软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(12):3030−3050 [doi: 10.13328/j.cnki.jos.004924]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

多尺度数据挖掘方法*

柳萌萌[1,2], 赵书良[1,2], 韩玉辉[1,2], 苏东海[3], 李晓超[1,2], 陈 敏[1,2]

[1](河北师范大学 数学与信息科学学院,河北 石家庄 050024)

[2](河北省计算数学与应用重点实验室(河北师范大学),河北 石家庄 050024)

[3](冀广传媒集团 河北广电无限传媒有限公司,河北 石家庄 050000)

通讯作者: 赵书良, E-mail: zhaoshuliang@sina.com

摘 要: 多尺度理论已被引入到数据挖掘领域,但人们对其研究仍不够深入和完善,缺乏普适性理论与方法.随着大数据处理应用的不断深入,其研究变得更加迫切.针对上述问题,进行了普适的多尺度数据挖掘理论和方法的研究.首先,基于概念分层理论给出了数据尺度划分和数据尺度的定义以及多尺度数据集之间的上下层尺度数据集关系;其次,阐明了多尺度数据挖掘的定义、研究实质和方法分类;最后,提出了多尺度数据挖掘算法框架,给出其理论基础,并将此框架应用于关联规则挖掘,提出了多尺度关联规则挖掘算法 MSARMA(multi-scale association rules mining algorithm),实现了多尺度数据集之间知识的跨尺度推导.利用 IBM T10I4D100K 数据集和 H 省全员人口真实数据集对 MSARMA 算法进行了实验和分析,实验结果表明:算法具有较高的覆盖率、精确度和较低的支持度估计误差,是可行且有效的.

关键词: 多尺度;频繁项集;关联规则;尺度转换;多尺度关联规则挖掘

中图法分类号: TP182

中文引用格式: 柳萌萌,赵书良,韩玉辉,苏东海,李晓超,陈敏.多尺度数据挖掘方法.软件学报,2016,27(12):3030−3050. http://www. jos.org.cn/1000-9825/4924.htm

英文引用格式: Liu MM, Zhao SL, Han YH, Su DH, Li XC, Chen M. Research on multi-scale data mining method. Ruan Jian Xue Bao/Journal of Software, 2016,27(12):3030−3050 (in Chinese). http://www.jos.org.cn/1000-9825/4924.htm

Research on Multi-Scale Data Mining Method

LIU Meng-Meng[1,2], ZHAO Shu-Liang[1,2], HAN Yu-Hui[1,2], SU Dong-Hai[3], LI Xiao-Chao[1,2], CHEN Min[1,2]

[1](Mathematics & Information Science, Hebei Normal University, Shijiazhuang 050024, China)

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

[3](Grand Media Group, Hebei Broadcasting Wireless Media Co. Ltd., Shijiazhuang 050000, China)

Abstract: Many researches of data mining have paid close attention to multi-scale theory. However the study of multi-scale data mining still comes short on universal theories and approaches. To overcome this limitation, this paper conducts a study of universal multi-scale data mining on theoretical and methodological aspect. First, the paper lays out the definition of data-scale-partition and data-scale based on concept hierarchy, and characterizes the relationship of upper-layer and lower-layer datasets between multi-scale datasets. Next, it illustrates the definition and essence of multi-scale data mining, and presents the classification of multi-scale data mining methods. Finally, it introduces the algorithm framework and its theoretical basis of multi-scale data mining, and proposes an algorithm named MSARMA (multi-scale association rules mining algorithm) to realize the transition of knowledge in multi-scale data expressions. Experiments are

* 基金项目: 国家自然科学基金(71271067); 国家社会科学基金(13BTY011, 13&ZD091)

Foundation item: National Natural Science Foundation of China (71271067); National Social Science Foundation of China (13BTY011, 13&ZD091)

收稿时间: 2015-02-12; 修改时间: 2015-05-11, 2015-09-10, 2015-09-25; 采用时间: 2015-10-08; jos 在线出版时间: 2015-11-27

CNKI 网络优先出版: 2015-11-26 16:06:08, http://www.cnki.net/kcms/detail/11.2560.TP.20151126.1606.002.html

carried out to test MSARMA with the help of IBM T10I4D100K dataset and demographic dataset from H province, and the results indicate that MSARMA is effective and feasible with better coverage rate, better accuracy and lower average support error .

Key words: multi-scale; frequent item-set; association rule; scale conversion; multi-scale association rules mining

多尺度现象普遍存在于客观世界中,近年来受到了学术界的广泛关注,并逐渐发展为一门独立的研究课题——多尺度科学.数学、物理学、化学等领域的学者已将多尺度理论引入到本学科中进行了一系列相关研究[1].数据融合[2]技术和关联模型[3]应用的快速发展,很大程度上促进了多尺度领域的研究,多信息源的采集、传输、综合、过滤、关联及合成极大地降低了尺度转换的时间消耗,提高了数据结果精度.针对数据挖掘领域的多尺度研究,国内方面:文献[4]利用空间数据固有的多尺度特性及概念层次关系,提出了点对象的多尺度空间关联规则挖掘算法,实现了空间数据关联规则的尺度上推;文献[5]结合地学领域的尺度转换机制,提出了基于加权向量提升的多尺度聚类挖掘算法;文献[6]采用分级的多尺度组合分类方法,提出了一种自适应多尺度分割的组合分类算法,减少了训练时间,得到了一个性能更好的分类器;文献[7]针对现有自相似网络的长相关性不稳定问题,以离散小波变换(DWT)多尺度分析为依据建立网络流模型并进行流量预测.国外方面:文献[8]基于最小平均峰度的快速小波变换 WLMK,考虑输入流量和小波变换的冗余特征,对自相似网络流量进行了多尺度分析和预测;文献[9]将数据进行四元数表示,基于改进的马氏距离度量进行快速多尺度聚类,实现了电子背向散射衍射EBSD数据分割模型;文献[10]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;文献[11]提出了一种迭代交互式层次多尺度分类器,实现了遥感图像的多尺度分割,相比一般分割算法,提高了分类精度;文献[12]基于条件随机域 CRF(conditional random field),实现了不同时期、分辨率下,地理光学遥感图像的多时间、空间尺度的上下文分类.随着大数据时代的到来,有关海量数据处理应用的研究越来越深入,且基本致力于探究新技术、新方法在效率和准确度上实现跨越.多尺度数据挖掘研究刚刚起步,探索能够处理海量数据的多尺度数据挖掘方法以支持多尺度决策,提高效率和准确度显得尤为重要和迫切.

从目前研究来看,多尺度数据挖掘的研究并未完全展开,所涉及的数据类型多集中于空间数据,这无疑限制了多尺度数据挖掘的发展空间.为了拓展多尺度理论在数据挖掘领域的研究范围与深度,本文从以下 3 方面进行探讨:(1) 将多尺度科学的基本理论引入到一般数据集,提出更为普适性的多尺度数据理论;(2) 阐述多尺度数据挖掘的定义、实质和分类;(3) 提出多尺度数据挖掘算法框架并给出其理论基础,然后将此算法框架应用到关联规则挖掘中,进一步提出多尺度关联规则挖掘算法,实现数据中隐含关联规则的跨尺度推导,将单一尺度数据集上的知识通过上推和下推,转换为其他尺度层面上的知识,实现一次挖掘,即可得到多层次知识的目的.同时,对算法的准确性从理论上进行评估.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)真实人口数据集和IBM T10I4D100K数据集进行实验,证明算法的可行性和效率.

1 数据的多尺度

广义上讲,尺度是研究对象的单位或测量工具.针对以数据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.与统计学中测量尺度以变量为基准衡量数据类似,数据挖掘中的尺度测量基准应为数据的固有属性.当研究者就某一范畴考察数据时,往往对应于数据在该范畴的一个属性集,此属性集通常可以形成偏序结构明确的概念分层,依据概念分层中相关概念对数据进行划分,可以形成具有多尺度特性的数据集.

1.1 概念分层

定义 1(概念分层). 概念分层 H 是一个偏序关系集$(H,\prec)$,其中,H 为有限概念集,$\prec$表示 H 所包含概念之间的一种偏序关系[13].

数据某一范畴的属性集可形成偏序结构明确的概念分层:属性集$H=\{h_1,\ldots,h_i,\ldots,h_n\}$中所有属性$h_i(i=1,\ldots,n)$均可作为有限概念集中的概念;各属性间可依领域知识形成某种偏序关系,对应有限概念集中概念间的偏序关系;某一属性$h_i\in H$实例化后对应若干具体属性值,记为$V_{h_i}=\{v_1,v_2,...,v_{m_i}\}$($v_j$表示具体的离散值或连续区间),这些

属性值在语义上的高度抽象形成了属性(概念)h_i,称属性值 $v_j \in V_{h_i}(j=1,...,m_i)$ 在语义上属于 h_i,记作 $v_j \overset{s}{\in} h_i$.实际应用中,地域范畴的属性集可形成概念分层:($H_{location}$,$\prec$)={村$\prec$乡$\prec$县$\prec$市$\prec$省};时间范畴的属性集亦可形成概念分层:(H_{time},$\prec$)={日$\prec$月$\prec$年}.

1.2 数据尺度的基本概念

若概念分层(H,$\prec$)中,偏序关系"$\prec$"表示概念涉及范围、粒度的相对大小,或者时间幅度的相对长短,那么称此概念分层(H,$\prec$)具有多尺度特性.例如,概念分层($H_{location}$,$\prec$)={村$\prec$乡$\prec$县$\prec$市$\prec$省}表示了地域范围的由小到大;(H_{time},$\prec$)={日$\prec$月$\prec$年}表示了时间幅度的由短到长.这些概念分层都有能力表示所衡量数据粒度的相对大小,故它们都具备多尺度特性.以具备多尺度特性的概念分层为标准进行数据集的划分,可以形成多尺度数据集.下面给出数据尺度划分的定义:

定义 2(数据尺度划分). 数据集 DS 某一范畴的属性集形成具备多尺度特性的概念分层:(H,$\prec$)={$h_1\prec...\prec h_i\prec...\prec h_n$},有限概念集 H 中某概念 $h_i(i=1,\ldots,n)$的属性值集为 $V_{h_i}=\{v_1^i,...,v_{m_i}^i\}$,依据概念 h_i 的不同属性值 v_j^i ($j=1,\ldots,m_i$)对 DS 进行划分,具有相同属性值的数据形成独立的子数据集,记为 $ds_{h_i-v_j^i}$;数据集 DS 被划分为 m_i 个子数据集,形成以概念 h_i 为划分粒度的一组数据集,此过程称为以概念 h_i 为基准的数据尺度划分.

数据尺度划分过程中,尺度本质上是具有一定语义的度量单位,并且其语义与具备多尺度特性的概念分层中的概念相关.在此,给出数据尺度的定义:

定义 3(数据尺度). 数据集 DS 在数据尺度划分中所依据的概念 h_i 为该数据集在本次划分中的数据尺度,记作 $S_{DS}=h_i$.

定义 4(基准尺度数据集). 数据集 DS 以概念分层(H,$\prec$)中的概念 h_i 为数据尺度进行数据尺度划分的结果中,所有子数据集 $ds_{h_i-v_j^i}(i=1,2,...,n;j=1,2,...,m_i)$ 为数据集 DS 在数据尺度 $S_{DS}=h_i$ 下的元尺度数据集.若其他尺度数据集可以由该元尺度数据集合并或分解得到,那么该元尺度数据集称为基准尺度数据集,与基准尺度数据集对应的概念 h_i 为基准尺度,记为 BS.

