

# 多尺度聚类挖掘算法

韩玉辉 赵书良 柳萌萌 罗 燕 丁亚飞

(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)

(河北师范大学移动物联网研究院 石家庄 050024)

**摘 要** 数据挖掘领域在多尺度研究上已取得了一些进展。然而,当前研究主要集中于空间、图像数据的多尺度挖掘,并且传统的聚类挖掘并未对数据集的多尺度特性进行单独的研究。针对存在的问题,进行了普适性的多尺度聚类挖掘理论和方法的研究。首先,根据概念分层理论扩展尺度定义并构建多尺度数据集;其次,阐述尺度转换原因、分类,归纳多尺度聚类的定义;然后,以克里格法为理论基础,给出多尺度聚类尺度上推算算法 MSCSUA 和多尺度聚类尺度下推算算法 MSCSDA;最后,利用公用 UCI 聚类数据集和 H 省全员人口真实数据集对算法进行实验验证,结果表明 MSCSUA 和 MSCSDA 是有效、可行的。

**关键词** 多尺度,聚类,尺度转换,多尺度聚类挖掘,克里格法

**中图法分类号** TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.8.049

## Multi-scale Clustering Mining Algorithm

HAN Yu-hui ZHAO Shu-liang LIU Meng-meng LUO Yan DING Ya-fei

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

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

(Institute of Mobile Internet of Things, Hebei Normal University, Shijiazhuang 050024, China)

**Abstract** Data mining field has made some progress on the multi-scale research. However, the current research mostly focuses on the multi-scale mining of the space or image data. And traditional clustering mining has not separately studied the multi-scale characteristic of datasets. According to existing problems, this paper carried on the general study of multi-scale clustering mining theories and methods. Firstly, we extended scale definition on the basis of the concept hierarchy and built multi-scale datasets. Secondly, we expounded the reasons and classification of scale conversion, meanwhile concluded the definition of the multi-scale clustering. Then, we introduced multi-scale clustering scaling up algorithm and multi-scale clustering scaling down algorithm based on the kriging theories. Finally, simulation experiments tested MSCSUA and MSCSDA with the help of public UCI clustering datasets and demographic dataset from H province. And the experimental results show that MSCSUA and MSCSDA are effective and feasible.

**Keywords** Multi-scale, Clustering, Scale conversion, Multi-scale clustering mining, Kriging

## 1 引言

随着数学、物理学、地理学、化学等领域学者对多尺度问题的广泛研究,多尺度科学已成为一门独立的跨学科研究课题。国内外很多学者针对多尺度聚类进行研究,提出了很多新颖的多尺度聚类理论和方法。国内方面,文献[1]基于空间数据的多尺度特性,构建了多尺度空间数据挖掘框架,将多尺度分析思想引进到空间数据挖掘领域;文献[2]结合多尺度模糊聚类与粒子群优化构建方向梯度矢量流 DGVF 模型的初

始轮廓来分割颈动脉超声造影图像;文献[3]结合地学领域尺度知识和聚类挖掘,提出了基于加权向量提升的多尺度聚类算法,实现了尺度转换。国外方面,文献[4]使用数据流聚类算法来抽取多时间尺度上的趋势;文献[5]提出了用于不规则数据集的多尺度聚类算法,其在图像分割上具有不错的聚类效果;文献[6]结合 SOM 编码原始数据和多尺度谱聚类算法进行图像分割;文献[7]使用多尺度潜在狄利克雷分配模型解决了超高分辨率卫星全色图像的聚类问题。

目前研究成果显示,多尺度聚类研究实际上大多应用于

到稿日期:2015-06-12 返修日期:2015-08-21 本文受国家自然科学基金项目(71271067),国家社会科学基金项目(13BTY011),国家社科基金重大项目(13&ZD091),河北省高等学校科学技术研究项目(QN2014196),河北师范大学硕士基金(201402002)资助。

韩玉辉(1989—),男,硕士生,CCF 学生会会员,主要研究方向为数据挖掘、智能信息处理,E-mail: hanyuhui1989@sina.com;赵书良(1967—),男,博士,博士生导师,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);柳萌萌(1988—),女,硕士生,主要研究方向为数据挖掘、智能信息处理;罗 燕(1993—),女,硕士生,主要研究方向为数据挖掘、智能信息处理;丁亚飞(1988—),女,硕士生,主要研究方向为数据挖掘、智能信息处理。

