

基于加权向量提升的多尺度聚类挖掘算法

苏东海 赵书良 柳萌萌 苏嘉庚 李 妍

(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 多尺度聚类挖掘在指导人们进行多尺度决策方面有着不可取代的作用,然而传统的多尺度聚类挖掘算法有一个致命的弱点,即需要在每个用户感兴趣的尺度上应用聚类挖掘算法。为克服此缺陷,定义了一种将数据的多尺度特性进行向量化的方法;结合地学领域的尺度转换机制,提出了一种新的知识的多尺度转换机制——基于加权向量提升的多尺度聚类挖掘算法 WVB-MSCA(Weight Vector Based Multi-scale Clustering Algorithm)。算法首先在选定的基准尺度上进行聚类挖掘,获取聚类结果,并借助尺度转换机制将基准尺度的聚类结果反演到其它感兴趣的尺度上。实验表明,算法 WVB-MSCA 是可行且有效的。

关键词 多尺度,尺度转换,聚类挖掘,多尺度聚类

中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.4.054

Weight Vector Based Multi-scale Clustering Algorithm

SU Dong-hai ZHAO Shu-liang LIU Meng-meng SU Jia-geng LI Yan

(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 Multi-scale clustering plays an important role in multi-scale decision making, which cannot be replaced, while almost of the traditional multi-scale clustering algorithms have encountered the same deadly problem that clustering algorithms are applied on every scale which every user is interested in. To overcome the problem occurred, this paper gave a method of vectorization for the multi-scale characteristic in data and proposed a new multi-scale conversion mechanism of knowledge named WVB-MSCA(Weight Vector Based Multi-Scale Clustering Algorithm) with the help of scale conversion mechanism in earth science. WVB-MSCA applies clustering algorithm on the basic scale selected in advance, and inverses the knowledge gained from basic scale to other scales which users are interested in. Experimental results show that WVB-MSCA is feasible and effective.

Keywords Multi-scale, Scale conversion, Clustering data mining, Multi-scale clustering

各行各业针对不同的应用所做出的不同决策都受到尺度的影响,具有明显的尺度特性。如对某一地区的植被覆盖情况进行研究时所采用的多尺度影响数据,地图导航应用中的城镇、街区、街道所表现出的多尺度特点,政府相关部门针对不同的群体实施的决策,某大型超市对历年商品销量进行分析时所遇到的时间上的尺度问题等都具有多尺度特性。聚类挖掘作为数据挖掘系统中的一个重要模块,既可以单独作为数据库中发现数据分布规律的开发工具,也可以为其他数据挖掘分析提供较好的预处理工作。因此,如何针对研究对象的多尺度特性进行相应的聚类挖掘,对于指导人们进行多尺度决策有着不可替代的作用。

目前,多尺度聚类挖掘研究已有一定的发展,人们提出了很多多尺度聚类挖掘算法和应用。然而,这些多尺度聚类挖掘算法均需要在人们感兴趣的所有尺度上应用聚类算法,从而造成了资源的浪费和聚类时间的增加,常常不能及时地为人们制定决策提供可靠的信息,造成很多不必要的损失。

本文提出一种新型算法来解决传统的多尺度聚类挖掘中所遇到的问题,实现在某单一尺度上进行聚类,再将此聚类结果或知识规则转换到其它尺度上。该算法避免了在所有尺度上聚类的问题,提高了聚类效率,且可得到在自己感兴趣尺度上的聚类结果,为人们进行多尺度决策提供指导。

到稿日期:2014-05-28 返修日期:2014-09-08 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)资助。

苏东海(1988—),男,硕士生,主要研究方向为数据挖掘、智能信息处理,E-mail:d_h_sue@ymail.com;赵书良(1964—),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;苏嘉庚(1986—),男,硕士生,主要研究方向为数据挖掘、智能信息处理;李妍(1988—),女,硕士生,主要研究方向为数据挖掘、智能信息处理。