一般地,在概念上某数据尺度 $S_{DS}=h_i$ 的所有元尺度数据集可简记为 ds_{h_i},以表示该数据尺度划分的划分粒度为 h_i.将概念分层中不同的概念作为数据尺度对同一数据集进行多粒度、多层次划分,可得到多分辨率、多尺度的数据集.以多尺度或多粒度的方式考察数据可以得到数据在不同尺度和不同层面下隐含的知识,从而达到从多个尺度分析数据的目的.本文提出的数据尺度、多尺度数据集的相关概念在表象上与数据立方体[14]类似,但本文中多尺度数据集同数据立方体在本质上的关注点不同:前者关注的是具有多尺度特性的数据集本身,后者则关注聚集值.

1.3 多尺度数据集之间的关系

定义 5(上层尺度数据集和下层尺度数据集). 数据集 DS 分别取 $S_{DS}=h_x$ 和 $S'_{DS}=h_y$ 进行数据尺度划分,得到的数据集分别为 ds_{h_x} 和 ds_{h_y},若 $h_x\prec h_y$,则数据尺度 $S'_{DS}=h_y$ 下的所有元尺度数据集 ds_{h_y} 为数据尺度 $S_{DS}=h_x$ 下的所有元尺度数据集 ds_{h_x} 的上层尺度数据集,数据尺度 $S_{DS}=h_x$ 下的所有元尺度数据集 ds_{h_x} 为数据尺度 $S'_{DS}=h_y$ 下的所有元尺度数据集 ds_{h_y} 的下层尺度数据集.

上、下层尺度数据集是相对的概念:上层尺度数据集的数据尺度概念层级较高,划分粒度较大,数据集所包含的数据含义更为概括泛化;下层尺度数据集数据尺度的概念层级较低,划分粒度较小,故其包含的数据含义更为具体明确.上层尺度数据集显然比下层尺度数据集包含了更大范围的数据,故上层尺度数据集与其下层尺度数据集相比为大尺度数据集;相应的,下层尺度数据集较其上层尺度数据集为小尺度数据集.大尺度数据集和小尺度数据集对应的数据尺度分别称为大尺度和小尺度,也是相对的概念.若数据尺度 $S_{DS}=h_x$ 和 $S'_{DS}=h_y$ 在概念分层中 $h_x\prec h_y$,则 h_x 为小尺度,h_y 为大尺度,记作 $S_{DS}\prec S'_{DS}$.

2 多尺度数据挖掘

2.1 多尺度数据挖掘的定义

多尺度数据挖掘的主要任务有两方面,即,数据的多尺度实现和知识的多尺度发掘:前者属于数据预处理,利用数据尺度划分即可实现;后者则需要改进具体的挖掘技术,在数据多个尺度的表现形式中发掘知识,分析、推导知识间的相互联系.在此给出多尺度数据挖掘的定义:

定义 6(多尺度数据挖掘). 多尺度数据挖掘指对数据进行多尺度处理,形成多尺度数据集,使用或改进数据挖掘技术发掘多尺度数据集中的隐含知识,并分析、推导数据不同尺度表现形式背后隐含的知识间相互关系的过程.

2.2 多尺度数据挖掘的实质

尺度转换是多尺度科学研究的核心内容,文献[5]曾针对空间数据挖掘提出 3 种尺度转换的途径,概括起来包括两方面内容,即,数据的多尺度转换和知识的多尺度转换.实际上,多尺度空间数据挖掘的研究方式同样适用于一般的多尺度数据挖掘研究.不难看出:数据的多尺度转换实现原理简单,但需要对数据的多尺度表现形式分别进行挖掘,工作量大;知识的多尺度转换工作量小,只需挖掘单一尺度的数据,但需要解决尺度效应问题,原理复杂.尺度效应是指某一尺度上得出的结论不能无差别地适用于另一尺度[15],即:当研究对象改变其表现范围、粒度或幅度时,分析结果也会随之变化[16].同理,在某一尺度数据集中挖掘得到的知识不可能无差别地适用于其他尺度的数据,必须利用领域知识或多尺度数据集之间的关系对挖掘结果进行推演、归算,才可能实现真正意义上的知识的多尺度转换.很明显,知识的多尺度转换是多尺度数据挖掘的实质.下面给出知识尺度转换的定义:

定义 7(知识尺度转换). 设 S,S'是数据集 DS 的数据尺度,将由尺度 S 下数据集的知识推导尺度 S'下数据集知识的过程称作知识尺度转换,记作 $S\xrightarrow{K}S'$.若尺度 S 较尺度 S'大,即 $S\succ S'$,则由大尺度 S 到小尺度 S'所做的知识的尺度转换称为尺度下推,简记为 $S\rightarrow\downarrow S'$;反之,由小尺度 S'到大尺度 S 所做的知识的尺度转换称为尺度上推,简记作 $S'\rightarrow\uparrow S$.

由于实际应用中收集到的数据通常只表现为单一尺度,并且没有必要将数据处理为多个尺度的表现形式分别进行挖掘,故本文将以知识的多尺度转换作为多尺度数据挖掘的研究实质:研究方法或算法,对单一尺度数据集的知识进行推导、归算,得到在一定误差范围内适于其他尺度数据集的知识,而不对这些尺度的数据集进行直接挖掘[17].

2.3 多尺度数据挖掘算法分类

依据知识尺度转换的定义和地学领域的尺度转换分类[5],从知识的多尺度转换方向角度,将多尺度数据挖掘算法分为尺度上推挖掘算法和尺度下推挖掘算法两种:

1) 尺度上推挖掘算法(scaling-up mining algorithm):利用从下层尺度数据集中得到的知识、领域知识以及上、下层尺度数据集之间的关系,推导上层尺度数据集中隐含的知识,而不对上层尺度数据集进行直接挖掘.尺度上推挖掘算法旨在利用数据的微观片面的知识推导数据的宏观全面的知识;
2) 尺度下推挖掘算法(scaling-down mining algorithm):利用从上层尺度数据集中得到的知识、领域知识以及上、下层尺度数据集之间的关系,推导下层尺度数据集中隐含的知识,而不对下层尺度数据集进行直接挖掘.尺度下推挖掘算法旨在利用数据的宏观全面的知识推导数据的细节局部的知识.

3 多尺度数据挖掘算法

3.1 算法框架

本文基于基准尺度与尺度转换机制,提出了多尺度数据挖掘算法 MSDMA(multi-scale data mining algorithm)的基本框架.该框架的基本思想是:首先选择基准尺度 BS,在基准尺度数据集上应用数据挖掘算法得

到挖掘结果;然后,对于其他任意目标尺度 S_O,通过运用尺度转换机制和转换方法将基准尺度 BS 上的挖掘结果或知识反演到目标尺度 S_O 上,直接得到目标尺度数据集背后隐含知识的近似结果,而不对目标尺度数据集进行直接挖掘,最终实现多尺度数据挖掘的目的.该算法框架的具体步骤如下:

算法 1. 多尺度数据挖掘算法 MSDMA 框架.

INPUT:具有多尺度特性的数据集 DS;

OUTPUT:用户感兴趣的目标尺度 S_O 上的数据挖掘结果.

1. Choose the Basic Scale of DS: BS
2. Apply data mining algorithm on datasets of DS on the scale BS
3. While the scale S_O of DS, which is not BS
4. If $S_O \prec BS$
5. Execute Scaling-Down Mining Algorithm: $BS \rightarrow \downarrow S_O$
6. //Deduce mining result of S_O based on the result of scale BS
7. Else if $S_O \succ BS$
8. Execute Scaling-Up Mining Algorithm: $BS \rightarrow \uparrow S_O$
9. //Deduce mining result of S_O based on the result of scale BS
10. Exit until there is no any scale to mine

算法中:如果目标尺度 S_O 比基准尺度 BS 小,则调用尺度下推挖掘算法,进行知识的向下尺度转换,通过基准尺度 BS 进行尺度下推挖掘,进而反演出尺度 S_O 上的挖掘结果;如果尺度 S_O 比基准尺度 BS 大,则调用尺度上推挖掘算法,进行知识的向上尺度转换,通过基准尺度 BS 进行尺度上推挖掘,进而反演出尺度 S_O 上的挖掘结果;若用户感兴趣的尺度均已挖掘完毕,则算法结束并返回各个尺度上的挖掘结果.

3.2 理论基础

多尺度数据挖掘的任务是,利用尺度转换机制将基准尺度数据集上的挖掘结果反演到其他目标尺度数据集上.本节将详细介绍尺度转换机制,即,知识尺度转换的理论基础.

尺度转换方法中,最常用的方法是基于空间统计学的地统计法[18].地统计法属于一种优化估计技术,其基本假设是建立在空间相关性的先验模型之上的.主要思想为:距离较近的采样点的相关参数比距离较远的采样点的相关参数更为相似,而其相似程度或空间协方差的大小,则可以通过比较被特定滞后距离分隔的同一随机变量的不同值及多个尺度上对区域化随机变量的变异性的度量来确定,故而成为尺度转换的一种良好解决方案.

3.2.1 克里格法可用于一般数据集进行多尺度数据挖掘的实质

克里格插值法[18]是基于采样数据反映的区域化变量的结构信息(变异函数和协方差函数提供),根据待估点或块段的有限邻域内的采样点数据,考虑样本点的空间相互位置关系、与待估点的空间位置关系,对待估点进行的一种无偏最优估计,并且能给出估计精度.由于研究目的和条件的不同,相继产生了各种各样的克里格法:在满足二阶平稳(或本征)假设时,可用普通克里格法;在非平稳现象中,可用泛克里格法;在计算可采储量时,要用非线性估计量,就可用析取克里格法;当区域化变量服从对数正态分布时,可用对数克里格法;当数据较少、分布不大规则、对估计精度要求不高时,可用随机克里格法等.

克里格法的实质是待估点的估值,是由临近点的加权求和,因此核心在于加权系数λ的确定.空间数据是指用来表示空间实体的位置、形状、大小及其分布特征诸多方面信息的数据,它具有时间尺度、空间尺度关系等特性.数据之间存在空间相互位置关系,并且有一定的分布规律.因此,使用克里格法进行空间数据的尺度上推、下推是非常合适的.

对于一般数据集,从理论上来说并不具备明显的空间、时间尺度特性,因为大多数用于分类、聚类、关联规则或者其他数据挖掘方法的数据集中的属性值没有空间数据的大小、位置信息.传统意义的尺度是指在研究某一物体或现象时所采用的空间或时间单位,然而随着研究的深入,我们发现,原始的尺度定义不够准确或不够

全面.本文提出使用概念分层区分大小尺度,实际上扩展了尺度概念,即,概念分层中具有偏序包含的概念都可认为是尺度,例如学校的行政组成结构(学校、学院、专业、系、班级)等.根据概念分层知识,我们可以认为,任意数据集都是具有数学偏序关系的多尺度数据集.其中,最特殊的情况是概念分层的尺度之间均为并列关系,即,任意尺度之间均不存在包含关系.从严格意义上讲,一些数据的多尺度化不具有实际意义,即,在实际应用中,并非一般数据的多尺度化都具有现实意义和实际意义.通过以上分析,我们可以得出尺度的实质是某一度量单位所涵盖的大小范围.虽然尺度在空间数据中表现为时间、空间范畴,在一般数据中表现为其他尺度范畴,但究其原理和本质并没有发生改变.即,目标尺度数据都是通过基准尺度数据的加权求和确定的.因此,一般数据集在进行适当的概念分层形成多尺度数据集后,样本数据之间就产生了区域化变量的结构信息,再根据数据的分布情况选择合适的克里格法,进而实现一般数据的尺度转换.