空间、图像数据,这明显限制了多尺度科学在聚类技术上的应用与推广。为解决存在的问题,本文主要从以下几个方面展开:1)扩展尺度定义,构建多尺度数据集;2)详细介绍尺度转换原因和分类,概括多尺度聚类的定义;3)提出多尺度聚类尺度转换算法;4)实验验证算法的有效性和可行性。

2 多尺度聚类

2.1 尺度

最初尺度只包含空间尺度和时间尺度,然而随着多尺度研究的深入,我们发现原始尺度定义有一定的局限性。结合概念分层的理论,给出扩展尺度定义。

定义 1(概念分层, Concept Hierarchy) 概念分层指概念以偏序(partial order)的方式组成,定义了由底层概念(具体概念)到高层概念(一般概念)的映射序列<sup>[8]</sup>。

图 1 示出一个教育行政划分的概念分层结构。

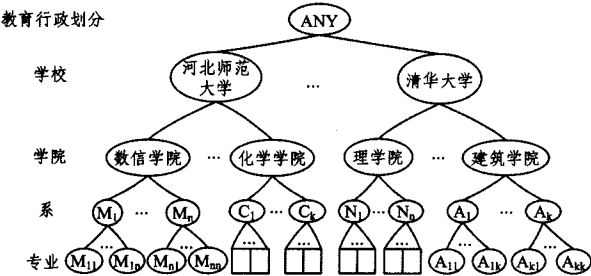


图 1 教育行政划分的概念分层结构

定义 2(尺度, Scale) 尺度指研究某一物体或现象时所采用的概念分层所用偏序关系中概念的范围大小单位。

其中,所有具有偏序包含关系的属性或维度都可认为具有尺度特性。

2.2 多尺度数据集

为了扩展多尺度聚类的适用范围,本文使用间隔尺度和名义尺度之间的转换来构建多尺度数据集。

定义 3(名义尺度, Nominal Scale) 名义尺度也称列名尺度或定类测量,它所测量的对象按质的特征进行分类。

定义 4(间隔尺度, Interval Scale) 间隔尺度也称计量水准或定距测量,在测量时用数字表示。

由于样本属性多,单一间隔尺度或名义尺度都无法完整描述特性。例如对表 1 样本进行聚类,若用间隔尺度,就会忽略性别、学历;若用名义尺度,就会忽略年龄、身高、体重。

表 1 间隔尺度与名义尺度示例

| 间隔尺度  |        |        | 名义尺度 |    |
|-------|--------|--------|------|----|
| 年龄(岁) | 身高(cm) | 体重(kg) | 性别   | 学历 |
| 35    | 173    | 68     | 男    | 高中 |
| 41    | 177    | 77     | 男    | 大学 |
| 37    | 183    | 67     | 女    | 初中 |

传统的聚类方法只能处理间隔尺度数据,本文通过将名义尺度转化为间隔尺度来解决此问题。针对表 1,对名义尺度进行量化处理,如表 2 所列。

表 2 名义尺度量化处理

| 指标 | 量化处理           |
|----|----------------|
| 性别 | 男:1,女:2        |
| 学历 | 初中:1,高中:2,大学:3 |

根据表 2 得出最终普适性的多尺度数据集,如表 3 所列。

表 3 名义尺度转换为间隔尺度后的数据集

| 间隔尺度  |        |        | 名义尺度 |    |
|-------|--------|--------|------|----|
| 年龄(岁) | 身高(cm) | 体重(kg) | 性别   | 学历 |
| 35    | 173    | 68     | 1    | 2  |
| 41    | 177    | 77     | 1    | 3  |
| 37    | 183    | 67     | 2    | 1  |

2.3 尺度转换

结合多尺度科学和数据挖掘知识,给出尺度转换定义。

定义 5(尺度转换, Scale Transformation) 尺度转换也称尺度推绎,是指利用数据挖掘技术将当前研究尺度获得的信息或知识,结合多尺度科学理论和方法,推演或归算到目标尺度的过程。