1 相关工作

1.1 多尺度数据挖掘

近些年,随着数据挖掘技术在其它科学领域的普遍应用和多尺度科学的兴起,多尺度数据挖掘受到了各界学者的广泛关注。Tsoukatos 等人基于 DFS 搜索思想提出了 DFS_MINE 方法,其可挖掘数据库中的多尺度频繁模型^[1];Cao 等人对串树结构和 Apriori 算法进行改进,将序列变成子序列,从而挖掘出多尺度频繁模式^[2];Lee 等人通过扫描多尺度数据生成映像图和轨道信息表,以 DFS 搜索方式进行频繁模式和多尺度事件的挖掘^[3];地学领域也做出了一些多尺度空间数据挖掘的研究,文献^[4]对地学领域尺度转换的一般理论与方法进行了总结,并把多尺度空间数据挖掘归为 3 种可行的途径:

(1)在挖掘前将单一尺度的数据转换成多个尺度的数据,然后对多个尺度的数据分别进行挖掘,即实现数据的多尺度转换;

(2)在挖掘算法中引入调节尺度的操作部件,以控制挖掘出知识的尺度;

(3)将挖掘出的单一尺度的知识转换成多个尺度的知识,即完成知识的多尺度转换。

多尺度空间数据挖掘的 3 种途径也同样适用于数据挖掘领域的聚类挖掘。目前通过前两种途径实现的多尺度聚类挖掘已取得了很好的研究成果,而如何通过第 3 种途径实现多尺度聚类挖掘还很少涉足。

1.2 传统多尺度聚类挖掘算法

多尺度聚类挖掘旨在实现在每一个用户感兴趣的尺度上获取聚类结果,可用于天气预测、交通拥挤预测、动物迁移分析、移动计算和异常点分析等方面。

Gaffney 等人提出了一种基于回归混合模式概率的算法,实现了对轨迹数据的多尺度聚类,并利用 EM 算法来确定模式中簇的隶属关系^[5];Chudova 等人将对象的时间和空间尺度方面的轨迹漂移作为混合模型的参数来研究数据的多尺度聚类^[6];Nanni 等人基于时态语义和时间聚焦的方法提出了一种轨迹聚类算法,其核心思想是使用一种基于密度的聚类算法对轨迹进行聚类^[7];更具代表性的多尺度聚类挖掘算法是施培蓓提出的多尺度谱聚类算法,该算法通过引入控制尺度的参数来实现在各个尺度上进行聚类挖掘^[8],充分地展现了多尺度数据挖掘的前两种途径各自的特点。

不难看出,多尺度数据挖掘的前两种途径实现的多尺度聚类算法都需要在每一个可检测或用户感兴趣的尺度上应用聚类挖掘算法,这类方法简单易行,却存在着一个致命的弱点,即其所需要的时间复杂度和空间复杂度往往让人难以忍受。通过对多尺度数据挖掘的 3 种可行途径进行分析,发现第三种途径只需要在特定的尺度上应用更少次甚至一次聚类算法,再将该尺度上的聚类结果反演到其它感兴趣的尺度上,这恰恰克服了前两种途径存在的问题。而这也是本文提出基于加权向量的多尺度聚类挖掘算法的基本思想,即实现知识的多尺度转换。

2 WVB-MSCA 算法思想及相关定义

2.1 多尺度数据中尺度的定义及其向量表示

2.1.1 尺度的相关定义

定义 1(尺度,Scale) 是指对某一事物或现象 O (Object)

进行研究时所采用的可度量维度信息,记作 S 。某一事物或现象所有的互不相同的尺度 $S_1, S_2, \dots, S_n$ 构成的集合 $\{S_1, S_2, \dots, S_n\}$ 称为研究 O 时的尺度集,记作 Ψ 。