3.2.2 面域普通克里格和点普通克里格

最典型的尺度上推和尺度下推方法分别是面域普通克里格法和点普通克里格法,我们将其统称为克里格法.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:

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

其中,$Z(x_i)$为样点数据;而 $Z_0^{\#}(x)$ 为待估计值;λ_i 为各个样点的权重,也称为克里格系数,且有 $\sum_{i=1}^{n+1}\lambda_i=1$.对于任意一个估计,实际值和估计值之间均存偏差,而此处的 $Z_0^{\#}(x)$ 实际上是真实值 $Z_0(x)$的一种线性无偏最优估计.公式(1)称为克里格方程,克里格系数λ可以表示成以下矩阵乘积的形式:

$$\lambda=K^{-1}D \tag{2}$$

下面分别介绍点普通克里格法和面域普通克里格法[19−21]及其在尺度下推和尺度上推方法中的应用方式,作为多尺度数据挖掘中尺度上推和尺度下推算法的理论依据.

1) 尺度下推挖掘算法的理论依据——点普通克里格法

使用点普通克里格法实现尺度下推时,关键需要确定克里格系数,矩阵 K 中 c_{ij} 为元尺度 S 中样点 i 与样点 j 之间的协方差,而 D 中 $c(x_i,x)$为元尺度 S 中样点 i 与目标尺度 S'中待估计点之间的协方差.矩阵 K 与 D 的具体形式如下:

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

$$D=\begin{bmatrix} c(x_1,x) \\ c(x_2,x) \\ \vdots \\ c(x_n,x) \\ 1 \end{bmatrix} \tag{4}$$

那么在尺度下推挖掘算法中,权重矩阵λ的具体计算公式如下:

$$\lambda=\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ -v \end{bmatrix}=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}^{-1}\times\begin{bmatrix} c(x_1,x) \\ c(x_2,x) \\ \vdots \\ c(x_n,x) \\ 1 \end{bmatrix} \tag{5}$$

例如,利用点普通克里格法进行降尺度计算,设在元尺度 S 上有 9 个相邻的已知像元及其测量值(如图 1 所示),若需要对 5 号像元进行降尺度计算,则首先将 5 号像元向下分解成目标尺度 S'上的 4 个像元,即,空心 1~空心 4 号像元,则空心 1 号像元的相关参数可以通过已知的相邻上层元尺度 S 上的实心 1、空心 2、空心 4、空心 5 号像元的相关参数加权计算获取.

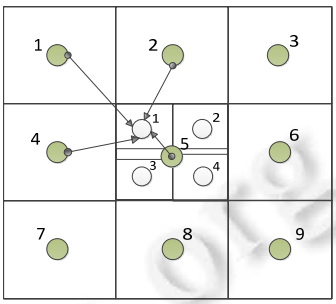


Fig.1 Scaling-Down calculation of point kriging method

图 1 点克里格法的尺度下推计算

2) 尺度上推挖掘算法的理论依据——面域普通克里格法

使用面域普通克里格法实现尺度上推,同样需要确定克里格系数λ.在此,矩阵 K 的确定方式与下推方法中点普通克里格法相同,即,由元尺度 S'中各个样点间的协方差确定 K.D 的确定方式完全不同,而是由元尺度 S'中各个样点对目标尺度 S 中待估计点的影响因子确定.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} 为元尺度 S'中样点 i 与样点 j 之间的协方差.而 D 中元素由下推方法中点普通克里格法的 $c(x_i,x)$替换为元尺度 S'中样点 i 对目标尺度 S 中待估计点 0 的影响因子 Inf_{i0},如公式(6),用面域克里格法确定克里格系数的公式为式(7):

$$D=\begin{bmatrix}\mathrm{inf}_{10}\\ \mathrm{inf}_{20}\\ \vdots\\ \mathrm{inf}_{n0}\\ 1\end{bmatrix} \tag{6}$$

$$\lambda=\begin{bmatrix}\lambda_1\\ \lambda_2\\ \vdots\\ \lambda_n\\ -v\end{bmatrix}=\begin{bmatrix}c_{11} & c_{12} & \cdots & c_{1n} & 1\\ c_{21} & c_{22} & \cdots & c_{2n} & 1\\ \vdots & \vdots & & \vdots & \vdots\\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1\\ 1 & 1 & \cdots & 1 & 0\end{bmatrix}^{-1}\times\begin{bmatrix}\mathrm{inf}_{10}\\ \mathrm{inf}_{20}\\ \vdots\\ \mathrm{inf}_{n0}\\ 1\end{bmatrix} \tag{7}$$

如图 2 所示,利用面域克里格法进行尺度上推计算时,由元尺度 S'下 4 个样点(空心 1~空心 4 号像元)的相关参数上推目标尺度 S 下的某点(实心 5 号像元)的相关参数,除了矩阵 K 中 S'下 4 个样点间的协方差对克里格系数有影响外,还需要确定矩阵 D 中元尺度 S'下的 4 个样点对目标尺度 S 下点的影响信息,即,占比信息.

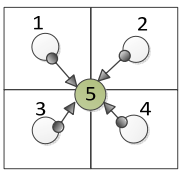


Fig.2 Scaling-Up calculation of region kriging method

图 2 面域克里格法尺度上推计算

如上文所述,克里格系数中的矩阵 D 不再是目标 S 尺度上点与元尺度上 S'已知样点之间的协方差,而是元尺度 S'上已知样点对目标尺度 S 上目标点的影响信息.

3.3 多尺度关联规则挖掘

我们以面域克里格法和点克里格法作为尺度上推和尺度下推的理论基础,将多尺度数据挖掘算法框架应用于关联规则挖掘中,提出多尺度关联规则挖掘算法 MSARMA(multi-scale association rules mining algorithm),将多尺度数据挖掘方法付诸实践.

多尺度数据挖掘算法框架可概括为 3 个关键步骤:(1) 确定基准尺度 BS,处理基准尺度数据集;(2) 对比目标尺度 S_O 和基准尺度 BS,确定多尺度挖掘任务的挖掘方向,即,判断是尺度上推挖掘还是尺度下推挖掘;(3) 进行相应方向上知识的多尺度转换,推导得到目标尺度数据集背后的知识.算法的核心为知识的多尺度转换.第 3.2 节已经详细介绍了尺度转换机制的理论基础,其关键在于确定克里格方程中的权重系数矩阵λ,即:不论是尺度上推还是尺度下推,都需要首先确定基准尺度数据集相关的协方差矩阵 K.而对于矩阵 D 的确定,上推算法中,D 为基准尺度数据集对目标尺度数据集的影响因子矩阵;下推算法中,D 为基准尺度数据集与目标尺度数据集之间的协方差矩阵.

本节将多尺度数据挖掘算法框架具化为多尺度关联规则挖掘算法.首先明确关联规则挖掘中的目标知识为关联规则,而挖掘关联规则过程中的时间和计算资源消耗主要集中于频繁项集挖掘中,故多尺度关联规则挖掘算法需要解决的核心问题是利用基准尺度数据集中的频繁项集推导目标尺度数据集中的频繁项集,进行频繁项集的多尺度转换,而不对目标尺度数据集进行直接挖掘.

这样,多尺度关联规则挖掘算法的基本思想是:首先明确基准尺度 BS,选择合适的频繁项集挖掘算法挖掘基准尺度数据集 ds_{BS},这里可以选择任意高效的频繁项集挖掘算法;然后,明确目标尺度 S_O 和目标尺度数据集 ds_{S_O},确定多尺度关联规则挖掘方向;最后,计算各基准尺度数据集 ds_{BS} 挖掘结果对目标尺度数据集 ds_{S_O} 知识的权重系数矩阵λ,这里的知识即频繁项集,λ用于估算目标尺度数据集中频繁项集的关键性参数——支持度.这里,我们对标准克里格法中权重系数的计算做一些改进:将 K 指定为基准尺度数据集之间的相似度矩阵,而非协方差矩阵,能够更准确地反映基准尺度数据集之间的相似性和相关性.在上推过程中,将矩阵 D 指定为基准尺度数据集对目标尺度数据集的影响因子矩阵;在下推过程中,将矩阵 D 指定为由领域知识确定的基准尺度数据集同目标尺度数据集之间的相似度矩阵.多尺度关联规则挖掘算法的问题描述和基本步骤如下:

- 问题描述

数据集 DS,挖掘任务的目标尺度为 S_O,目标尺度数据集为 ds_{S_O},希望通过多尺度关联规则挖掘得到目标尺度数据集 ds_{S_O} 满足最小支持度 min_sup 与最小置信度 min_$conf$ 的关联规则.

- 基本步骤

1. 选择基准尺度 BS,明确基准尺度数据集 ds_{BS}.以利于挖掘为原则,根据可用计算资源,选择能够发挥现有计算资源最大效用的尺度大小为基准尺度.

2. 以最小支持度 min_sup 挖掘所有基准尺度数据集,得到各基准尺度数据集频繁项集的集合 FI_i(基准数据集 ds_{BS}^i 的频繁项集集合),并求取上述若干频繁项集集合的并集 FI_c,作为目标尺度数据集 ds_{S_O} 频繁项集的候选项集集合.此候选项集集合能够最大程度地反映目标尺度数据集中隐含频繁项集的情况.

3. 由于无论是尺度上推还是尺度下推,都需要确定基准尺度数据集之间的相似度矩阵 K,故在此步计算基准尺度数据集之间的相似度;从统计学的观点来看,频繁项集是数据集的一种统计结果,在一定程度上代表了数据集本身的分布与特性;统计学中的 Jaccard 相似性系数主要用来比较有限样本集之间的相似性和分散性.有限集合之间的 Jaccard 系数等于两集合的交集所含元素个数与并集所含元素个数之比,见公式(8):

$$Jaccard(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{8}$$

已有研究者将 Jaccard 系数运用到数据挖掘领域的实际研究中[22,23].本文用基准尺度数据集频繁项集集合

之间的相似性估计其所属原数据集之间的相似性.