2.3.1 尺度转换原因

综合多尺度的实际研究,可概括尺度转换原因如下。

1) 尺度效应现象

尺度效应也称生态谬论,指在某一尺度下得出的结论不能无差别地适用于另一尺度的研究中。

2) 研究尺度与应用尺度不一致

在实际研究中,数据集本身的尺度特性往往是事先确定的,但研究尺度与最终应用尺度之间存在着不一致现象<sup>[9]</sup>。

3) 目标尺度样本过多,难以高效处理

当前研究所用数据集的容量日趋增大,想要直接高效地处理数据集是十分困难的。尺度转换可将数据集“分而治之”,然后推绎得到源数据集的知识。

2.3.2 尺度转换分类

按照转换的方向可分为尺度上推、尺度下推和尺度互推。其中,尺度互推是本文新增的尺度转换方向。

1) 尺度上推(Scaling-Up)

尺度上推指将概念分层中的低层次尺度知识推绎得到高层次尺度知识的过程,意味着将知识由微观趋向宏观。

2) 尺度下推(Scaling-Down)

尺度下推指将概念分层中的高层次尺度知识推绎得到低层次尺度知识的过程,意味着将知识由宏观趋向微观。

3) 尺度互推(Scaling-Mutual)

尺度互推指将概念分层中兄弟尺度间知识转化的过程,意味着尺度的大小不发生变化。

2.4 多尺度聚类

根据 2.1 节尺度相关知识和聚类挖掘过程,概括多尺度聚类定义如下。

定义 6(多尺度聚类, Multi-Scale Clustering) 多尺度聚类指利用尺度知识将原始数据构建多尺度数据集,使用或改进聚类方法挖掘基准尺度知识,并将基准尺度的聚类结果尺度转换到目标尺度的过程。

3 多尺度聚类尺度上推算法 MSCSUA

多尺度聚类尺度上推算法 MSCSUA(Multi-Scale Clustering Scaling UP Algorithm)的实现核心在于尺度转换过程中如何保留宏观特征而忽略微观特征。目前尺度上推已有很多理论和方法,而其中最为成熟和完善的是地统计学中的块状克里格法 BK(Block Kriging)。

3.1 块状克里格法 BK

克里格法是建立在变异函数理论及结构分析的基础上,

在有限区域内对区域化变量取值的线性无偏最优估计。

克里格线性估计量如式(1)所示:

$$Z_0^*(x) = \sum_{i=1}^n \lambda_i Z(x_i) \quad (1)$$

其中,  $Z_0^*(x)$  为待估点,  $\lambda_i$  为各个样本点的权重系数,  $Z(x_i)$  为样本点数据。克里格法的重点在于各样本点权重系数  $\lambda_i$  的确定, 克里格方程矩阵表示如下:

$$\lambda = K^{-1} D \quad (2)$$

其中:

$$K = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \quad (3)$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ 1 \end{bmatrix}, D = \begin{bmatrix} c(x_1, x) \\ c(x_2, x) \\ \vdots \\ c(x_n, x) \\ 1 \end{bmatrix} \quad (4)$$

式(3)和式(4)中,  $C_{ij}$  为第  $i$  个已知样本和第  $j$  个已知样本的协方差,  $C(x_i, x)$  为待估点  $x$  和已知样点  $x_i$  的协方差。

本文利用块状克里格法尺度上推, 由于数据结构不变, 式(2)中克里格矩阵  $K$  不变, 矩阵  $D$  改变。

$$D = \begin{bmatrix} \bar{c}(x_1, V) \\ \bar{c}(x_2, V) \\ \vdots \\ \bar{c}(x_n, V) \\ 1 \end{bmatrix} \quad (5)$$

其中,  $\bar{C}(x_i, V)$  为离散化块段中心点的平均协方差函数。每个块段  $V$  又被离散成 4 个小块段(见图 2), 即通过临近已知样点来估计, 以估计点  $x_0$  作为中心块段的平均值, 实现尺度一致, 减小检验误差。

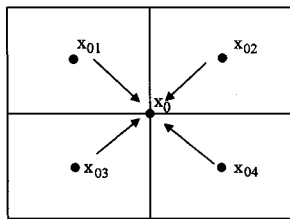


图2 待估块  $x_0$  的离散化示意图

已知样本点  $x_i$  与待估块  $V$  中 4 个离散化中心点的平均协方差函数如式(6)所示:

$$\bar{C}(x_i, V) = \frac{1}{4} (C_{01-i} + C_{02-i} + C_{03-i} + C_{04-i}) \quad (6)$$