定义 2(大,小尺度) 设 S, S' 是研究某一事物或现象 O 时所采用的两个不同的尺度,若尺度 S 所表示的维度信息比尺度 S' 大,则称尺度 S 比尺度 S' 大,记作 $S >_O S'$ ($S <_O S'$),而尺度 S' 比尺度 S 小,记作 $S <_O S'$ 。基于这两种尺度进行研究时的数据称为大尺度数据和小尺度数据。

定义 3(基准尺度,Basic Scale) 设尺度 S 是研究某一事物或现象 O 时所采用的尺度,在对事物或现象的其它任意尺度 S' 进行研究时,若可以将尺度 S' 表示成尺度 S 的某种线性或非线性函数,则尺度 S 称为研究 O 时的基准尺度,记作 BS_O 。

定义 4(尺度转换,Scale Convert) 设 S, S' 是研究某一事物或现象 O 时所采用的尺度,且尺度 S 比尺度 S' 大,即有 $S >_O S'$ 成立,将从尺度 S 到尺度 S' 所做的转换过程称作尺度转换,记作 $S \rightarrow S'$ 。由大尺度 S 到小尺度 S' 所做的尺度转换称为尺度下推,记作 $S \rightarrow_{\downarrow} S'$;反之,由小尺度 S' 到大尺度 S 所做的尺度转换称为尺度上推,记作 $S' \rightarrow_{\uparrow} S$ 。

定义 5(多尺度聚类,Multi-Scale Clustering) 设 $\Psi = \{S_1, S_2, \dots, S_n\}$ 是研究某一事物或现象 O 时所采用的尺度集,借助数据挖掘技术在尺度集 Ψ 中的每一个尺度 S_i ($i=0, 1, \dots, n$) 上实现聚类的过程称作多尺度聚类。

2.1.2 尺度的向量表示

定义 6(尺度的向量表示,Vectorization of Scale) 将研究某一事物或现象 O 时所采用的尺度集 Ψ 以 N 维空间向量的形式表示的过程称为尺度的向量表示。

尺度的向量表示分为两种形式:空间尺度的向量表示(Vectorization of Scale in Space)和时间尺度的向量表示(Vectorization of Scale in Time)。这两种尺度向量表示的基本思想是将需要表示的尺度进行二进制化,并以二进制向量的形式表示出来。经过向量表示的尺度有以下好处:进行尺度间大小比较时可以转而比较相应向量的维数大小。

2.2 WVB-MSCA 算法

2.2.1 WVB-MSCA 算法描述

本文基于基准尺度与尺度转换机制,借助于 N 维空间向量的知识提出了基于加权向量提升的多尺度聚类挖掘算法 WVB-MSCA (Weight Vector Based Multi-Scale Clustering Algorithm),该算法的基本思想是:首先选择基准尺度 BS_O ,并在此基准尺度上应用特定的聚类挖掘算法获取聚类结果;其次,对于其它任意的尺度 S ,通过加权向量提升的方法,借助于尺度转换机制将基准尺度 BS_O 上的聚类结果或知识反演到尺度 S 上,并最终实现多尺度聚类挖掘的目的。该算法的具体步骤如下:

INPUT:具有多尺度特性的数据研究对象

OUTPUT:用户感兴趣的多尺度上的聚类结果

1. Choose the Basic Scale of O : BS_O

2. Make clusters of O on the scale BS_O

3. For the other Scale S of O , which is not BS_O

4. If $S <_O BS_O$

5. $SD(S <_O BS_O)$

6. Inverse cluster of O on scale S

7. Else if $S >_O BS_O$

8. $SU(S >_0 BS_0)$
9. Inverse cluster of O on scale S
10. Exit until there is no any scale to cluster

对于在该算法的第3行,如果尺度 S 比基准尺度 BS_0 更小,则调用向下的尺度转换方法,通过基准尺度 BS_0 下推出尺度 S ,进而反演出尺度 S 上的聚类结果;如果尺度 S 比基准尺度 BS_0 更大,则调用向上的尺度转换方法,通过基准尺度 BS_0 上推出尺度 S ,进而反演出尺度 S 上的聚类结果;若用户感兴趣的尺度均已聚类完毕(第10行),则算法推出并返回各个尺度上的聚类结果。如图1所示,本文WVB-MSCA算法共分为3步:基准尺度 BS_0 的选取、应用聚类挖掘算法及挖掘结果的反演。