利用公式(9)计算 FI_i 之间的 Jaccard 相似性系数,用此相似性系数作为数据集 $ds_{BS}^i(i=1,...,n)$ 之间相似度的估计值,以基准尺度数据集频繁项集集合之间的相似性估计其所属原数据集之间的相似性,并构建基准尺度数据集相似度矩阵 K,其中,元素 M_{ij} 表示数据集 ds_{BS}^i 与 ds_{BS}^j 之间的相似性:

$$M_{ij}=Jaccard(FI_i,FI_j)=\frac{|FI_i\cap FI_j|}{|FI_i\cup FI_j|} \tag{9}$$

4. 判断挖掘方向,确定矩阵 D.若 $S_O\succ BS$,即进行尺度上推,显然, ds_{S_O} 为 ds_{BS} 的上层尺度数据集,则 $D=[\mathrm{Inf}_{01},\ldots,\mathrm{Inf}_{0i},\ldots,\mathrm{Inf}_{0n}]^T$,其中,$\mathrm{Inf}_{i0}$ 为 ds_{BS}^i 对 ds_{S_O} 的影响参数;若 $S_O\prec BS$,即进行尺度下推,显然, ds_{S_O} 为 ds_{BS} 的下层尺度数据集,则 $D=[s_{01},\ldots,s_{0i},\ldots,s_{0n}]^T$,其中,$S_{0i}$ 为 ds_{BS}^i 与 ds_{S_O} 之间的相似度.此处一般遵循如下原则:若 ds_{BS}^i 与 ds_{S_O} 存在祖孙关系或父子关系,那么与其对应的影响因子参数或相似度较大.影响因子是基准尺度数据集在数据量上对上层尺度数据集的占比,体现了其对上层尺度数据集表现趋势的影响程度,例如:以“民族”尺度划分某地区人口数据,同时汉族人口占该地区人口的绝大多数,那么“民族”基准尺度数据集中汉族人口对整体人口数据的占比较大,等同于该地区汉族人口对整体人口在人口数据分析中的影响程度,而汉族人口表现出的数据特性也反映了整体人口数据表现趋势.上推时,矩阵 D 的元素主要为基准尺度数据集在上层尺度数据集中数量上的占比;下推时,D 的元素即为两者的 Jaccard 相似性系数.

5. 确定基准尺度数据集对目标尺度数据集的权重矩阵:$\lambda=K^{-1}D=[\lambda_1,\ldots,\lambda_i,\ldots\lambda_n]^T$,其中,$\lambda_i$ 为基准尺度数据集 ds_{BS}^i 对目标尺度数据集 ds_{S_O} 的权重系数.对于候选项集集合中每个项集 $f\in FI_c$,估计其在 ds_{S_O} 中的支持度.方法如下:设 f 在目标尺度数据集中的支持度估计值为 $\sup_0^{\#}$,利用 f 在基准尺度数据集 ds_{BS}^i 中的支持度 $\sup_i$ 和对应权重 λ_i 进行加权求和,估算 $\sup_0^{\#}$,见公式(10):

$$\sup_0^{\#}=\sum_{i=1}^{n}\lambda_i\cdot\sup_i \tag{10}$$

公式(10)为克里格方程的变形.按上述方式处理所有候选项集.

6. 筛选目标尺度数据集最终的频繁项集,并生成关联规则.将所有候选项集的估计支持度 $\sup_0^{\#}$ 同最小支持度 min_sup 进行比较,选择 $\sup_0^{\#}$ 不小于 min_sup 的频繁项集组成 ds_{S_O} 的最终频繁项集集合 FI,并依据最小置信度 min_*conf* 产生关联规则.

算法伪代码如下:

算法 2. 多尺度关联规则挖掘算法 MSARMA.

INPUT:数据集 *DS*,支持度阈值 min_sup,置信度阈值 min_*conf*;

OUTPUT:用户感兴趣的目标尺度 S_O 数据集中的关联规则.

1. Choose the Basic Scale of *DS*: *BS*;
2. Execute association rules mining on datasets ds_{BS}^i, and get candidate frequent itemsets of ds_{S_O}:

   For each ds_{BS}^i do begin

   $FI_i=getFrequentItemset(ds_{BS}^i,\min\_sup)$;

   EndFor

   $FI_c=getUnionSet(FI_1,\ldots,FI_i,\ldots,FI_n)$;
3. Calculate similarity M_{ij} between every ds_{BS}^i and ds_{BS}^j, and build the similarity matrix K:

   For each FI_i, FI_j do begin

   $M_{ij}=getJaccardindex(FI_i,FI_j)=\frac{|FI_i\cap FI_j|}{|FI_i\cup FI_j|}$;

   $M_{ij}\rightarrow K$;

EndFor

4. While the scale S_O of DS, which is not BS

Judge the direction of association rules mining, and determine the matrix D:

If $S_O \succ BS$ then do scaling-up mining: $BS \to \uparrow S_O$

Influence info matrix: $D=[\mathrm{Inf}_{01},\ldots,\mathrm{Inf}_{0i},\ldots,\mathrm{Inf}_{0n}]^T$;

Else if $S_O \prec BS$ then do scaling-down mining: $BS \to \downarrow S_O$

Similarity matrix: $D=[s_{01},\ldots,s_{0i},\ldots,s_{0n}]^T$;

5. Calculate weight matrix λ, estimate the support value of every candidate frequent itemset in ds_{S_O}:

$\lambda=K^{-1}D=[\lambda_1,\ldots,\lambda_i,\ldots\lambda_n]^T$;

For each $f \in FI_c$ do begin

$$\sup_0^{\#}=\sum_{i=1}^{n}\lambda_i\cdot\sup_i;$$

EndFor

6. Filter candidate frequent itemsets, get final frequent itemsets of ds_{S_O}, and generate association rules:

$FI = getFinalFrequentItemset(FI_c, \min\_\sup)$;

$AssociationRules\_of\_ds_{S_O} = getAssociationRules(FI, \min\_conf)$;

Exit until there is no any scale to mine

3.4 算法评估

本节将对多尺度关联规则挖掘算法作用于数据集时的错误率做出评估与分析.首先,给出与算法评估相关的几个基本概念.

- 假正项集:在目标尺度数据集中并不频繁但被 MSARMA 算法推导出的项集,表示为 *fp*;
- 假负项集:在目标尺度数据集中确实为频繁项集但被 MSARMA 算法漏掉的项集,表示为 *fn*;
- 项集空间:算法作用于数据集时所有相关项集,包括正确推导的项集与错误推导的项集(假正项集和假负项集),表示为 *FI*;
- 错误率:假正项集与假负项集在算法项集空间中所占的比例,见公式(11),n 为项集空间中项集个数:

$$error=\frac{|fp|+|fn|}{n}\times 100\% \tag{11}$$

已知算法作用于整体数据集 DS 中的一部分 ds 的错误率 $error_S$,算法作用于 DS 其他子集或 DS 整体时的错误率 $error$ 如何估计.下面依据统计学原理和评估假设理论给出两个定理.

定理 1. 若算法在数据集 DS 中某一子集 ds 上错误率的观察值为 $error_S$,则 $error_S$ 为算法作用于 DS 其他子集或 DS 整体时错误率 $error$ 的无偏估计量.

证明:设算法 MSARMA 作用于某一数据集后,相应的项集空间为 FI.算法对任意项集 $f \in FI$ 的推导只有判断正确或错误两种结果,设随机变量 $X=0$ 表示判断正确,$X=1$ 表示判断错误,算法出错概率为 p,p 即为需要估计的 $error$.

设 FI 包含 n 个项集,则算法进行了 n 次判断,结果可形成随机变量序列 $X_i(i=1,2,\ldots,n)$.令随机变量 R 表示判读错误的项集个数,即 $R=\sum_{i=1}^{n}X_i$,则恰有 r 个出错的概率为 $P(R=r)=C_n^r p^r(1-p)^{n-r}(r=0,1,...,n)$.故算法在某一数据集上的出错次数服从二项分布 $R\sim B(n,p)$.所以,算法在样本数据集上的出错次数亦服从二项分布,此出错次数对应错误率 $error_S=r/n$.不同的样本数据集对应不同的出错次数和错误率,易得样本错误率 $error_S$ 服从二项分布.在此,$error_S$ 实则为 $error$ 的估计量.由于 $R\sim B(n,p)$,期望 $E[R]=np$,显然,$error_S$ 的期望为 p,那么 $error_S$ 为错误率 $error$ 的无偏估计量.定理 1 得证. □

定理 2. 若算法在数据集 DS 中某一子集 ds 上错误率的观察值为 $error_S$,则算法作用于 DS 其他子集或 DS

整体时的错误率 $error \geqslant error_S + z_N\sqrt{\frac{error_S(1-error_S)}{n}}$ 的概率至多为(1–*N*%)/2.

其中,*n* 为算法作用于子集 *ds* 所产生项集空间的项集个数,z_N 为表 1 中与 *N*%对应的常数.

Table 1 Value of z_N in *N*% two-sided confidence interval

表 1 双侧 *N*%置信区间的 z_N 值

置信度 *N*%	50%	68%	80%	90%	95%	98%	99%
常量 z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

证明:因为出错次数 *R* 和错误率 $error_S$ 均服从二项分布,*R* 的方差为 *np*(1–*p*),则 $error_S$ 的标准差为公式(12):

$$\sigma_{error_S} = \frac{\sigma_R}{n} = \sqrt{\frac{p(1-p)}{n}} \tag{12}$$

由于 $error_S$ 为 *p* 的无偏估计量,则 $error_S$ 标准差的近似见公式(13):

$$\sigma_{error_S} \approx \sqrt{\frac{error_S(1-error_S)}{n}} \tag{13}$$

由于 $error_S$ 服从二项分布,期望为 *error*,标准差为 σ_{error_S}.由中心极限定理可得:当 $n \geqslant 30$ 时,$error_S$ 近似服从正态分布.由于正态分布中均值μ有 *N*%的可能性落入区间 $x \pm z_N\sigma$,可得均值 *error* 有 *N*%的可能性落入区间:

$$error_S \pm z_N\sqrt{\frac{error_S(1-error_S)}{n}} \tag{14}$$

由公式(14)便可以确定 *N*%置信区间的上、下界阈值 *u* 和 *l*,见公式(15)和公式(16):

$$l = error_S - z_N\sqrt{\frac{error_S(1-error_S)}{n}} \tag{15}$$

$$u = error_S + z_N\sqrt{\frac{error_S(1-error_S)}{n}} \tag{16}$$

真实错误率 *error* 有 *N*%的可能性落入到此区间内.定理 2 得证. □

3.5 算法效率分析

本文在挖掘基准尺度数据集的频繁项集时选择了经典的Apriori算法,但此步骤不限于Apriori算法,可以选择任意的频繁项集挖掘算法,包括一些时间复杂度低且高效的算法.下面以Apriori算法为例,从时间复杂度方面对多尺度关联规则挖掘算法 MSARMA 进行分析.若选择其他的挖掘算法挖掘基准尺度数据集的频繁项集,其算法效率分析类同.

1) 算法的尺度上推部分

设下层尺度数据集的个数为 *m*,平均问题规模为 *n*.MSARMA 挖掘下层基准尺度数据集的时间复杂度是 $O(n^k)$,在处理基准尺度数据集挖掘结果部分的时间复杂度为 $O(n)$,故 MSARMA 算法的时间复杂度为 $O(mn^k+n)$;用 Apriori算法直接挖掘上层尺度数据集整体的时间复杂度为 $O((mn)^k)$.可见,MSARMA 算法的尺度上推部分在时间复杂度方面优于 Apriori 算法.

2) 算法的尺度下推部分

设上层尺度数据集的个数为 *m*,平均问题规模为 *n*,目标下层尺度数据集的问题规模为 *x*.MSARMA 挖掘上层基准尺度数据集的时间复杂度是 $O(n^k)$,在处理基准尺度数据集挖掘结果部分的时间复杂度为 $O(n)$,故 MSARMA 算法的时间复杂度为 $O(mn^k+n)$;若使用 Apriori 算法达到相同的效果,其时间复杂度将为 $O(mn^k+x^k)$.可见,MSARMA 算法的尺度下推部分在时间复杂度方面也是优于 Apriori 算法.

