Partial data of real population in H province

对象	年龄/岁	身高/cm
v_1	22	167
v_2	10	122
v_3	22	157
v_4	24	165
v_5	50	154
v_6	99	163
v_7	23	190

文献[16]提出的多尺度化方法和 MSCSUA[12] 使用的多尺度化方法对年龄尺度进行划分,结果如图 4 所示。

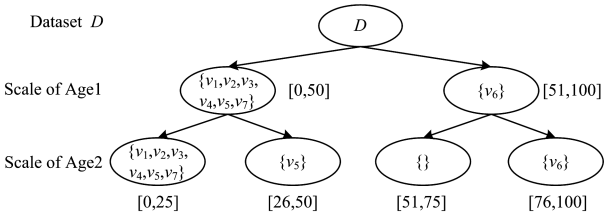

图 4 年龄的多尺度层次结构(MSCSUA)

Fig. 4 Multi-scale hierarchical structure of age(MSCSUA)

文献[16]提出的多尺度化方法和 MSCSUA[12] 使用的多尺度化方法对身高尺度进行划分,结果如图 5 所示。

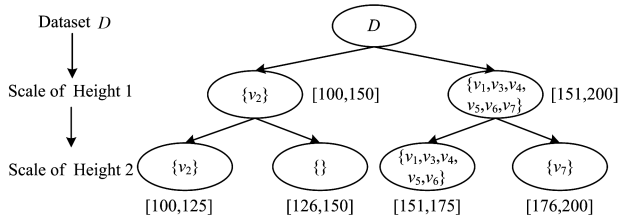

图 5 身高的多尺度层次结构(MSCSUA)

Fig. 5 Multi-scale hierarchical structure of height(MSCSUA)

由图 4、图 5 可知,该方法仅能通过先验知识设定,不能形成层次结构且不能体现数据集中数据的原有尺度特性,从而会影响多尺度数据挖掘结果。

多尺度划分算法(MSA)对年龄尺度进行划分,结果如图 6 所示。

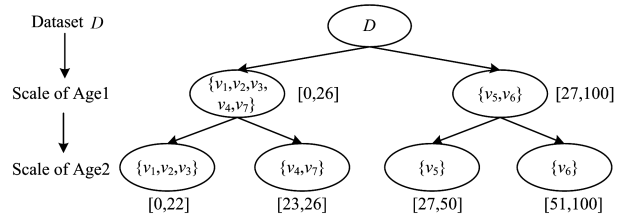

图 6 年龄的多尺度层次结构(MSA)

Fig. 6 Multi-scale hierarchical structure of age(MSA)

多尺度划分算法(MSA)对身高尺度进行划分,结果如图 7 所示。

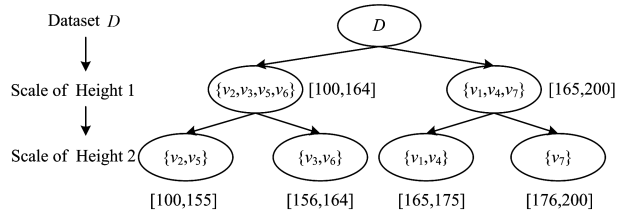

图 7 身高的多尺度层次结构(MSA)

Fig. 7 Multi-scale hierarchical structure of height(MSA)

构建多尺度数据集算法(BMDSA)组合了身高、年龄的尺度层次结构,实验结果如图 8 所示。

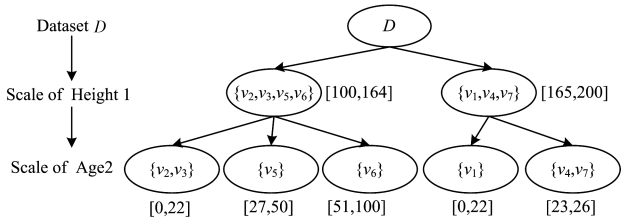

图 8 组合尺度层次结构

Fig. 8 Composite scale hierarchy

对比文献[16]提出的多尺度化方法和 MSCSUA[12] 使用的多尺度化方法,本文提出的多尺度划分算法(MSA)和构建多尺度数据集算法(BMDSA)使用评分函数对每层划分点进行划分,生成尺度 S 对应的 m 层多尺度划分结果序列 S_m,不仅考虑了数据本身的尺度特性,而且可以组合多个尺度生成多尺度层次树。

