ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.10, October 2010, pp.2395−2409
doi: 10.3724/SP.J.1001.2010.03745

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

基于谱聚类的多数据流演化事件挖掘*

杨 宁, 唐常杰+, 王 悦, 陈 瑜, 郑皎凌

(四川大学 计算机学院,四川 成都 610065)

Mining Evolutionary Events from Multi-Streams Based on Spectral Clustering

YANG Ning, TANG Chang-Jie+, WANG Yue, CHEN Yu, ZHENG Jiao-Ling

(College of Computer Science, Sichuan University, Chengdu 610065, China)

+ Corresponding author: E-mail: yneversky@gmail.com

Yang N, Tang CJ, Wang Y, Chen Y, Zheng JL. Mining evolutionary events from multi-streams based on spectral clustering. *Journal of Software*, 2010,21(10):2395−2409. http://www.jos.org.cn/1000-9825/3745.htm

Abstract: To solve the problem of mining evolutionary events from multi-streams, this paper proposes a spectral clustering algorithm, SCAM (spectral clustering algorithm of multi-streams), to generate the clustering models of Multi-Streams. The similarity matrix in the clustering models of Multi-Streams are based on Coupling Degree, which measures the dynamic similarity between two streams. In addition, this paper also proposes an algorithm, EEMA (evolutionary events mining algorithm), to discover the evolutionary event points based on the drift of clustering models. EEMA takes the index of Clustering Model Quality as the optimization objective in determing the number of clusters automatically. The Clustering Model Quality combines the matrix perturbation theory and the Clustering Cohesion, which has a sound upper bound and is used to measure the compactness of a clustering model. Finally, this paper presents O-EEMA (optimized-EEMA) as the optimization of EEMA with the temporal complexity of $O(cn^2/2)$, and the results of extensive experiments on the synthetic and real data set show that EEMA and O-EEMA are effective and practicable.

Key words: multi-streams; spectral clustering; evolutionary event; matrix perturbation

摘 要: 为解决从多数据流挖掘演化事件这一难题,提出了一种多数据流上的谱聚类算法 SCAM(spectral clustering algorithm of multi-streams),其相似矩阵基于耦合度构造,而耦合度衡量了两个数据流的动态相似性.提出了算法 EEMA(evolutionary events mining algorithm),该算法基于聚类模型的演变挖掘多数据流的演化事件.定义了聚类模型凝聚度,用以衡量聚类的紧凑程度,并证明了凝聚度的上界.基于到上界的距离和规范化相似矩阵的特征间隙,定义了聚类模型质量,并作为 EEMA 的优化目标自动地确定聚簇数 k.设计了 O-EEMA 作为 EEMA 的优化实现,其时间复杂度为 $O(cn^2/2)$.在合成和真实数据集上的实验结果表明,EEMA 和 O-EEMA 是有效的、可行的.

关键词: 多数据流;耦合聚类;演化事件;矩阵扰动

中图法分类号: TP181 **文献标识码**: A

随着信息技术的发展,数据流日益广泛地出现在网络分析[1]、传感器网络监测[2]、移动目标跟踪[3]、金融

* Supported by the National Natural Science Foundation of China under Grant No.600773169 (国家自然科学基金); the 11th Five Years Key Programs for Science & Technology Development of China under Grant No.2006BAI05A01 (国家"十一·五"科技支撑计划)

Received 2009-04-22; Revised 2009-08-12; Accepted 2009-10-10

数据分析[4]和科学数据处理[5]等应用中.面向数据流的聚类分析和事件发现,是研究数据流内在规律的重要任务[6,7].目前的研究成果主要局限于单数据流的分析处理,而在大量现实应用中需要在多个相关数据流之间进行演化聚类分析,既需要发现各个时间局部的聚类模型,又需要挖掘出全局事件点.