实际上,在具体应用中,计算机的处理能力是有限的,随着问题规模的增加,当计算机负荷增加到一定程度时,处理效率会急剧下降.MSARMA 算法这种分层次、分尺度的处理方式以及有效利用阶段性挖掘结果的方法,能够降低问题规模,提高处理效率,避免不必要的计算,具有实用价值.

3.6 算法分析

与经典的数据挖掘算法相比,MSARMA 的优势在于:MSARMA 只对基准尺度数据集进行一次挖掘,就可以得到多个不同层级尺度数据集背后隐含的关联规则.若使用经典数据挖掘算法达到相同的效果,则需要分别对多个尺度的数据集进行挖掘,即,进行多次挖掘,这将造成巨大的时间和空间开销.

另外,MSARMA 算法本身的执行机制为先处理基准尺度数据集,再对挖掘结果进行多尺度推导,若将并行运算的思想应用于 MSARMA 中,并行挖掘若干基准尺度数据集,再收集并行处理的结果进行尺度向上或尺度向下转换和推导,效率将会有较为明显的大幅提升.这对于大数据处理是非常有益的.由此可见,MSARMA 算法十分适于并行运算,多尺度数据挖掘的并行解决方案是可行且具有实践意义的.

4 多尺度关联规则挖掘算法实例

本节以尺度上推为例,通过实例展示多尺度关联规则挖掘算法 MSARMA 的执行过程.尺度下推的执行过程与此类同.

假设现在有一项挖掘任务,分析某跨国企业的购物篮数据,找到顾客的购买模式.现有数据粒度是该企业依据各门店的地理位置相关的属性集 $H_{location}$={*street*,*city*,*country*}所形成的概念分层($H_{location}$,$\prec$)={*street*$\prec$*city*$\prec$*country*}中的概念 *country*,即,现有数据尺度为 $S_{DS_{sale}}$ = *country* ,数据以国家为单位存在.与尺度 *country* 对应的属性值集为 $V_{country}$={USA,Canada,China,Japan},即,该企业只有这 4 个国家的数据.不同国家拥有各自的数据集,分别为 ds_{USA},ds_{Canada},ds_{China} 和 ds_{Japan},为方便描述,依次记为 ds_1,ds_2,ds_3 和 ds_4.目前,该企业希望通过 4 个国家的数据得到全球范围内的购买模式,即,挖掘全球数据得到整体的购买模式.可以注意到:*country* 的最近祖先概念虽然在概念分层中省略了,但应为全部 *All*,所以全球数据 ds_{all} 即为上述 4 个国家数据集的祖先尺度数据集,亦为父尺度数据集,$ds_{all}=ds_{\mathrm{USA}}\cup ds_{\mathrm{Canada}}\cup ds_{\mathrm{China}}\cup ds_{\mathrm{Japan}}$.此挖掘任务为由子尺度数据集(4 个国家的销售数据)的知识推导父尺度数据集(全球的销售数据)的关联规则,可利用尺度上推关联规则挖掘算法完成,算法应用于此例的详细过程如下所述.4 个国家的数据集 ds_1,ds_2,ds_3 和 ds_4 见表 2~表 5 所示(为了简化描述,只使用少量事务加以说明).

Table 2 USA sales data ds_1 (ds_{USA})

表 2 USA 的销售数据 ds_1(ds_{USA})

TID	商品 ID 列表
T01	*I*1, *I*2, *I*4
T02	*I*1, *I*2, *I*3, *I*4
T03	*I*2, *I*3, *I*4
T04	*I*1, *I*2, *I*3
T05	*I*1, *I*2
T06	*I*1, *I*2, *I*3, *I*4
T07	*I*1, *I*2, *I*4, *I*5
T08	*I*3, *I*4, *I*5
T09	*I*2, *I*3, *I*4, *I*5
T10	*I*1, *I*2, *I*3, *I*4

Table 3 Canada sales data ds_2 (ds_{Canada})

表 3 Canada 的销售数据 ds_2(ds_{Canada})

TID	商品 ID 列表
T01	*I*2, *I*3, *I*4
T02	*I*1, *I*2, *I*4
T03	*I*1, *I*2, *I*3
T04	*I*1, *I*2, *I*3, *I*4
T05	*I*1, *I*2, *I*3, *I*4
T06	*I*1, *I*2, *I*4
T07	*I*2, *I*3, *I*4
T08	*I*1, *I*2, *I*4, *I*5
T09	*I*1, *I*2, *I*4, *I*5
T10	*I*1, *I*2, *I*3, *I*4, *I*5

Table 4 China sales data ds_3 (ds_{China})

表 4 China 的销售数据 ds_3(ds_{China})

TID	商品 ID 列表
T01	*I*1, *I*2, *I*3, *I*4
T02	*I*2, *I*3, *I*4
T03	*I*1, *I*2, *I*3, *I*4
T04	*I*2, *I*3, *I*4
T05	*I*1, *I*3, *I*4
T06	*I*1, *I*2, *I*4
T07	*I*1, *I*2, *I*3, *I*4
T08	*I*1, *I*3, *I*4, *I*5
T09	*I*2, *I*3, *I*4, *I*5
T10	*I*2, *I*3, *I*4, *I*5

Table 5 Japan sales data ds_4 (ds_{Japan})

表 5 Japan 的销售数据 ds_4(ds_{Japan})

TID	商品 ID 列表
T01	*I*1, *I*3, *I*4, *I*5
T02	*I*1, *I*2, *I*3, *I*4
T03	*I*1, *I*2, *I*4
T04	*I*1, *I*2, *I*3, *I*4
T05	*I*1, *I*2, *I*3, *I*4
T06	*I*1, *I*2, *I*3, *I*5
T07	*I*1, *I*2, *I*3
T08	*I*1, *I*2, *I*3, *I*5
T09	*I*1, *I*2, *I*4
T10	*I*1, *I*2, *I*3, *I*5

需要通过挖掘 4 个子尺度数据集,找到全球销售数据的满足最小支持度 min_sup=0.7 和最小置信度 min_*conf*=0.85 的关联规则.应用第 3.3 节描述的多尺度关联规则挖掘算法 MSARMA,通过下面 6 步实现多尺度关联规则挖掘.

1) 计算各子尺度数据集的最小支持度,以此最小支持度挖掘各子尺度数据集,得到各自的频繁项集集合.

在此,min_sup=0.7,$|ds_1|=|ds_2|=|ds_3|=|ds_4|=10$,取 $p=0.2,4$ 个子尺度数据集的最小支持度均为

$$\min\_\sup_i = 0.7 - \sqrt{\frac{1}{2\times 10}\cdot In\frac{1}{0.2}} \approx 0.4163(i=1,2,3,4).$$

以 $\min\_\sup_i$ 作为支持度阈值挖掘数据集 ds_1,ds_2,ds_3 和 ds_4,得到它们各自的频繁项集集合 FI_1,FI_2,FI_3,FI_4,见表 6~表 9.

Table 6 ds_1 frequent itemsets FI_1(USA)

表 6 ds_1 的频繁项集集合 FI_1(USA)

项集	支持度计数
{*I*1}	7
{*I*2}	9
{*I*3}	7
{*I*4}	8
{*I*1,*I*2}	7
{*I*1,*I*4}	5
{*I*2,*I*3}	6
{*I*2,*I*4}	7
{*I*3,*I*4}	6
{*I*1,*I*2,*I*4}	5
{*I*2,*I*3,*I*4}	5

Table 7 ds_2 frequent itemsets FI_2(Canada)

表 7 ds_2 的频繁项集集合 FI_2(Canada)

项集	支持度计数
{*I*1}	8
{*I*2}	10
{*I*3}	6
{*I*4}	9
{*I*1,*I*2}	8
{*I*1,*I*4}	7
{*I*2,*I*3}	6
{*I*2,*I*4}	9
{*I*3,*I*4}	5
{*I*1,*I*2,*I*4}	7
{I2,I3,I4}	5

Table 8 ds_3 frequent itemsets FI_3(China)

表 8 ds_3 的频繁项集集合 FI_3(China)

项集	支持度计数
{*I*1}	6
{*I*2}	8
{*I*3}	9
{*I*4}	10
{*I*1,*I*3}	5
{*I*1,*I*4}	6
{*I*2,*I*3}	7
{*I*2,*I*4}	8
{*I*3,*I*4}	9
{*I*1,*I*3,*I*4}	5
{*I*2,*I*3,*I*4}	7

Table 9 ds_4 frequent itemsets FI_4(Japan)

表 9 ds_4 的频繁项集集合 FI_4(Japan)

项集	支持度计数
{*I*1}	10
{*I*2}	9
{*I*3}	8
{*I*4}	6
{*I*1,*I*2}	9
{*I*1,*I*3}	8
{*I*1,*I*4}	6
{*I*2,*I*3}	7
{*I*2,*I*4}	5
{*I*1,*I*2,*I*3}	7
{*I*1,*I*2,*I*4}	5

2) 计算子尺度数据集之间的相似度,得到子尺度数据集相似度矩阵.

$$M_{11} = Jaccard(FI_1,FI_1) = \frac{|FI_1 \cap FI_1|}{|FI_1 \cup FI_1|} = 1;$$

$$M_{12} = M_{21} = Jaccard(FI_1,FI_2) = \frac{|FI_1 \cap FI_2|}{|FI_1 \cup FI_2|} = 1;$$

$$M_{13} = M_{31} = Jaccard(FI_1,FI_3) = \frac{|FI_1 \cap FI_3|}{|FI_1 \cup FI_3|} \approx 0.6923;$$

$$M_{14} = M_{41} = Jaccard(FI_1,FI_4) = \frac{|FI_1 \cap FI_4|}{|FI_1 \cup FI_4|} \approx 0.6923;$$

$$M_{22} = Jaccard(FI_2,FI_2) = \frac{|FI_2 \cap FI_2|}{|FI_2 \cup FI_2|} = 1;$$

$$M_{23}=M_{32}=Jaccard(FI_2,FI_3)=\frac{|FI_2\cap FI_3|}{|FI_2\cup FI_3|}\approx 0.6923;$$

$$M_{24}=M_{42}=Jaccard(FI_2,FI_4)=\frac{|FI_2\cap FI_4|}{|FI_2\cup FI_4|}\approx 0.6923;$$

$$M_{33}=Jaccard(FI_3,FI_3)=\frac{|FI_3\cap FI_3|}{|FI_3\cup FI_3|}=1;$$

$$M_{34}=M_{43}=Jaccard(FI_3,FI_4)=\frac{|FI_3\cap FI_4|}{|FI_3\cup FI_4|}\approx 0.5714;$$

$$M_{44}=Jaccard(FI_4,FI_4)=\frac{|FI_4\cap FI_4|}{|FI_4\cup FI_4|}=1.$$

子尺度数据集相似度矩阵:

$$Sim\_Matrix=\begin{bmatrix}M_{11}&M_{12}&M_{13}&M_{14}\\M_{21}&M_{22}&M_{23}&M_{24}\\M_{31}&M_{32}&M_{33}&M_{34}\\M_{41}&M_{42}&M_{43}&M_{44}\end{bmatrix}=\begin{bmatrix}1&1&0.6923&0.6923\\1&1&0.6923&0.6923\\0.6923&0.6923&1&0.5714\\0.6923&0.6923&0.5714&1\end{bmatrix}.$$

3) 计算父尺度数据集的候选项集.

$$\begin{aligned}Candidate\_FI&=FI_1\cup FI_2\cup FI_3\cup FI_4\\&=\{\{I1\},\{I2\},\{I3\},\{I4\},\{I1,I2\},\{I1,I3\},\{I1,I4\},\{I2,I3\},\{I2,I4\},\\&\quad\{I3,I4\},\{I1,I2,I3\},\{I1,I2,I4\},\{I1,I3,I4\},\{I2,I3,I4\}\}.\end{aligned}$$

4) 构建候选项集支持度计数信息存储结构 *Candidate_supcntInfo*(见表 10).