为验证本文方法的有效性,本文采用 UCI 数据集、IBM 数据集和 H 省真实人口数据集进行实验,分别验证多尺度划

分算法和基准尺度选择方法的有效性和构建多尺度数据集模型的多尺度关联规则和多尺度分类挖掘的可行性。实验环境为 CPU Intel Core i7-6700 3.41 GHz、内存 16 GB、操作系统 Windows10x86，数据库环境为 Oracle 10g，实验算法使用 Python 语言编写，开发工具为 Python3.6、pycharm2017.3、Eclipse jdk 1.7 和 Pydv3.9。实验中，挖掘基准尺度数据的频繁项集所采用的算法是经典 Apriori 算法。

表 3、表 4 为实验数据集的相关信息，其中包括数据集名称、项目数、事务数、平均事务长度、样本数、特征数、类别信息。其中，项目数指数据集中不同项集的个数，事务数指数据集中记录的条数，样本数指数据集中的样例个数，特征数指特征属性个数。将文献[10-11]中的 MSARSUA 和 SU-ARMA 算法表现相对较好的参数设置作为本实验参数设置，如表 5 所列，实验数据集的划分尺度信息如表 6 所列。

表 3　MSARSUA 和 SU-ARMA 实验数据集的相关信息对比

Table 3　Relevant information of MSARSUA and SU-ARMA experimental datasets

数据集	项目数	事务数	平均事务长度
Connect	128	67557	43
Pumsb	2113	49046	74
T10I4D100K	870	80000	27
H 省真实人口数据	36	140663	7

表 4　MSCSUA 和 SLAD 实验数据集的相关信息

Table 4　Relevant information of MSCSUA and SLAD experimental datasets

数据集	样本数	特征数	类别数
Ionosphere	351	34	2
PID	768	8	2
Spambase	4601	57	2
wine	178	13	3
H 省部分真实人口数据	6311	7	3

表 5　实验参数

Table 5　Experimental parameters

数据集	MSARSUA 最小支持度	SU-ARMA 最小支持度
Connect	0.28	0.28
Pumsb	0.23	0.24
T10I4D100K	0.033	0.034
H 省真实人口数据	0.26	0.24

表 6　实验数据集划分尺度和目标尺度

Table 6　Division scale and target scale of experimental datasets

数据集	MSARSUA 划分尺度	SU-ARMA 划分尺度	本文划分尺度	目标尺度
Connect	2	2	2	1
Pumsb	2	2	2	1
T10I4D100K	2	2	2	1
H 省真实人口数据	3	3	3	1

本文实验主要从覆盖率、F1-measure、平均支持度估计误差以及运行时间 4 个方面对文献[11]提出的 MSARSUA 和应用构建多尺度数据集方法（BMDSA）和尺度划分方法（MSA）后的 MSARSUA 方法进行对比实验，并对基准尺度的选择方法做了对比验证。从正确率（Accuracy）、F1-Measure、标准化互信息（NMI）以及运行时间 4 个方面对文献[21]提出的 SLAD 和应用 BMDSA 和 MSA 后的 SLAD 方法进行对比实验。

由表 7、表 8 可知，在数据集 Connect，Pumsb，T10I4D100K 和 H 省真实人口数据集上，MSARSUA 相对于 SU-ARMA 算法有较高的覆盖率和 *F1-Measure*，平均支持度误差较低。文献[10]采用概念分层构建多尺度化数据集，本质上与文献[11]定距离散化构建多尺度数据集相同，因此将构建多尺度数据集（BMDSA）算法应用于 MSARSUA 算法，并进行分析比较。

表 7　在数据集 Connect 和 Pumsb 上的实验结果

Table 7　Experimental results on Connect and Pumsb datasets

数据集	Connect			Pumsb		
基准尺度	2	2	2	2	2	2
评价指标	MSARSUA	SU-ARMA	MSARSUA+（BMDSA+MSA）	MSARSUA	SU-ARMA	MSARSUA+（BMDSA+MSA）
覆盖率/%	92.91	83.46	94.17	100	97.50	100.00
F1-Measure/%	91.83	88.70	93.78	93.92	99.38	99.38
平均支持度误差/%	3.66	4.72	2.92	1.55	2.86	1.55
运行时间/s	4.42	4.13	5.05	3.32	3.01	5.36