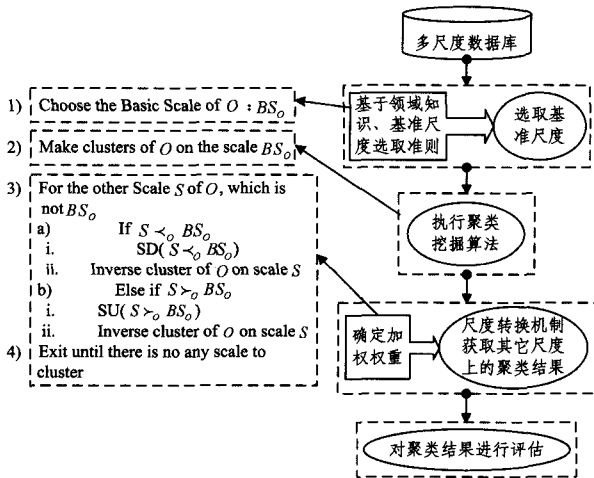


图1 基于加权向量提升的多尺度聚类挖掘算法

2.2.2 算法中需要解决的两个问题

在算法WVB-MSCA中,有两个重要的问题需要解决,即基准尺度的选取和尺度转换方法的选择。其中基准尺度的选取决定了整个算法的运行效率和结果,因此有必要对基准尺度的选择准则进行研究,在进行基准尺度选取时,除了要考虑领域知识以外,还应当遵循以下3个准则^[9]。

准则1(距离最大准则) 选取基准尺度时,若所选择的基准尺度使得在所有目标尺度上聚类结果的距离和达到最大,则该基准尺度即为最佳基准尺度。两个基准尺度间的距离度量定义如下:

$$\Delta d(BS_i, BS_{i+1}) = \sum_{s_i \in \Psi_{BS_i}, s_j \in \Psi_{BS_j}} [f(s_i) - f(s_j)]^2 \quad (1)$$

其中, f 是基准尺度在各个目标尺度上产生的聚类结果。式(1)遵循了聚类挖掘的定义和评价标准,即非同类对象间尽可能相异,同类对象间尽可能相似。

准则2(能量最大准则) 对于各个类簇,当它们具有相同的均值或均值较为接近时,其特征向量的平方和便提供了一种分类信息;不同类别的特征向量之间的能量相差越大,则类别的可分性就越好。因此可以定义如下的能量差异函数,选取某一基准尺度使得以其为参数的能量差异最大,则该基准尺度便为最佳的基准尺度。

$$\Delta E(BS_i, BS_{i+1}) = \sum_{s_i \in \Psi_{BS_i}, s_j \in \Psi_{BS_j}} |f(s_i)^2 - f(s_j)^2| \quad (2)$$

其中, f 是基准尺度在各个目标尺度上所具有的能量。

准则3(相关系数最小准则) 若所选取的基准尺度使得研究对象在所有目标尺度上的聚类结果的相关系数最小,则这一基准尺度便为最佳的基准尺度。相关系数最小准则本质

上是用相关系数代替通常含义下聚类标准中的相异性,即类间对象相关性系数尽可能小,而类内对象相关性系数尽可能大。该准则也是聚类挖掘标准评价的另一种表述。

选择好基准尺度后,需借助尺度转换机制将基准尺度上的聚类结果反演到目标尺度上,故所选择的尺度转换方法关系着知识转换的合理性和输出结果的优劣。尺度转换最初出现于地学领域,因此地学领域的尺度转换理论与方法相对成熟和完善。本文选择地统计学领域的面域普通克里格法作为尺度上推转换方法的理论依据,选择点普通克里格法作为尺度下推转换方法的理论依据。