Table 10 Candidate itemsets support counting information

表 10 候选项集支持度计数信息

候选项集	*m*	*Supcnt*_1 (USA)	*Supcnt*_2 (Canada)	*Supcnt*_3 (China)	*Supcnt*_4 (Japan)
{*I*1}	4	7	8	6	10
{*I*2}	4	9	10	8	9
{*I*3}	4	7	6	9	8
{*I*4}	4	8	9	10	6
{*I*1,*I*2}	3	7	8	0	9
{*I*1,*I*3}	2	0	0	5	8
{*I*1,*I*4}	4	5	7	6	6
{*I*2,*I*3}	4	6	6	7	7
{*I*2,*I*4}	4	7	9	8	5
{*I*3,*I*4}	3	6	5	9	0
{*I*1,*I*2,*I*3}	1	0	0	0	7
{*I*1,*I*2,*I*4}	3	5	7	0	5
{*I*1,*I*3,*I*4}	1	0	0	5	0
{*I*2,*I*3,*I*4}	3	5	5	7	0

5) 构建候选项集支持度计数估计信息存储结构 *est_Candidate_supcntInfo*.

对于每一个候选项集,依据结构 *Candidate_supcntInfo* 中的信息做如下操作:

① 若候选项集对应的 $supcnt_i\neq 0$,则其 est_supcnt_i 取值为 $supcnt_i$.

② 若候选项集对应的 $supcnt_i=0$,说明其在数据集 ds_i 中的支持度计数未知,需利用兄弟尺度数据集之间的相似度和其已知的精确支持度计数值对其未知的支持度计数值进行估计.例如,项集 $\{I1,I2\}\in Candidate\_FI$,需估计其在 ds_3(China)中的支持度计数:

$$est\_supcnt_3=|ds_3|\cdot\frac{1}{m}\cdot\sum_{j=1}^{4}M_{ij}\cdot\frac{supcnt_j}{|ds_j|}=10\cdot\frac{1}{3}\cdot\left(0.6923\cdot\frac{7}{10}+0.6923\cdot\frac{8}{10}+1\cdot\frac{0}{10}+0.5724\cdot\frac{9}{10}\right)=5.1787.$$

由于(est_supcnt_3=5.1787)>(min_sup$_3\cdot|ds_3|$=4.1632),所以应有 est_supcnt_3=min_sup$_3\cdot|ds_3|$=4.1632.但在实际处理过程中,由于支持度计数为整数,且这种情况下该项集在数据集中的实际支持度计数应小于 min_sup$_3\cdot|ds_3|$.所

以,这里不妨对结果 $\min\_sup_3\cdot|ds_3|$=4.1632 进行取整操作,令 est_supcnt_3=4.

对于候选项集{*I*1,*I*2,*I*4}∈*Candidate_FI*,需估算其在 ds_3(China)中的支持度计数:

$$est\_supcnt_3=|ds_3|\cdot\frac{1}{m}\cdot\sum_{j=1}^{4}M_{ij}\cdot\frac{supcnt_j}{|ds_j|}=10\cdot\frac{1}{3}\cdot\left(0.6923\cdot\frac{5}{10}+0.6923\cdot\frac{7}{10}+1\cdot\frac{0}{10}+0.5724\cdot\frac{5}{10}\right)=3.7232.$$

由于(est_supcnt_3=3.7232)<($\min\_sup_3\cdot|ds_3|$=4.1632),所以 est_supcnt_3=3.7232.

对于在所有子尺度数据集中均表现频繁的候选项集,即 *m*=4 的候选项集,将其在所有的子尺度数据集中的支持度计数值相加,计为其在父尺度数据集中的精确支持度.例如,项集{*I*1}∈*Candidate_FI*,将它在 4 个国家的子尺度数据集中的支持度计数值相加:*sum*=7+8+6+10=31,作为其在全球销售数据中的精确支持度计数;对于在某些子尺度数据集中并非频繁项集的候选项集,即 *m*<4 的候选项集,将其支持度精确值和估计值相加,作为其在父尺度数据集中支持度计数的估计值,例如,候选项集{*I*1,*I*2,*I*4}的 *sum*=5+7+7.7232+5=20.7232.处理所有候选项集,形成的支持度计数估计信息存储结构 *est_Candidate_supcntInfo* 见表 11.

Table 11 Candidate itemsets support counting estimate information

表 11 候选项集支持度计数估计信息

候选项集	*sum*	*Est_Supcnt*_1 (USA)	*Est_Supcnt*_2 (Canada)	*Est_Supcnt*_3 (China)	*Est_Supcnt*_4 (Japan)
{*I*1}	31	7	8	6	10
{*I*2}	36	9	10	8	9
{*I*3}	30	7	6	9	8
{*I*4}	33	8	9	10	6
{*I*1,*I*2}	28	7	8	4	9
{*I*1,*I*3}	21	4	4	5	8
{*I*1,*I*4}	24	5	7	6	6
{*I*2,*I*3}	26	6	6	7	7
{*I*2,*I*4}	29	7	9	8	5
{*I*3,*I*4}	24	6	5	9	4
{*I*1,*I*2,*I*3}	19	4	4	4	7
{*I*1,*I*2,*I*4}	20.723 2	5	7	3.723 2	5
{*I*1,*I*3,*I*4}	14.780 2	3.461 5	3.461 5	5	2.857 1
{*I*2,*I*3,*I*4}	20.641	5	5	7	3.641

6) 筛选父尺度数据集的最终频繁项集

计算每个项集的支持度值,见表 12.依据最小支持度 min_sup=0.7,筛选出父尺度数据集最终的频繁项集,见表 13.

最后,由表 13 中的频繁项集产生关联规则,并依据置信度阈值 min_conf=0.85 筛选出父尺度数据集最终的关联规则,见表 14.

Table 12 Candidate itemsets and support

表 12 候选项集及其支持度

候选项集	支持度计数	支持度
{*I*1}	31	0.775
{*I*2}	36	0.9
{*I*3}	30	0.75
{*I*4}	33	0.825
{*I*1,*I*2}	28	0.7
{*I*1,*I*3}	21	0.525
{*I*1,*I*4}	24	0.6
{*I*2,*I*3}	26	0.65
{*I*2,*I*4}	29	0.725
{*I*3,*I*4}	24	0.6
{*I*1,*I*2,*I*3}	19	0.475
{*I*1,*I*2,*I*4}	20.723 2	0.548 1
{*I*1,*I*3,*I*4}	14.780 2	0.369 5
{*I*2,*I*3,*I*4}	20.641	0.516 0

Table 13 Father scale datasets final frequent itemsets

表 13 父尺度数据集最终的频繁项集

频繁项集	支持度
{*I*1}	0.775
{*I*2}	0.9
{*I*3}	0.75
{*I*4}	0.825
{*I*1,*I*2}	0.7
{*I*2,*I*4}	0.725

Table 14 Father scale datasets association rules

表 14 父尺度数据集的关联规则

关联规则	支持度	置信度
$I1\Rightarrow I2$	0.7	0.9032
$I4\Rightarrow I2$	0.725	0.8788

实例的结果,与以同样的支持度与置信度阈值直接挖掘父尺度数据集(全球销售数据)所得到的结果完全一致.而尺度上推挖掘算法只通过一次挖掘数据库,不仅得到了各个子尺度数据集上的频繁项集,也得到了父尺度数据集的频繁项集和关联规则.尺度上推关联规则挖掘算法达到了从单一尺度数据集出发,发掘数据集多尺度表现形式中隐含的关联规则的目的.在本例中,通过使用尺度上推关联规则挖掘算法,不仅得到了 4 个国家销售数据中的购买模式,以对各个国家的销售决策提供信息,同时还可得到全球的购买模式,达到了在宏观上制定全球销售决策的目的,从而高效地实现了多尺度分析数据、多尺度制定决策的愿景.

5 实 验

5.1 实验准备

本文使用H省全员人口数据集和IBM T10I4D100K数据集验证MSARMA算法的可行性、准确性和效率.H省全员人口数据集完整地记录了人口的管理地和户籍地等空间属性信息,地域属性可形成概念分层$(H_{location},\prec)$={村≺乡≺县≺市≺省},具有良好的多尺度特性.人口数据集的实验中,从地域范畴出发,选择概念“乡”作为基准尺度,“县”和“村”分别作为上推和下推部分的目标尺度.抽取 H 省某一县级单位 15 周岁以上的女性人口数据进行实验,该县级单位下辖的乡级单位 15 周岁以上的女性人口数据作为基准尺度数据集,该县级单位的数据作为 MSARMA 算法中上推部分的目标尺度数据集,该县级单位下辖的某个村级单位数据作为 MSARMA 算法中下推部分的目标尺度数据集.而 IBM T10I4D100K 数据集作为人口数据集实验的补充,为了验证算法的普适性.T10I4D100K 数据集并无明显的多尺度特性,因此我们对其进行简单的顺序划分,依据划分规模形成 3 个尺度:$(H_{location},\prec)=\{S_1\prec S_2\prec S_3\}$;尺度 S_1 和 S_2 分别将数据集划分为 20 个和 10 个等规模的数据子集,S_3 为最大尺度,即,数据集整体;S_2 作为基准尺度,S_3 和 S_1 分别作为上推和下推部分的目标尺度,尺度 S_3 的整体数据集作为上推部分的目标尺度数据集,尺度 S_1 下随机选取某一子集作为下推部分的目标尺度数据集.

实验的运行环境为 Lenovo M7300 工作站,CPU Pentium 3.40GHz,4G 内存,Windows 7 操作系统,ORACLE 10g 数据库系统,使用 Octave 实现算法.本文在使用 MSARMA 算法挖掘基准尺度数据集时采用经典的 Apriori 算法[24,25].