表 8　在数据集 T10I4D100K 和 H 省真实人口数据上的实验结果

Table 8　Experimental results on T10I4D100K and H provincial population data

数据集	T10I4D100K			H 省真实人口数据			
基准尺度	2	2	2	3	3	2	3
评价指标	MSARSUA	SU-ARMA	MSARSUA+（BMDSA+MSA）	MSARSUA	SU-ARMA	MSARSUA+（BMDSA+MSA+BSSA）	MSARSUA+（BMDSA+MSA）
覆盖率/%	97.78	97.56	97.78	96.72	96.36	100	98.03
F1-Measure/%	98.88	97.56	98.88	98.33	98.15	99.10	98.72
平均支持度误差/%	1.59	1.49	1.50	3.25	4.03	2.71	3.15
运行时间/s	26.05	26.00	26.10	4.09	4.01	5.80	5.80

由表 7、表 8 可知，在数据集 Connect、Pumsb 和 H 省真实人口数据 3 个数据集上，MSARSUA+（BMDSA+MSA）相对于 MSARSUA 明显提高了覆盖率和 *F1-Measure*，明显降低了平均支持度误差。在数据集 T10I4D100K 上，MSARSUA+（BMDSA）相对于 MSARSUA 的覆盖率和 *F1-Measure* 基本没有提高，平均支持度误差也没有降低。分析 T10I4D100K 数据集中表征范围的尺度属性分布情况属于均匀分布，使用多尺度划分算法（MSA）在划分效果上与文献

[11]定距离散化基本相同,推断多尺度划分算法(MSA)更加适用于属性值分布并不均匀的表征尺度的属性上。

根据实验数据,多尺度划分算法(MSA)和构建多尺度数据集(BMDSA)算法的运行时间在数据集 T10I4D100K 上比在 Connect、Pumsb 和 H 省真实人口数据 3 个数据集上要少,在数据集 Pumsb 上的运行时间增长比例最多。观察评价指标可以发现,在数据集 Pumsb 上,覆盖率、*F1-Measure* 和平均支持度误差提高比例最大,分析对应尺度化属性可以发现,数据集 Pumsb 的属性值分布不均匀。这从另一方面验证了多尺度划分算法(MSA)更加适用于属性值分布并不均匀的尺度推理。

由于 Connect,Pumsb 和 T10I4D100K 这 3 个数据集均划分两个尺度层次,为验证基准尺度选择方法(BSSM)的有效性,将 H 省真实人口数据集划分为 3 个尺度,按照基准尺度选择方法(BSSM)得到基准尺度为第 2 层,对比使用第 3 层尺度的 MSARSUA 算法可以发现,使用第 2 层尺度比使用第 3 层尺度在覆盖率和 *F1-Measure* 均有所提高,也降低了平均支持度误差,验证了基准尺度选择方法(BSSM)的有效性。

文献[21]中的 MSCSUA 算法将实验数据集划分为两层尺度。为验证同等条件下 BMDSA 和 MSA 的有效性,本文也将 Ionosphere、PID、Spambase、wine、H 省部分真实人口数据等数据集划分为两层尺度,通过分析和对比发现,应用 BMDSA 和 MSA 的 MSCSUA 算法在标准化互信息(NMI)上提高明显,平均提高了 8.07%,标准互信息主要衡量通过真实类别标签集与实验得到的类别标签集之间的差异大小,当差异增大时,NMI 值会迅速降低,即放大差异。应用 BMDSA 和 MSA 的 MSCSUA 算法的 *Accuracy* 平均提高了 2%,Accuracy 在数据集 PID 上提高最明显,分析发现,PID 尺度化属性的属性值分布不均匀,其对应尺度化时间的消耗较长。

表 9、表 10 主要是将本文提出的应用 BMDSA 和 MSA 算法后的 SLAD 和文献[21]中提出的方法 SLAD 进行对比。SLAD 算法基于 LAD 框架,通过利用数据的统计信息来提高分类性能。文献[21]中的 *NMI* 值都保留了小数点后两位数字,因此不能与本文的 *NMI* 作精确比较,但是可以看出,SLAD 算法和应用 BMDSA 和 MSA 算法的 SLAD 算法的 *NMI* 值差异不大,且提高了准确度。由此可见,本文提出的算法具有一定的可行性。