### 3.2 MSCSUA 算法实现

MSCSUA 的基本思想是: 首先将原始数据集通过间隔尺度和名义尺度转换, 形成多尺度数据集, 并在此基础上使用 PCA 选择最佳属性; 其次, 在此属性上选取基准尺度, 并使用已有的聚类算法(如 K-means, EM 等)在基准尺度上挖掘聚类结果; 然后, 将基准尺度结果当作已知样本点, 利用块状克里格法获得目标尺度与样本点间的权重关系; 最终, 加权平均得到目标尺度聚类结果。

MSCSUA 具体步骤如下。

INPUT: Original Datasets

OUTPUT: The Clustering results of Large Scale

1. Convert Interval Scale and Nominal Scale
2. Build Multi-scale Datasets
3. Use PCA to Choose the Best attribute
4. Select the Basic Scale:  $BS_0$
5. Make Clusters on  $BS_0$
6. Use BK to Obtain Weight coefficients
7. Scaling-Up the Target Scale:  $TS_0$

## 4 多尺度聚类尺度下推算法 MSCSDA

多尺度聚类尺度下推算法 MSCSDA (Multi-Scale Clustering Scaling Down Algorithm) 的核心在于如何扩展微观特征而忽略宏观特征。尺度下推是从宏观到微观的过程, 这个过程需要更多的数据信息。在实际研究中, 面尺度数据拥有的信息是有限的, 仅靠单一面尺度数据尺度下推, 其结果是不可信的。

### 4.1 回归面到点克里格法 ATPRK

普通的 ATPK (Area To Point Kriging) 是用已知面尺度数据对未知点进行插值的尺度下推方法, 基本原理为未知点值为其所在面以及邻近面的数据线性加权和。具体如图 3 所示。

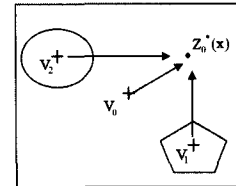


图3 ATPK 示意图

估计方程组及方差分别如下:

$$Z_0^*(x) = \sum_{i=1}^K \lambda_i(v_i) Z(v_i) \quad (7)$$

$$\begin{cases} \sum_{j=1}^K \lambda_j(v_j) \bar{C}(v_i, v_j) + \mu_i = \bar{C}(v_i, x), i=1, 2, \dots, K \\ \sum_{j=1}^K \lambda_j(v_j) = 1 \end{cases} \quad (8)$$

其中,  $Z_0^*(x)$  为待估点估计,  $v_i$  和  $v_j$  为已知面,  $Z(v_i)$  为已知面的值,  $\lambda_i$  为权重,  $\mu_i$  为拉格朗日乘子。

相比普通克里格, ATPK 公式中用面点协方差  $\bar{C}(v_i, x)$  和面面协方差  $\bar{C}(v_i, v_j)$  替换点点协方差。

面点协方差  $\bar{C}(v_i, x)$ 、面面协方差  $\bar{C}(v_i, v_j)$  的计算公式如下所示。

$$\bar{C}(v_i, x) = \frac{1}{|v_i|} \int_{s \in v_i} \text{Cov}\{z(s), z(x)\} \quad (9)$$

$$\bar{C}(v_i, v_j) = \frac{1}{|v_i|} \frac{1}{|v_j|} \int_{s \in v_i} \int_{s' \in v_j} \text{Cov}\{z(s), z(s')\} \quad (10)$$

ATPK 进行尺度下推时不仅考虑数据间的关系, 还给出了估计精度。而回归面到点克里格 ATPRK (Area To Point Regression Kriging) 是普通 ATPK 的改进, 主要步骤如下。

步骤 1 利用已知面数据提取相关的辅助数据;

步骤 2 以已知面为因变量, 辅助数据为自变量, 建立回

归方程,得到回归系数和面尺度残差;

步骤3 根据辅助数据和回归系数确定目标变量趋势;

步骤4 使用 ATPK 对面尺度残差尺度下推;

步骤5 合并步骤3和步骤4即得到 ATPRK 尺度下推结果。

4.2 MSCSDA 算法实现

MSCSDA 的基本思想大致与 MSCSUA 相似,只是在尺度转换部分不同。尺度下推首先根据已知聚类结果获取辅助数据,并构建两者之间的回归方程;其次,利用 ATPK 尺度下推已知结果得到结果  $R_1$ ;然后,辅助数据根据回归系数尺度下推得到结果  $R_2$ ;最后合并  $R_1$  和  $R_2$  得到目标尺度聚类结果。