实验将 MSARMA 算法的实验结果与直接使用 Apriori 算法挖掘目标尺度数据集的结果进行对比,以验证算法的正确性和可行性.显然,MSARMA 算法可能产生 3 种误差,即:假正项集、假负项集以及项集支持度的估计误差.其中,假正项集和假负项集在第3.4节中已经说明,项集支持度估计误差为MSARMA算法的挖掘结果中项集支持度的估计值与该项集在目标尺度数据集中支持度精确值之间的差异.为验证算法的准确性,实验采用覆盖率、精确度和平均支持度估计误差作为衡量标准.覆盖率表示 MSARMA 算法挖掘结果覆盖目标尺度数据集中真实频繁项集的比例,精确度反映了假正项集和假负项集对实验结果精确性的影响[25],平均支持度估计误差为算法中支持度是估计所得的频繁项集的支持度估计误差的平均值.以上 3 项评价标准计算公式如下:

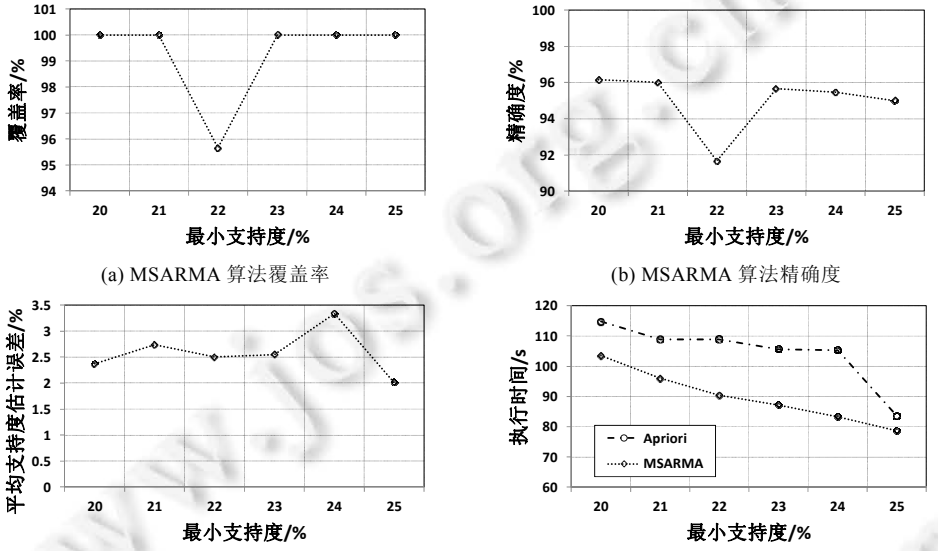
- 覆盖率: $Coverage\_Rate=\dfrac{|FI_{ms}-fp|}{|FI_o|}\times 100\%$;
- 精确度: $Accuracy=\left(1-\dfrac{|fp|+|fn|}{|FI_{ms}|+|FI_o|}\right)\times 100\%$;
- 平均支持度估计误差: $Avg\_Sup\_Error=\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{|sup_i-sup_i^{\#}|}{sup_i}\times 100\%$.

其中,FI_{ms} 表示算法 MSARMA 挖掘得到的频繁项集,FI_o 表示目标尺度数据集所包含的真实频繁项集,fp 与

fn 分别表示假正项集与假负项集,*n* 表示支持度是估计所得的频繁项集的个数,sup_i 与 $sup_i^{\#}$ 分别表示项集在目标尺度数据集中的精确支持度和 MSARMA 算法得到的估计支持度.

5.2 实验结果与分析

对比人口数据集图 3 和图 4 的实验结果可以发现:MSARMA 算法上推部分的准确度明显优于下推部分,但综合分析,上推和下推部分的准确度都比较高.

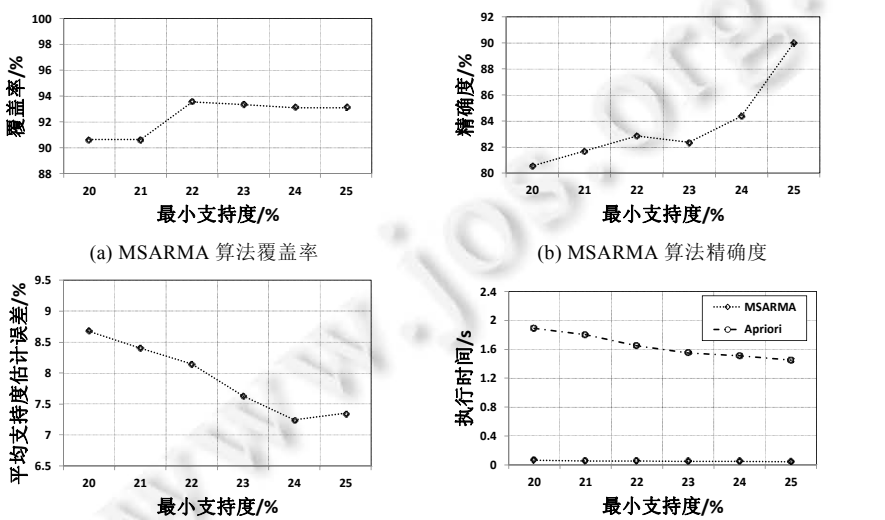


(a) MSARMA 算法覆盖率 (b) MSARMA 算法精确度

(c) MSARMA 算法平均支持度估计误差 (d) MSARMA 算法与 Apriori 算法执行时间对比

Fig.3 Scaling-Up experimental results of MSARMA applying to demographic dataset

图 3 MSARMA 算法应用于人口数据集尺度上推部分的实验结果



(a) MSARMA 算法覆盖率 (b) MSARMA 算法精确度

(c) MSARMA 算法平均支持度估计误差 (d) MSARMA 算法与 Apriori 算法执行时间对比

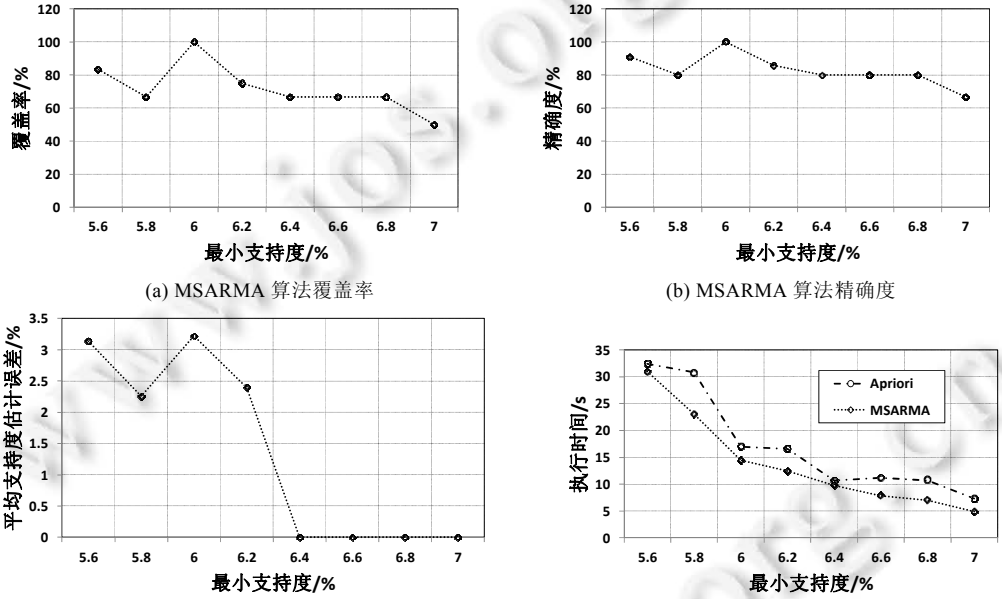
Fig.4 Scaling-Down experimental results of MSARMA applying to demographic dataset

图 4 MSARMA 算法应用于人口数据集尺度下推部分的实验结果

由图 3(a)、图 3(b)可知:上推部分的覆盖率和精确度均在 90%以上,有些情况下甚至到达了 100%.由图 4(a)、

图 4(b)中反映的覆盖率和精确度略差于上推部分,但也全部在 80%以上,验证了算法的准确性和可行性,说明 MSARMA 算法在很大程度上能够由基准尺度数据集蕴含的频繁项集得到目标尺度数据集中真实的频繁项集.图 3(a)、图 3(b)在支持度为 22%时出现陡降现象,是由于 y 轴粒度很细,这样,y 值细微的变化在图中会显示很大,实际上只下降了 4%.图 4(b)显示:随着最小支持度的增加,精度呈上升趋势,是由于假正和假负项集在总体结果的频繁项集中所占的比例减小所致.图 3(c)和图 4(c)展示的平均支持度估计误差均较低,尤其是上推算法的实验结果,基本在 3.5%以下,说明 MSARMA 算法在估计项集支持度方面表现良好.执行效率方面,图 3(d)上推部分,通过挖掘下层尺度数据集的频繁项集推导上层尺度数据集的频繁项集所用时间比直接使用 Apriori 算法挖掘上层尺度数据集更短;图 4(d)下推部分,由上层尺度数据集的挖掘结果直接推导下层尺度数据集中蕴含的频繁项集比直接使用 Apriori 算法挖掘对应下层尺度数据集耗时更短,说明算法效率较高.

图 5 和图 6 展示了 IBM T10I4D100K 数据集的实验结果.



(a) MSARMA 算法覆盖率 (b) MSARMA 算法精确度

(c) MSARMA 算法平均支持度估计误差 (d) MSARMA 算法与 Apriori 算法执行时间对比

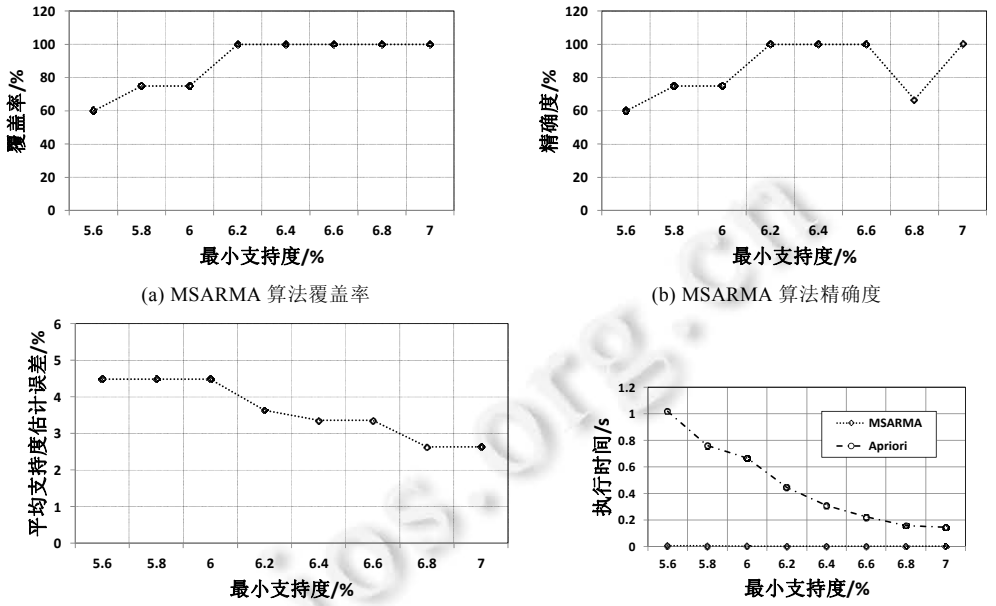
Fig.5 Scaling-Up experimental results of MSARMA applying to IBM T10I4D100K dataset

图 5 MSARMA 算法应用于 IBM T10I4D100K 数据集尺度上推部分的实验结果

相比人口数据集的实验结果,不难发现:在覆盖率和精确度反映的准确性方面,此实验结果明显要差一些,说明 MSARMA 算法更适用于尺度特性明显的数据集.图 5(d)和图 6(d)所表现的 MSARMA 在效率方面的优势十分明显.