例如,在股票交易分析中,每支股票的交易数据都是一个数据流.用户既需要分析多支股票交易数据在某个时间段的局部关系,又需要根据局部关系的演化来发现全局的重要交易事件,以获得对整个行情的全面认识. 多个数据流之间存在与时俱进的耦合关系(即具有一致的变化趋势),根据这种关系,可以在多数据流之间进行聚类划分.耦合关系也会随时间发生变化,从而导致多数据流间聚类模型的变化,这种变化往往意味着一个全局事件的出现.图 1 显示了 3 支股票数据流 $\boldsymbol{x}_1$,$\boldsymbol{x}_2$ 和 $\boldsymbol{x}_3$ 的涨跌曲线,横坐标为时间,纵坐标为涨跌幅度.根据耦合关系的变化,整个时间区间在 t_1 和 t_2 处被分为 3 部分.当 $t<t_1$ 时,$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 的变化率更加一致,因此 3 个数据流被聚类成两部分:$\{\boldsymbol{x}_1,\boldsymbol{x}_2\}$和$\{\boldsymbol{x}_3\}$.当 $t_1<t<t_2$ 时,耦合关系发生了变化,此时 3 支股票变化率都比较一致,因此聚类模型变成了$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\}$.当 $t>t_2$ 时,$\boldsymbol{x}_1$ 和 $\boldsymbol{x}_3$ 的变化率更加一致,此时聚类模型变为$\{\boldsymbol{x}_2\}$和$\{\boldsymbol{x}_1,\boldsymbol{x}_3\}$.聚类模型的转折点 t_1 和 t_2 即为两个演化事件点.

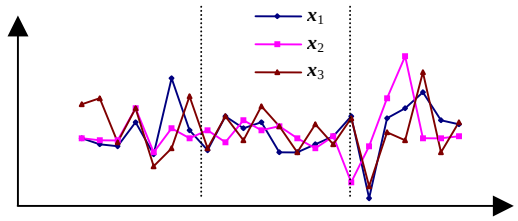


Fig.1 Coupling between streams

图 1 数据流间的耦合

上述示例中,多数据流间与时俱进的耦合关系,导致多数据流间聚类模型的演化.为揭示这种演化和全局事件出现的规律,本文做了下列工作:(1) 定义了耦合度的概念,从数据变化的角度衡量两个数据流之间的相似程度,并以此为基础把多数据流集合建模为耦合矩阵和耦合图模型,从而将以数据流为单位的多数据流聚类问题归结为耦合图的聚类,并给出基于耦合图和谱聚类的多数据流聚类算法;(2) 定义了聚类模型的凝聚度,用以评价聚类模型的整体紧凑程度,并证明了凝聚度的上界;(3) 结合矩阵扰动理论,基于凝聚度上界定义了聚类模型质量,并以此作为优化目标,自动确定聚簇数 k;(4) 设计了多数据流的演化事件挖掘算法 EEMA(evolutionary events mining algorithm)及其优化实现 O-EEMA(Optimized-EEMA),该算法根据聚类模型的变化发现全局演化事件点.

1 相关工作

本文工作涵盖数据流演化聚类分析、图的谱聚类两个方面.

数据流演化聚类分析.文献[7,8]提出了基于 k-median 的数据流聚类算法 STREAM.文献[9]进一步提出了 CluStream 算法,将数据流聚类分为在线数据统计和离线聚类两个模块.文献[10]提出了统一的数据流演化聚类框架.文献[11]研究了分布式噪声环境下当数据不完整或不可靠时的数据流聚类问题,提出了基于 EM (expectation maximization)的聚类算法.文献[12]提出的基于 PCA(principal component analysis)的算法 SPIRIT,通过对多数据流进行汇总分析实现了趋势预测和隐藏主变量的挖掘.文献[13]基于 Jaseen-Shannon Divergence 构造了两个数据分布的距离度量,设计了检测数据流分布变化的算法.文献[14,15]研究了基于小波分析的多数据流耦合模式挖掘问题.现有关于数据流的研究工作存在下述缺点中的一个或多个:

(1) 局限于单数据流中数据点的聚类问题,缺乏对以数据流为单位的多数据流之间的聚类划分的研究;

(2) 缺乏对多数据流的演化分析,尤其是根据多数据流之间的聚类模型变化挖掘演化事件;

(3) 在衡量相似性时,主要基于数据属性的当前值,没有考虑多数据流之间历史变化趋势的相似性.