MSCSDA 具体步骤如下。

INPUT:Original Datasets

OUTPUT:The Clustering results of Small Scale

- 1. Convert Interval Scale and Nominal Scale
- 2. Build Multi-scale Datasets
- 3. Use PCA to Choose the Best attribute
- 4. Select the Basic Scale;BS<sub>0</sub>
- 5. Make Clusters on BS<sub>0</sub>
- 6. Obtain Auxiliary data
- 7. Construct the Regression Equation
- 8. Scaling-Down results; $R_1$  with ATPK
- 9. Use Auxiliary data and Coefficients to  $R_2$
- 10. Merge  $R_1$  and  $R_2$  to TS<sub>0</sub>

5 实验分析

为了验证 MSCMA 的有效性及其他性能,本文利用具有不同实际应用的数据集进行验证。实验环境为 Windows 10 Pro Insider Preview 系统,主频 3.40GHz,4GB RAM,算法采用 JAVA 语言进行具体实现。实验设计思路如下:首先利用 K-Means,EM,FarthestFirst(以下简称 FF)等聚类算法直接在数据集目标尺度上进行聚类挖掘;然后使用本文算法进行尺度变换得到目标尺度聚类结果;最后利用聚类评价指标对比本文算法和其他算法,以验证本文算法的有效性。

5.1 UCI 数据集

本文选取 UCI 公用数据集 Iris, Wine, Segmentation, Waveform, Magic gamma 和真实数据即 H 省全员数据共 6 组数据集进行测试。表 4 列出实验数据集的相关信息,包括数据集名称、样本数目、属性数、类别数和数据集编号等信息。

表 4 UCI 数据集相关信息

| 数据集          | 样本数    | 属性数 | 类别数 | 编号 |
|--------------|--------|-----|-----|----|
| Iris         | 150    | 4   | 3   | D1 |
| Wine         | 178    | 14  | 3   | D2 |
| Segmentation | 2310   | 19  | 7   | D3 |
| Waveform     | 5000   | 21  | 3   | D4 |
| Magic gamma  | 19020  | 10  | 2   | D5 |
| H 省全员数据      | 140663 | 7   | 8   | D6 |

5.2 聚类评价指标

本文采用 CA(Clustering Accuracy),F1-Measure 和运行时间等 3 种聚类评价指标对比实验结果。下面将详细介绍 CA 和 F1-Measure 评价指标。

5.2.1 CA

CA 为聚类正确划分样本所占的比率,定义为式(11):

$$CA=\sum_{l=1}^{\zeta}n_l/N$$
 (11)

其中, $n_l$  为样本中被正确划分为类标  $l$  的样本个数, $N$  为数据集样本总数。

5.2.2 F1-Measure

F1-Measure 是信息检索、人工智能、数据挖掘等领域的重要指标,其值越大表明聚类效果越好。

$$F_1=\frac{2PR}{P+R}$$
 (12)

其中, $P$  为查准率(Precision), $R$  为查全率(Recall); $P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN}$ 。

$TP,FP$  和  $FN$  的值由表 5 混淆矩阵(Confusion Matrix)确定。

表 5 混淆矩阵

|      | 预测正类 | 预测负类 |
|------|------|------|
| 实际正类 | TP   | FN   |
| 实际负类 | FP   | TN   |

5.3 实验结果分析

本文主要根据 CA,F1-Measure 和运行时间 3 个指标对不同聚类算法及本文算法在尺度上推和尺度下推两个转换方向上进行具体的实验对比和分析。

5.3.1 尺度上推实验分析

从表 6 可以看出,MSCSUA 算法除了在 Waveform 数据集上的 CA 度量指标略低于 K-means 外,在其他数据集上均优于其他算法。通过计算不同算法的 CA 平均值可知,MSCSUA 要优于其他算法,在平均性能上较 FF 有 7%的提升。

表 6 尺度上推——不同算法的 CA 值比较

| 数据集          | K-means | EM   | FF   | MSCSUA |
|--------------|---------|------|------|--------|
| Iris         | 0.89    | 0.91 | 0.85 | 0.93   |
| Wine         | 0.98    | 0.99 | 0.96 | 1.00   |
| Segmentation | 0.92    | 0.85 | 0.84 | 0.95   |
| Waveform     | 0.93    | 0.88 | 0.87 | 0.92   |
| Magic gamma  | 0.91    | 0.86 | 0.85 | 0.92   |
| H 省全员数据      | 0.93    | 0.92 | 0.91 | 0.94   |
| 平均值          | 0.93    | 0.90 | 0.88 | 0.94   |