5.3 多尺度数据挖掘算法的适用领域

从第 5.2 节的实验结果看,算法在人口数据集上的实施效果明显优于 IBM T10I4D100K 数据集.两数据集的最大区别在于:人口数据集具备明显的多尺度特性,实验中,我们从地域范畴对其进行了尺度划分;IBM T10I4D100K 数据集不具备明显的多尺度特性,为了完成实验,我们只进行了简单的顺序划分.



(a) MSARMA 算法覆盖率 (b) MSARMA 算法精确度

(c) MSARMA 算法平均支持度估计误差 (d) MSARMA 算法与 Apriori 算法执行时间对比

Fig.6 Scaling-Down experimental results of MSARMA applying to IBM T10I4D100K dataset

图 6 MSARMA 算法应用于 IBM T10I4D100K 数据集尺度下推部分的实验结果

这里的多尺度特性指数据集本身的某些属性或者属性集是空间地域相关的,或是时间相关的,或是其他能够表示范围、粒度相对大小和明确尺度含义的.实际上,多尺度数据挖掘的研究主体即为数据的多尺度实现和知识的多尺度转换.本文以多尺度数据理论初步达成了数据的多尺度实现这一目标,从这个角度上分析,具备多尺度特性的数据集的尺度划分过程具有明确依据,划分结果中的数据集也具备明确的尺度含义;而对于类似 IBM T10I4D100K 数据集这样多尺度特性不甚明显的数据集进行的顺序尺度划分,过程和结果含义并不十分明确.我们以多尺度数据挖掘算法框架和具体的多尺度关联规则挖掘实现了知识的多尺度转换,从算法角度分析,算法在具备多尺度特性的数据集上实施,无论从过程还是结果看,都更具备实际意义,尤其对于需要进行多尺度决策的情形.因此,无论在理论上还是算法实践上,多尺度数据理论和多尺度数据挖掘算法更适用于具备多尺度特性的数据集.

6 结束语

本文将多尺度科学的基本思想引入到数据挖掘领域.给出了基于概念分层的数据尺度划分及数据尺度定义,给出了多尺度数据集之间上下层尺度数据集的关系,奠定了多尺度数据理论的基础;基于知识多尺度转换的研究核心,给出了多尺度数据挖掘的定义和分类;提出了多尺度数据挖掘算法框架,同时给出了算法框架中知识多尺度转换的理论基础,并将此算法框架应用于多尺度关联规则挖掘中,提出了多尺度关联规则挖掘算法,算法利用基准尺度数据集的挖掘结果以及基准尺度数据集对于目标尺度数据集的影响权重,推导目标尺度数据集背后的关联规则,实现了知识的跨尺度推导,为多尺度决策提供了可能.最后,本文采用 H 省全员人口的真实数据集和 IBM T10I4D100K 构造数据集对算法的准确性和效率进行验证,实验结果表明:算法具有较高的覆盖率和精确度,具有较低的支持度估计误差,效率也较传统的 Apriori 算法有较大的提升.

下一步工作中,我们将致力于以下方面:研究多尺度数据挖掘的跨尺度上推和下推,分析跨尺度情况下累积误差和精确度的变化规律;研究更加完善的多尺度数据理论体系;将多尺度数据挖掘算法框架应用于诸如分类、聚类等其他挖掘任务中,探究多尺度分类和聚类挖掘算法,并完善这些算法的数据实验;进一步提高多尺度

关联规则挖掘算法的覆盖率、精确度和算法在实际应用中的效率,同时,从理论和实践上探索更优秀的权重系数计算方式,降低支持度估计误差.

References:

[1] Chai LH. Recent progress of multiscale science. Progress in Chemistry, 2005,17(2):186−189 (in Chinese with English abstract).

[2] Sakai Y, Yamanishi K. Data fusion using restricted Boltzmann machines. In: Proc. of the 2014 IEEE Int'l Conf. on Data Mining (ICDM). IEEE Computer Society, 2014. 953−958. [doi: 10.1109/ICDM.2014.70]

[3] Huo Y, Wang T, Maunder RG, Hanzo L. Motion-Aware mesh-structured trellis for correlation modelling aided distributed multi-view video coding. IEEE Trans. on Image Processing, 2013,99(1):319−331. [doi: 10.1109/TIP.2013.2288913]

[4] Chen JP, Li PX. Research on multi-scale spatial association rule mining of point object. Computer Applications, 2004,24(7):18−21 (in Chinese with English abstract).

[5] Su DH, Zhao SL, Liu MM, Su JG, Li Y. Weight vector based multi-scale clustering algorithm. Computer Science, 2015,42(4): 263−267 (in Chinese with English abstract).

[6] Li XR, Wang XJ. Remote image classification based on multi-scale boost classifier. Computer Engineering and Applications, 2013, 51(5):187−192 (in Chinese with English abstract).

[7] He KL, Ding XF. Predixtion of self-similar network traffic based on multi-scale DWT analysis. Computer Engineering and Design, 2015,36(4):915−919,961 (in Chinese with English abstract).

[8] Zhao Z. Multiscale analysis and prediction of network traffic. In: Proc. of the 2009 IEEE Int'l Conf. on Performance Computing and Communications (IPCCC). 2009. 388−393. [doi: 10.1109/PCCC.2009.5403856]

[9] McMahon C, Soe B, Loeb A, Vemulkar A, Ferry M, Bassman L. Boundary Identification in EBSD Data with A Generalization of Fast Multiscale Clustering. Ultramicroscopy, 2013,133:16−25. [doi: 10.1016/j.ultramic.2013.04.009]

[10] Lee A, Chen YA, Weng-Chong IP. Mining frequent trajectory patterns in spatial-temporal databases. Information Sciences, 2009, 179(13):2218−2231. [doi: 10.1016/j.ins.2009.02.016]

[11] Alex J, Sylvie PH, Torres RDS, Falcao AX. Interactive multiscale classification of high-resolution remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2013,6(4):2020−2034. [doi: 10.1109/JSTARS.2012. 2237013]

[12] Hoberg T, Rottensteiner F, Feitosa RQ, Heipke C. Conditional random fields for multitemporal and multiscale classification of optical satellite imagery. IEEE Trans. on Geoscience and Remote Sensing, 2015,53(2):659−673. [doi: 10.1109/TGRS.2014. 2326886]

[13] Jin SN. Research on the knowledge discovery in conceptual hierarchy knowledge base based on the multiple-level association rules [Ph.D. Thesis]. Tianjin: Tianjin University, 2006 (in Chinese with English abstract).

[14] Han JW, Kamber M, Pei J. Data mining: Concepts and Techniques. Beijing: China Machine Press, 2012. 187−240.

[15] Hu YF, Xu YZ, Liu Y, Yan Y. A review of the scaling issues of geospatial data. Advances in Earth Science, 2013,28(3):297−304 (in Chinese with English abstract).

[16] Sun QX, Li MT, Lu JX, Guo DZ, Fang T. Scale issue and its research progress of geospatial data. Geography and Geo-Information Science, 2007,23(4):53−56 (in Chinese with English abstract).

[17] Xu BX, Ye PH. Research on the method of knowledge representation. Information Science, 2007,25(5):690−694 (in Chinese with English abstract).

[18] Liu AL, Wang PF, Ding YY. Introduction to Geostatistics. Beijing: Science Press, 2012 (in Chinese).

[19] Cheng JH, Wei FY, Xue HZ. Method for spatial data scale conversion. Geospatial Information, 2008,6(4):13−15 (in Chinese with English abstract).

[20] Wang L, Hu YM, Zhao YS, Liu ZH. Remote sensing scale transformation of soil moisture based on block kriging. Journal of Geo-Information Science, 2012,14(4):465−473 (in Chinese with English abstract). [doi: 10.3724/SP.J.1047.2012.00465]

[21] Wu H, Jiang XG, Xi XH, Li CR, Li ZL. Comparison and analysis of two general scaling methods for remotely sensed information. Journal of Remote Sensing, 2009,13(2):183−189 (in Chinese with English abstract).

[22] Moens S, Goethals B. Randomly sampling maximal itemsets. In: Proc. of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics. New York: ACM Press, 2013. 79−86. [doi: 10.1145/2501511.2501523]

[23] Aggarwal CC, Li Y, Yu PS, Jin RM. On dense pattern mining in graph streams. In: Proc. of the VLDB Endowment. VLDB Endowment, 2010. 975−984. [doi: 10.14778/1920841.1920964]

[24] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Proc. of the '93 ACM SIGMOD Int'l Conf. on Management of Data. Washington: ACM Press, 1993. 207−216. [doi: 10.1145/170035.170072]

[25] Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. of the 20th Int'l Conf. on Very Large Databases. Santiago: VLDB, 1994. 487−499.

[26] Chen B, Haas P, Scheuermann P. A new two-phase sampling based algorithm for discovering association rules. In: Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2002. 462−469. [doi: 10.1145/775047.775114]

附中文参考文献:

[1] 柴立和.多尺度科学的研究进展.化学进展,2005,17(2):186−189.

[4] 陈江平,李平湘.点对象的多尺度空间关联规则挖掘算法研究.计算机应用,2004,24(7):18−21.

[5] 苏东海,赵书良,柳萌萌,苏嘉庚,李妍.基于加权向量提升的多尺度聚类挖掘算法.计算机科学,2015,42(4):263−267.

[6] 李学荣,王秀娟.基于组合分类器的遥感图像多尺度分类研究.计算机工程与应用,2013,51(5):187−192

[7] 何凯霖,丁晓峰.基于多尺度小波变化的自相似网络流量预测.计算机工程与设计,2015,36(4):915−919,961.

[13] 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[博士学位论文].天津:天津大学,2006.

[15] 胡云锋,徐芝英,刘越,艳燕.地理空间数据的尺度转换.地球科学进展,2013,28(3):297−304.

[16] 孙庆先,李茂堂,路京选,郭达志,方涛.地理空间数据的尺度问题及其研究进展.地理与地理信息科学,2007,23(4):53−56.

[17] 徐宝祥,叶培华.知识表示的方法研究.情报科学,2007,25(5):690−694.

[18] 刘爱利,王培法,丁园圆.地统计学概论.北京:科学出版社,2012.

[19] 程结海,魏峰远,薛华柱.空间数据尺度转换方法与应用.地理空间信息,2008,6(4):13−15.

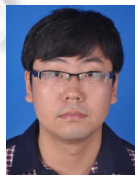
[20] 王璐,胡月明,赵英时,刘振华.克里格法的土壤水分遥感尺度转换.地球信息科学学报,2012,14(4):465−473. [doi: 10.3724/SP.J.1047.2012.00465]

[21] 吴骅,姜小光,习晓环,李传荣,李召良.两种普适性尺度转换方法比较与分析研究.遥感学报,2009,13(2):183−189.

柳萌萌(1988−),女,河北张家口人,硕士生,主要研究领域为数据挖掘,智能信息处理.

苏东海(1988−),男,硕士,主要研究领域为数据挖掘,智能信息处理.

赵书良(1967−),男,博士,教授,博士生导师,CCF 会员,主要研究领域为数据挖掘,智能信息处理.



李晓超(1986−),男,硕士生,主要研究领域为数据挖掘,智能信息处理.

韩玉辉(1989−),男,硕士生,主要研究领域为数据挖掘,智能信息处理.

陈敏(1988−),女,硕士生,主要研究领域为数据挖掘,智能信息处理.