克里格法是建立在变异函数理论和结构分析基础上的空间局部估计方法,是在有限区域内对区域化变量聚集的一种无偏最优估计,该方法首先定义一个线性估计量:

$$Z_0^\#(x) = \sum_{i=1}^{n+1} \lambda_i Z(x_i) \quad (3)$$

其中, $Z(x_i)$ 为样点数据,而 $Z_0^\#(x)$ 为待估计量, λ_i 为各个样点的权重,即克里格系数,且 $\sum_{i=1}^{n+1} \lambda_i = 1$ 。对于任意一个估计量,实际值和估计值之间都是存在偏差的,而此处的 $Z_0^\#(x)$ 实际上是 $Z_0(x)$ 真实值的一种线性无偏最优估计,式(3)为克里格方程。克里格系数 λ 可以表示成如下矩阵乘积形式:

$$\lambda = K^{-1} D \quad (4)$$

其中

$$K = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \quad (5)$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -j \end{bmatrix}, D = \begin{bmatrix} c(x_1, x) \\ c(x_2, x) \\ \vdots \\ c(x_n, x) \\ 1 \end{bmatrix} \quad (6)$$

其中, c_{ij} 是第 i 个已知样本和第 j 个已知样本的协方差,而 $c(x_i, x)$ 为待估计点和第 i 个已知样本之间的协方差。

使用面域普通克里格法实现尺度上推,关键需要确定克里格系数 λ ,方法如下:由源尺度 S' 中各个样点间的协方差确定 K ;由源尺度 S' 中各个样点对目标尺度 S 中待估计点的影响因子确定 D ,这里我们称此影响因子为占比信息 (Accounting Info) 或影响信息 (Influence Info) inf 。即矩阵 K 中的 c_{ij} 为源尺度 S' 中样点 i 与样点 j 之间的协方差,而 D 中 $c(x_i, x)$ 被替换为源尺度 S' 中样点 i 对目标尺度 S 中待估计点 0 的影响因子 inf_{i0} 。用面域克里格法确定克里格系数的公式如下:

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -u \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} inf_{10} \\ inf_{20} \\ \vdots \\ inf_{n0} \\ 1 \end{bmatrix} \quad (7)$$

使用点普通克里格法实现尺度下推时,矩阵 K 的确定方法与尺度上推的面域普通克里格法相同,即 c_{ij} 为源尺度 S' 中样点 i 与样点 j 之间的协方差,而 $c(x_i, x)$ 的含义与传统克里格法一致,即源尺度 S' 中样点 i 与目标尺度 S 中待估计点之间的协方差, λ 的具体计算公式如下:

$$\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -v \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \times \begin{bmatrix} c(x_1, x) \\ c(x_2, x) \\ \vdots \\ c(x_n, x) \\ 1 \end{bmatrix} \quad (8)$$

算法 WVB-MSCA 基于克里格法的基本思想,使用式(7)和式(8),将源尺度 S' 的聚类结果反演到目标尺度 S 上,从而实现了聚类结果的跨尺度转换。

2.2.3 WVB-MSCA 同其它多尺度聚类挖掘算法的比较

目前多尺度聚类挖掘的大多算法都是基于多尺度数据挖掘 3 种模式中的前两种,如文献[8]在研究多尺度谱聚类算法时在算法运行时引入了控制尺度的参数以引导挖掘结果的多尺度性;而更多普遍的多尺度聚类挖掘算法是先将挖掘对象进行多尺度化,再在每一个尺度数据上应用聚类挖掘算法,这些算法的一个共同特性就是需要在每一个尺度上应用聚类算法。由于进行了多次挖掘,这些算法效率普遍不高,并且不能有效利用已有尺度上的聚类结果。而本文提出的算法是先在某一尺度上进行聚类,并将该尺度的聚类结果通过尺度转换机制转换到其它尺度上,有效利用了已有聚类结果,不必进行多次挖掘,算法效率明显提高。