表 9 在不同数据集上 BMDSA+MSA 与 MSCSUA 的对比实验结果

Table 9 Comparison of experimental results among BMDSA, MSA and MSCSUA on different datasets

数据集	*Accuracy*		*F1-Measure*		标准化互信息(NMI)		运行时间/s	
	MSCSUA	MSCSUA (BMDSA+MSA)	MSCSUA	MSCSUA (BMDSA+MSA)	MSCSUA	MSCSUA (BMDSA+MSA)	MSCSUA	MSCSUA (BMDSA+MSA)
Ionosphere	86.2857	87.1524	0.8522	0.8601	0.4026	0.4823	0.002	0.013
PID	76.8333	79.6940	0.6887	0.7012	0.1545	0.1620	0.002	0.056
Spambase	82.1384	85.1120	0.8076	0.8871	0.3178	0.3319	0.002	0.023
wine	97.7528	97.8501	0.9780	0.9780	0.9088	0.9987	0.004	0.089
H 省部分人口数据	97.5071	97.5071	0.9699	0.9771	0.8879	0.9004	0.002	0.018

表 10 在不同数据集上 BMDSA+MSA 与 SLAD 的对比实验结果

Table 10 Comparison of experimental results among BMDSA, MSA and SLAD on different datasets

数据集	*Accuracy*		标准化互信息(NMI)	
	SLAD	SLAD (BMDSA+MSA)	SLAD	SLAD (BMDSA+MSA)
Ionosphere	85.14	86.7133	0.40	0.4087
PID	75.54	76.0169	0.15	0.1554

由表 7、表 8 可知,在 Connect、Pumsb 和 H 省真实人口数据 3 个数据集上,多尺度划分算法(MSA)和构建多尺度数据集算法(BMDSA)的运行时间较长,多尺度划分算法(MSA)的时间复杂度为 $O(m\log n)$,其中 m 为尺度层数,n 为最底层尺度划分块数。根据时间复杂性分析可知,尺度划分块数越少,多尺度划分算法(MSA)的运行时间越短,表征范围尺度的属性个数越少,越能有效减少基准尺度选择时间。

综合而言,本文提出的多尺度划分方法以及构建多尺度数据集方法具有较高的有效性与可行性。

结束语 本文总结了多尺度科学在数据挖掘领域的应用,分析并对比了多尺度数据挖掘中尺度划分的研究现状,并对目前多尺度划分方法进行了分析和完善。本文定义了尺度的含义,利用概率密度估计的离散化方法提出了多尺度划分算法,有效地扩展了尺度划分对象类型。根据多尺度化数据集模型定义多尺度化数据集方法,提出了构建多尺度数据集算法和基准尺度评分模型,为基准尺度的选择提供了一种有效的解决方法,削弱了多尺度数据挖掘中的尺度效应。本文通过实验验证了算法和模型的有效性和可行性。未来的工作中,应将多尺度划分方法更广泛地应用于数据挖掘研究领域,进一步研究构建多尺度数据集的方法,将一维尺度划分推广至多维尺度划分,并探索更加优秀的评价尺度方法。

参 考 文 献

[1] SUN Q X, LI M T, LU J X, et al. Scale of geospatial data and its research progress [J]. Geography and Geographic Information Science, 2007, 23(4): 53-56, 80. (in Chinese)
孙庆先,李茂堂,路京选,等. 地理空间数据的尺度问题及其研究进展[J]. 地理与地理信息科学,2007,23(4):53-56,80.

[2] LIU M M, ZHAO S L, HAN Y H, et al. Research on multi-scale data mining method [J]. Journal of Software, 2016, 27(12): 3030-3050. (in Chinese)
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.

[3] HAN Y H, ZHAO S L, LIU M M, et al. Multi-scale Clustering Mining Algorithm [J]. Computer Science, 2016, 43(8): 244-248. (in Chinese)

韩玉辉,赵书良,柳萌萌,等. 多尺度聚类挖掘算法[J]. 计算机科学,2016,43(8):244-248.