根据 F1-Measure 的定义,如果算法具有较大的 F1 值,则说明  $P$  和  $R$  均为较大值,即查准率和查全率较高。从表 7 可以看出,MSCSUA 在测试数据集上均具有较高的 F1 值,只是在 Waveform 数据集上其值略低。

表 7 尺度上推——不同算法的 F1-Measure 值比较

| 数据集          | K-means | EM    | FF    | MSCSUA |
|--------------|---------|-------|-------|--------|
| Iris         | 0.874   | 0.906 | 0.842 | 0.927  |
| Wine         | 0.972   | 0.989 | 0.956 | 1.000  |
| Segmentation | 0.903   | 0.864 | 0.837 | 0.944  |
| Waveform     | 0.918   | 0.861 | 0.857 | 0.902  |
| Magic gamma  | 0.892   | 0.874 | 0.863 | 0.906  |
| H 省全员数据      | 0.915   | 0.903 | 0.897 | 0.925  |
| 平均值          | 0.912   | 0.900 | 0.875 | 0.934  |

通过表 8 可以看到,当数据量增大时,EM 算法的运行时间也成倍增加;MSCSUA 在运行时间上略优于 FF,但参考表 6 和表 7,FF 的 CA 值和 F1 值均较低。

表8 尺度上推——不同算法的运行时间(s)比较

| 数据集          | K-means | EM    | FF   | MSCSUA |
|--------------|---------|-------|------|--------|
| Iris         | 0.02    | 0.09  | 0.03 | 0.01   |
| Wine         | 0.03    | 0.10  | 0.04 | 0.02   |
| Segmentation | 0.22    | 1.03  | 0.07 | 0.05   |
| Waveform     | 0.28    | 1.22  | 0.09 | 0.07   |
| Magic gamma  | 0.45    | 2.63  | 0.11 | 0.13   |
| H省全员数据       | 1.36    | 20.35 | 0.27 | 0.24   |
| 平均值          | 0.39    | 4.24  | 0.10 | 0.09   |

### 5.3.2 尺度下推实验分析

从表9可以看出,使用ATPRK方法的MSCSDA在正确率CA上具有明显的优势。CA平均值相对于K-means算法提高了7%的性能,相对EM算法提高了5.7%的性能,相对FF算法提高了13.6%的性能。

表9 尺度下推——不同算法的CA值比较

| 数据集          | K-means | EM   | FF   | MSCSDA |
|--------------|---------|------|------|--------|
| Iris         | 0.88    | 0.90 | 0.80 | 0.92   |
| Wine         | 0.85    | 0.89 | 0.77 | 0.90   |
| Segmentation | 0.83    | 0.85 | 0.83 | 0.91   |
| Waveform     | 0.85    | 0.86 | 0.82 | 0.89   |
| Magic gamma  | 0.89    | 0.84 | 0.84 | 0.94   |
| H省全员数据       | 0.86    | 0.87 | 0.82 | 0.95   |
| 平均值          | 0.86    | 0.87 | 0.81 | 0.92   |

尺度下推相较于尺度上推在查全率和查准率上有明显的降低。通过表10可以看出,其他算法的查全率和查准率下降较多,而本文MSCSDA算法在F1-Measure的度量上依然可以维持在0.9以上,可见该算法充分考虑了辅助数据的作用。

表10 尺度下推——不同算法的F1-Measure值比较

| 数据集          | K-means | EM    | FF    | MSCSDA |
|--------------|---------|-------|-------|--------|
| Iris         | 0.863   | 0.892 | 0.763 | 0.911  |
| Wine         | 0.851   | 0.887 | 0.758 | 0.897  |
| Segmentation | 0.816   | 0.846 | 0.814 | 0.903  |
| Waveform     | 0.847   | 0.870 | 0.806 | 0.884  |
| Magic gamma  | 0.881   | 0.823 | 0.833 | 0.922  |
| H省全员数据       | 0.862   | 0.874 | 0.796 | 0.937  |
| 平均值          | 0.853   | 0.865 | 0.795 | 0.909  |

通过表11可以看出,在运行效率上,FF算法具有很好的表现,MSCSDA略差于FF。但同时也发现FF的CA和F1度量指标是4个算法中最差的,MSCSDA因计算量相对较大,运行时间有所增加,但准确率很高,真正做到了时间和准确率的均衡。