3 实验及分析

本文将提出的算法 WVB-MSCA 应用于 H 省¹⁾流动人口犯罪分析中,以验证该算法的可行性及有效性。通过该算法在 H 省流动人口犯罪分析中的应用,可以更好地为政府部门制定相关决策提供有力的科学依据和判断标准。

3.1 H 省流动人口的多尺度聚类挖掘模式

明确定义多尺度聚类挖掘对象,认清多尺度聚类挖掘的目的是多尺度聚类挖掘的第一步。对目前 H 省流动人口犯罪特点及成因进行分析可以发现,对该省流动人口何时迁入、何时迁出、年龄特征、原户籍地、现居住地、教育背景、工作性质等相关信息进行合理地分析为更好地实施流动人口管理决策有着决定性的作用。

应用本文提出的算法 WVB-MSCA 对 H 省流动人口进行分析,通过对流动人口的性别、年龄、教育程度、居住地、管理地、迁入迁出时间等众多属性进行多尺度聚类挖掘,分别在村、乡、县、市等 4 个行政单位尺度上将该省 6 万多个村,2500 多个乡,199 个县和 11 个市聚成 3 类,即 3 种不同的流动人口犯罪监管级别,以便为政府部门进行相关的多尺度决策提供科学的指导依据。

3.2 实验流程

在此实验中,分别以村级单位、乡级单位、县级单位为基准尺度,选择前文描述的基于划分的 K-Means 和 K-Medoids 算法作为流动人口分析的聚类算法,在选定的基准尺度上进行聚类挖掘,并在此聚类结果的基础上反演出其它尺度上的聚类结果。将基于本文算法 WVB-MSCA 得到的聚类结果同在该尺度上直接应用聚类挖掘算法得到的聚类结果进行对比,以算法 WVB-MSCA 正确聚类的类簇个数同聚类结果的总类簇个数之比作为相符度,来衡量算法的准确率。主要的

实验流程如图 2 所示。

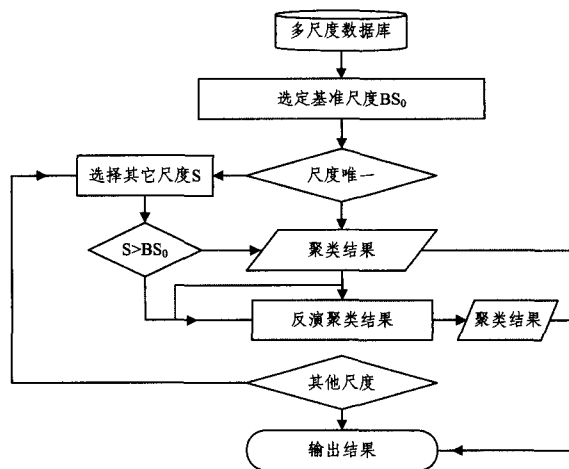


图 2 WVB-MSCA 算法在 H 省流动人口分析中的应用流程

3.3 结果分析

实验一:由县级尺度上推市级尺度

在本实验中,以县级单位为基准尺度,应用算法 WVB-MSCA 在 199 个县级单位上应用聚类挖掘算法,获取县级单位尺度的聚类结果,再反演出 11 个市级单位的聚类结果,并将该结果同直接在市级单位上应用聚类挖掘算法所得到的结果进行比较,结果如图 3 和图 4 所示。从聚类结果可以看出,在市级单位尺度上直接应用聚类算法对 11 个市进行聚类时,有 1 个城市属于 C1 类,6 个属于 C2 类,4 个属于 C3 类(见图 3);而 WVB-MSCA 算法以县级单位为基准尺度进行聚类时,将 C2 类中的一个城市聚类到了 C3 类,将 C3 类中的一个城市聚类到了 C2 类(见图 4)。这里采用市级直接聚类结果为基准,WVB-MSCA 算法的结果同直接在市级单位进行聚类的结果相符度为 81.82%,说明算法 WVB-MSCA 在市级和县级尺度上进行聚类结果反演的效果较好,准确率较高,能从县级尺度得到市级尺度的大部分正确聚类结果。