[4] LIU Q,HANG R,SONG H,et al. Learning Multi-Scale Deep Features for High-Resolution Satellite Image Classification[J]. IEEE Transactions on Geoscience & Remote Sensing,2016,PP(99):1-10.

[5] AZAMI H,FERNÁNDEZ A,ESCUDERO J. Refined multiscale fuzzy entropy based on standard deviation for biomedical signal analysis[J]. Medical & Biological Engineering & Computing,2017,55(11):2037-2052.

[6] LI Z,WEI Z,WEN C,et al. Detail-Enhanced Multi-Scale Exposure Fusion[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society,2017,26(3):1243-1252.

[7] SHEN L,SUN G,HUANG Q M,et al. Multi-Level Discriminative Dictionary Learning With Application to Large Scale Image Classification[J]. IEEE Transactions on Image Processing,2015,24(10):3109-3123.

[8] LIAO S,ZHU Q,QIAN Y,et al. Multi-granularity feature selection on cost-sensitive data with measurement errors and variable costs[OL]. https://www.onacademic.com/detail/journal_1000040426607310_1fb6.html.

[9] 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.

[10] LIU M M,ZHAO S L,CHEN M,et al. Scaling-up mining algorithm of multi-scale association rules mining [J]. Application Research of Computers,2015,32(10):2924-2929. (in Chinese)
柳萌萌,赵书良,陈敏,等. 多尺度关联规则挖掘的尺度上推算法[J]. 计算机应用研究,2015,32(10):2924-2929.

[11] LI C,ZHAO S L,ZHAO J P,et al. Scaling-up Algorithm of Multi-scale association rules [J]. Computer Science,2017,44(8):285-289. (in Chinese)
李超,赵书良,赵骏鹏,等. 多尺度关联规则尺度上推算法[J]. 计算机科学,2017,44(8):285-289.

[12] LI J X,ZHAO S L,AN L,et al. Scaling-up Algorithm of Multi-scale Classification Based on Fractal Theory[J]. Computer Science,2018,45(S1):453-459. (in Chinese)
李佳星,赵书良,安磊,等. 基于分形理论的多尺度分类尺度上推算法[J]. 计算机科学,2018,45(S1):453-459.

[13] LI J X,ZHAO S L,AN L,et al. Scaling-down Algorithm of Multi-scale Classification Based on Fractal Theory[J]. Application Research of Computers,2019(7):1-3. (in Chinese)
李佳星,赵书良,安磊,等. 基于广义分形插值理论的多尺度分类尺度下推算法[J]. 计算机应用研究,2019(7):1-3.

[14] PETRY F E,YAGER R R. Fuzzy Concept Hierarchies and Evidence Resolution[J]. IEEE Transactions on Fuzzy Systems,2014,22(5):1151-1161.

[15] KANG X,MIAO D. A study on information granularity in formal concept analysis based on concept-bases[J]. Knowledge-Based Systems,2016,105(C):147-159.

[16] HAO C,LI J,FAN M,et al. Optimal scale selection in dynamic multi-scale decision tables based on sequential three-way decisions[J]. Information Sciences,2017,415:213-232.

[17] ZHAO J P,ZHAO S L,LI C,et al. A multi-scale clustering algorithm based on grain calculation [J]. Application Research of Computers,2018,35(2):362-366. (in Chinese)
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.

[18] BIBA M,ESPOSITO F,FERILLI S,et al. Unsupervised discretization using kernel density estimation[C]// Proceedings of the International Joint Conference on Artificial Intelligence,Hyderabad,India,January. DBLP,2008:696-701.

[19] ZHOU C H,ZHANG J T. A geospatial data mining model based on information entropy [J]. Chinese Journal of Image and Graphics,1999,4(11):946-951. (in Chinese)
周成虎,张健挺. 基于信息熵的地学空间数据挖掘模型[J]. 中国图象图形学报,1999,4(11):946-951.

[20] GOU J,LIU J Y,WEI Z B,et al. Analysis of power energy flow complexity based on multi-scale entropy [J]. Acta Physica Sinica,2014(20):347-354. (in Chinese)
苟竞,刘俊勇,魏震波,等. 基于多尺度熵的电力能量流复杂性分析[J]. 物理学报,2014(20):347-354.

[21] BRUNI R,BIANCHI G. Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis [J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(9):2349-2361.