表11 尺度下推——不同算法的运行时间(s)比较

| 数据集          | K-means | EM    | FF   | MSCSDA |
|--------------|---------|-------|------|--------|
| Iris         | 0.01    | 0.07  | 0.01 | 0.05   |
| Wine         | 0.03    | 0.14  | 0.02 | 0.07   |
| Segmentation | 0.17    | 1.03  | 0.13 | 0.16   |
| Waveform     | 0.26    | 1.26  | 0.19 | 0.88   |
| Magic gamma  | 1.42    | 4.63  | 1.13 | 1.03   |
| H省全员数据       | 3.36    | 16.35 | 1.87 | 2.71   |
| 平均值          | 0.88    | 3.91  | 0.56 | 0.82   |

以上实验结果表明,在多尺度聚类挖掘中,在相同条件下,利用BK和ATPRK进行尺度上推和尺度下推的准确率高,执行速度快,显著提高了聚类挖掘的性能。

**结束语** 本文针对多尺度科学多应用于空间、图像数据的局限性现状,对多尺度科学在聚类挖掘领域应用的理论和

方法进行了分析和研究,提出了一种新的基于克里格法的多尺度聚类尺度转换算法。利用概念分层对扩展尺度定义,归纳总结尺度转换的原因、分类等内容;在给定多尺度聚类定义的基础上,详细介绍了克里格的相关理论。最后,通过实验验证了MSCSUA和MSCSDA的有效性和可行性。在下一步工作中,我们将研究如何结合多尺度科学和聚类挖掘评估知识,提出适用于多尺度聚类的评价方法;将多尺度聚类进行可视化表示;研究关联规则、分类、预测等其他多尺度数据挖掘方法。

## 参考文献

- [1] Sun Q X, Chen Q P, Fang T, et al. Multi-scale Spatial Data Mining Model Based on Fuzzy Clustering and its Application in Mine [J]. Journal of Shanghai Jiaotong University, 2008, 42(2): 194-197, 201 (in Chinese)  
孙庆先, 陈秋平, 方涛, 等. 基于模糊聚类的多尺度空间数据挖掘模型及其矿山应用[J]. 上海交通大学学报, 2008, 42(2): 194-197, 201
- [2] Zhang Q, Huang C C, Han H, et al. CEUS Image Segmentation of Carotid Arteries Using Multi-scale Fuzzy Clustering and DGVF Model [J]. Journal of Shanghai University (Natural Science), 2014, 20(5): 633-644 (in Chinese)  
张麒, 黄春春, 韩红, 等. 基于多尺度模糊聚类与DGVF模型分割颈动脉超声造影图像[J]. 上海大学学报(自然科学版), 2014, 20(5): 633-644
- [3] Su D H, Zhao S L, Liu M M, et al. Weight Vector Based Multi-scale Clustering Algorithm [J]. Computer Science, 2015, 42(4): 263-267 (in Chinese)  
苏东海, 赵书良, 柳萌萌, 等. 基于加权向量提升的多尺度聚类挖掘算法[J]. 计算机科学, 2015, 42(4): 263-267
- [4] Andreea B D. Stock Data Clustering and Multi-scale Trend Detection [J]. Methodology and Computing in Applied Probability, 2012, 14(1): 87-105
- [5] Tao G, Yong L Y, Tao X. An Online Multi-scale Clustering Algorithm for Irregular Data Sets [C] // 2011 International Conference on Future Computer Sciences and Application. 2011: 209-211
- [6] Ying D, Tao G, Lei L. Self-Organizing Map Based Multi-scale Spectral Clustering for Image Segmentation [C] // 2012 International Conference on Computer Science and Electronics Engineering. 2012: 329-333
- [7] Hong T, Li S, Yin F Q, et al. A Multi-scale Latent Dirichlet Allocation Model for Object-Oriented Clustering of VHR Panchromatic Satellite Images [J]. IEEE Transaction on Geoscience and Remote Sensing, 2013, 51(3): 1680-1692
- [8] Jin S N. Research on the Knowledge Discovery in Conceptual Hierarchy Knowledge Base Based on the Multiple-Level Association Rules [D]. Tianjin: Tianjin University, 2006 (in Chinese)  
金胜男. 基于多层关联规则的概念分层和知识库中知识发现的研究[D]. 天津: 天津大学, 2006
- [9] Alex J, Henri P, Philipp S, et al. Multi-scale Classification of Remote Sensing Images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2012, 50(10): 3764-3775