图的谱聚类.近年来,谱聚类算法引起了广泛的研究兴趣,并在图划分和聚类方面得到了成功的应用[16,17].与建立在凸分布基础上的传统聚类算法,如 k-means,EM 等相比,谱聚类算法的优势在于,当数据分布空间不为凸时,基于谱图理论的谱聚类算法也能收敛于全局最优.文献[18]将谱聚类应用到 VLSI 设计中,设计了以比例割为目标函数的图划分算法.文献[19]提出了以规范化割为目标函数的 2-路划分算法.文献[20]进一步提出了以规范化割目标函数的 k-路划分算法.文献[21]研究了谱聚类与随机游走的关系,指出谱聚类使随机游走停留在簇

中的概率最大.文献[22]研究了谱聚类的优化问题,提出了完美谱聚类的概念,分析、比较了确定聚簇数 k 的若干优化标准.文献[23]提出了时态一致的演化谱聚类算法.上述算法存在下述缺点中的一个或者多个:(1) 需要预先设定聚簇数 k;(2) 没有统一的优化标准;(3) 局限于静态数据,没有处理随时间动态变化的情况;(4) 即使考虑了动态变化,也局限于聚类模型的时态一致性,缺乏对聚类模型的时态差异性的考察和应用.

2 多数据流的图聚类模型

根据本文研究的数据流的特性,本文只考虑由离散的时刻构成的时间区间.本文借鉴时态数据库中关于时间的概念[24],将时刻定义为一个时间量子,即不再可分的最小时间间隔,其粒度可以是秒、小时、天等.

定义 1(数据流的变化率). 设数据流 $\boldsymbol{x}_i$ 在时刻 t 的均值为 $avg(\boldsymbol{x}_i,t)$,则数据流 $\boldsymbol{x}_i$ 在时刻 t 的变化率 $f(\boldsymbol{x}_i,t)$定义为 $f(\boldsymbol{x}_i,t)=[avg(\boldsymbol{x}_i,t)-avg(\boldsymbol{x}_i,t-1)]/avg(\boldsymbol{x}_i,t-1)$.

定义 2(耦合度). 两个数据流 $\boldsymbol{x}_k,\boldsymbol{x}_l$ 在给定时间区间 $I=[a,b]$内的耦合度$\omega_{kl}(I)$定义为

$$\omega_{kl}(I)=\sum_{i=a}^{b}(f(\boldsymbol{x}_k,i)-F_1)(f(\boldsymbol{x}_l,i)-F_2)\Big/\sqrt{\sum_{i=a}^{b}(f(\boldsymbol{x}_k,i)-F_1)^2\sum_{i=a}^{b}(f(\boldsymbol{x}_l,i)-F_2)^2},$$

其中,$F_1=[1/(b-a)]\sum_{i=a}^{b}f(\boldsymbol{x}_k,i),F_2=[1/(b-a)]\sum_{i=a}^{b}f(\boldsymbol{x}_l,i)$.当没有歧义时,$\omega_{kl}(I)$简记为$\omega_{kl}$.

耦合度反映了两个数据流在某个历史阶段的动态相关性,即变化趋势的相似性.容易证明,ω_{kl} 的取值范围为[−1,1].当ω_{kl}=0 时,数据流 $\boldsymbol{x}_k,\boldsymbol{x}_l$ 没有任何耦合关系;当ω_{kl}=1 时,数据流 $\boldsymbol{x}_k,\boldsymbol{x}_l$ 完全正耦合,即不仅变化的幅度,而且变化的方向都是相同的;当ω_{kl}=−1 时,数据流 $\boldsymbol{x}_k,\boldsymbol{x}_l$ 完全负耦合,此时变化的幅度相同但方向相反.总的说来,耦合度越接近 0,两个数据流的耦合关系越弱,动态的差异性就越大.