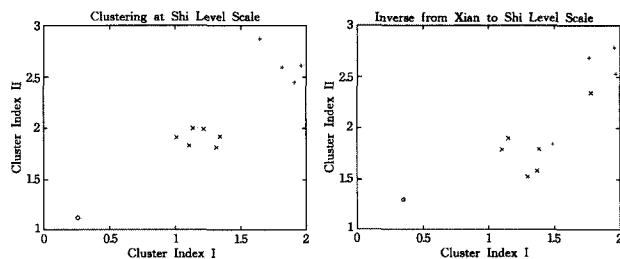


图 3 直接在市级单位尺度上进行 图 4 从县级单位上推市级单位聚类的结果

实验二:由乡级尺度上推县级尺度

在本实验中,以乡级单位为基准尺度,应用本文提出的算法 WVB-MSCA 在 2554 个乡级单位应用聚类挖掘算法,获取乡级单位尺度的聚类结果,再反演出 199 个县级单位的聚类结果,并同直接在县级单位的聚类结果进行比较,结果如图 5、图 6 所示。算法 WVB-MSCA 在县级和乡级尺度上进行聚类结果反演的相符度为 68.34%,与实验一结果进行对比,可以看出,随着聚类对象数目的增加,算法的准确率明显下降。乡级对象数量较多导致克里格方程中以各个乡级对象作为样

¹⁾ 应用户的要求,将省名隐去,而叫做 H 省。

点对县级聚类结果的影响更加分散,局部小误差的积累导致较大的误差,致使算法的准确率下降。

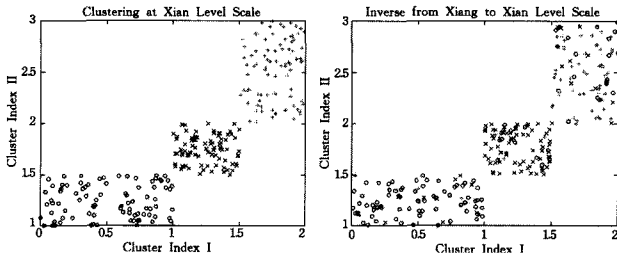


图5 直接在县级单位尺度上进行的聚类结果

实验三:由乡级尺度上推市级尺度

在本实验中,同样以乡级单位为基准尺度,应用 WVB-MSCA 算法在 2554 个乡级单位应用聚类挖掘算法,获取乡级单位尺度的聚类结果,不同的是在本实验中跨过县级尺度,算法直接反演出 11 个市级单位的聚类结果(见图 7),并将该结果同直接在市级单位上应用聚类挖掘算法所得到的结果(见图 3)进行比较,计算相符度,结果为 45.45%。算法准确率进一步下降,由乡级尺度的聚类结果直接反演市级尺度的聚类结果,隐藏了对中间尺度——县级尺度的利用,算法实际上进行了两次尺度转换操作。第一次为乡级尺度到县级尺度的聚类结果反演,得到中间结果,由实验二可知,聚类对象数量较多,影响较分散,导致中间结果的相符度为 68.34%,误差较大,使用此中间结果再对市级尺度对象进行聚类结果反演导致第一次尺度转换操作中的误差在第二次尺度转换时进一步放大,算法准确度下降。

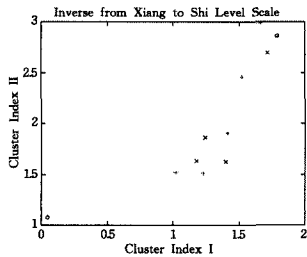


图7 从乡级单位上推市级单位的聚类结果

3.4 结论

通过对以上实验数据进行分析,可以得出以下结论:

1)WVB-MSCA 算法是可行和有效的。实验一和实验二的结果表明,由小尺度对象的聚类结果向上推导所得的大尺度对象的聚类结果在很大程度上与大尺度对象的直接聚类结果相同;实验一和实验二的结果对比说明基准尺度中聚类对象的个数同算法的准确度成反比,即聚类对象越多,算法准确度越低,由对象增多使克里格方程中小尺度样点对大尺度目标点的影响分散、误差积累所致。当基准尺度的聚类对象较少时,WVB-MSCA 算法能确保较高的准确度,效果较好,是可行且有效的。

2)WVB-MSCA 算法存在缺陷和不足。如 1)所述,算法在处理较多数据对象时准确度不高,基于克里格法的尺度转换机制仍需完善和加强;实验一、二与实验三的结果对比表明,算法在处理多层次尺度转换问题方面表现欠佳,即从小尺

度对象到大尺度对象进行聚类结果转换时,单次的转换效果要比多次的转换效果更好,这是由于多层次尺度转换时每一次转换操作的误差会逐层放大,导致最终结果的准确度较低。故 WVB-MSCA 算法在修正多层次尺度转换误差方面仍需改进和加强。

3)从总体上看,算法只在单一尺度上进行挖掘,利用得到的聚类结果,反演推算出其他尺度上的聚类结果,从而规避了对不同尺度的数据进行多次聚类,在牺牲一定精确度的情况下取得了效率的提高;并且,算法在处理相邻尺度数据聚类时表现出较高的精确度,因此算法在处理可容忍一定误差的相邻尺度聚类任务时由于其高效性,较其他多尺度聚类挖掘算法有较大的优势。

结束语 本文以多尺度聚类挖掘的第三种途径——知识的多尺度转换为基本思想,借助地统计学中的克里格法,提出了多尺度聚类挖掘算法 WVB-MSCA。算法通过单一尺度的聚类结果推导多个尺度的聚类结果,实现了聚类挖掘中知识的多尺度转换。最后通过具体实验,将本文的算法应用于 H 省流动人口犯罪分析当中,验证了本文算法的可行性和有效性。下一步工作中,我们将致力于解决 3 方面问题:改进具体的尺度转换机制,提高算法在单次尺度转换中的准确度;修正和减少单次尺度转换中的误差影响,优化算法在多层尺度转换中的表现;完善基准尺度的选择标准,在算法上实现基准尺度的自动选取。

参考文献

- [1] Tsoukatos I, Gunopulos D. Efficient mining of spatiotemporal patterns[C]// 7th International Symposium, SSTD 2001, Berlin; Springer Berlin Heidelberg, 2001:425-442
- [2] Cao Hui-ping, Mamoulis N, DW C. Mining frequent spatiotemporal sequential patterns[C]// Proceeding of Fifth IEEE International Conference on Data Mining, NJ; IEEE, 2005:82-89
- [3] Lee A, T J, Chen Y A, et al. Mining frequent trajectory patterns in spatial-temporal databases[J]. Information Sciences, 2009, 179(13):2218-2231
- [4] 孙庆先,方涛,郭达志. 空间数据挖掘中的尺度转换研究[J]. 计算机工程与应用, 2005, 16(3):17-20
- [5] Gaffney S, Smyth P. Trajectory clustering with mixtures of regression models[C]// Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York; ACM, 1999:63-72
- [6] Chudova D, Gaffney S, Mjolsness E, et al. Translation-invariant mixture models for cover clustering[C]// Proceeding of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington D C; ACM, 2003:79-88
- [7] Nanni M, Pedreschi D. Time-focused clustering of trajectories of moving objects[J]. Journal of Intelligent Information Systems, 2006, 27(3):267-289
- [8] 施培蓓,郭玉堂,胡玉娟,等. 多尺度的谱聚类算法[J]. 计算机工程与应用, 2011, 47(8):128-130
- [9] 赵学智,叶邦彦,陈统坚,等. 小波变换的尺度自适应选择[J]. 数据采集与处理, 2004, 19(1):46-50