从图论的观点来看,如果把 n-数据流集合 $\boldsymbol{X}^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$中的每个数据流 $\boldsymbol{x}_i(i=1,\ldots,n)$都看作一个结点,则耦合度就可看作两个结点间边的权重,它度量了两个数据流在变化上的相关性.由此可以得到多数据流的耦合图、耦合矩阵和图聚类模型的定义.

定义 3(耦合图、耦合矩阵和图聚类模型). 设 n-数据流集合为 $\boldsymbol{X}^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$,则:

(1) $\boldsymbol{X}^n$ 的耦合图定义为加权图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$,其中:

(a) 顶点集 $\boldsymbol{V}=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$,边集 $\boldsymbol{E}=\{e_{ij}|\boldsymbol{x}_i,\boldsymbol{x}_j,\omega_{ij}(I)\neq 0,i\neq j,i,j=1,\ldots,n\}$;

(b) 边 e_{ij} 在时间区间 $I=[a,b]$内的权重为$\omega_{ij}(I)$.无歧义时,结点 $\boldsymbol{x}_i$ 简记为 i.

(2) $\boldsymbol{X}^n$ 的耦合矩阵定义为图 $\boldsymbol{G}$ 在时间区间 $I=[a,b]$内的邻接矩阵$\boldsymbol{\Omega}(I)=(\omega_{ij}(I))_{i,j=1,\ldots,n}$.

(3) $\boldsymbol{X}^n$ 在时间区 $I=[a,b]$上的图聚类模型定义为$\boldsymbol{\Delta}_k(I)=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$,其中,$k$ 为聚簇数,$\boldsymbol{C}_i(i=1,\ldots,k)$为满足下列条件的簇:

(a) $\bigcup_{i=1}^{k}\boldsymbol{C}_i=\boldsymbol{X}^n$;

(b) $\forall i\neq j,\boldsymbol{C}_i\cap\boldsymbol{C}_j=\varnothing$;

(c) $\forall\boldsymbol{x}_i,\boldsymbol{x}_j\in\boldsymbol{X}^n$ 的相似性由耦合度$\omega_{ij}(I)$度量.无歧义时,$\boldsymbol{\Delta}_k(I)$简记为Δ_k,当不考虑聚簇数 k 时,简记为$\boldsymbol{\Delta}$.

定义 4(全局演化事件划分). 给定全局时间区间$[s,e]$,n-数据流集合 $\boldsymbol{X}^n$ 的全局演化事件划分定义为满足下列条件的时刻集合 $\boldsymbol{T}=\{t_1,t_2,\ldots,t_m,t_{m+1}\}$,其中:

(1) $t_1=s,t_{m+1}=e$;

(2) 全局时间被划分为 m 个区间 $I_1,I_2,\ldots,I_m$,其中,$I_i=[t_i,t_{i+1}]$;

(3) 相邻区间 I_i,I_{i+1} 的聚类模型不同,即$\boldsymbol{\Delta}_k(I_i)\neq\boldsymbol{\Delta}_k(I_{i+1})$.

问题界定. 本文的目标是根据耦合度对 n-数据流集合 $\boldsymbol{X}^n$ 进行聚类划分,从而得到它的图聚类模型,并根据图聚类模型的演变挖掘全局演化事件划分.

3 谱聚类的基本概念

设耦合图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$,其耦合矩阵$\boldsymbol{\Omega}=(\omega_{ij}),i,j=1,\ldots,n,n$ 为结点数.显然,$\omega_{ij}=\omega_{ji}$.每个结点 $\boldsymbol{x}_i\in\boldsymbol{V}$ 的度定义为 $d_i=\omega_{i1}+\omega_{i2}+\ldots+\omega_{in}$.称矩阵 $\boldsymbol{D}=\mathrm{diag}(d_1,d_2,\ldots,d_n)$为度矩阵.

谱聚类的思想来源于谱图理论[16],它利用图的 Laplacian 矩阵前 k 个最小特征值所对应的特征向量,在谱映射空间 $\mathbb{R}^k$ 中利用某种聚类算法(例如 k-Means)将图结点聚类成 k 个簇[17].Laplacian 矩阵有 3 种形式[17],其中,未规范化的 Laplacian 矩阵定义为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{\Omega}$,规范化且对称的 Laplacian 矩阵定义为 $\boldsymbol{L}_{sym}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{\Omega}\boldsymbol{D}^{-1/2}$,规范化但不对称的 Laplacian 矩阵定义为 $\boldsymbol{L}_{rw}=\boldsymbol{D}^{-1}\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{\Omega}$.实验和统计分析结果表明,使用 $\boldsymbol{L}_{sym}$ 和 $\boldsymbol{L}_{rw}$ 的聚类性能优于 $\boldsymbol{L}_{sym}$ 和 $\boldsymbol{L}_{rw}$,尤其是当聚簇数据倾斜分布时(有些聚簇密集,有些聚簇稀疏)[25,26].$\boldsymbol{L}_{sym}$ 和 $\boldsymbol{L}_{rw}$ 的特征值和特征向量具有下述关系:λ是 $\boldsymbol{L}_{rw}$ 的特征值且对应λ的特征向量为 $\boldsymbol{v}$ 当且仅当λ是 $\boldsymbol{L}_{sym}$ 的特征值且对应λ的特征向量 $\boldsymbol{w}=\boldsymbol{D}^{-1/2}\boldsymbol{v}$[22].因此对于谱聚类而言,$\boldsymbol{L}_{sym}$ 和 $\boldsymbol{L}_{rw}$ 是等效的.基于上述讨论,同时考虑到多数据流在耦合关系上可能出现倾斜分布的情况以及对称矩阵处理的便利,所以本文决定采用 $\boldsymbol{L}_{sym}$,并直接记为 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{\Omega}\boldsymbol{D}^{-1/2}$.

关于谱聚类有下述重要概念:

定义 5(谱映射)[22]. 对于耦合图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$和 n 维向量集合 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$,谱映射 $S:\boldsymbol{V}\to(v_{1i},v_{2i},\ldots,v_{ki})$将耦合图结点 $\boldsymbol{x}_i(i=1,\ldots,n)$映射为 $\mathbb{R}^k$ 中的一个点,其中 $v_{ri}(r=1,\ldots,k)$是第 r 个向量的第 i 个分量.称 $\mathbb{R}^k$ 为谱映射空间.

可以用随机游走的观点解释谱聚类[21].随机游走在一个簇中停留的概率较大,而在簇间转移的概率较小.在耦合图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$中,定义随机转移矩阵为 $\boldsymbol{P}=(p_{ij})_{i,j=1,\ldots,n}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$,结点 i 到结点 j 的一步转移概率为 $p_{ij}=\omega_{ij}/d_i$,从而 $\boldsymbol{P}$ 可看作是在图 $\boldsymbol{G}$ 上的一个 Markov 随机游走.显然,矩阵 $\boldsymbol{P}$ 是规范化的耦合矩阵.关于矩阵 $\boldsymbol{P}$ 和 Laplacian 矩阵 $\boldsymbol{L}$ 的关系有如下引理:

引理 1[22]. 如果$\mu_1<\mu_2<\ldots<\mu_k$是 $\boldsymbol{L}$ 的前 k 个最小的特征值,对应的特征向量为 $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k$,$\lambda_1>\lambda_2>\ldots>\lambda_n$是 $\boldsymbol{P}$ 的前 k 个最大特征值,对应的特征向量为 $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_k$,则对所有的 $i=1,\ldots,k$,有$\mu_i=1-\lambda_i,\boldsymbol{v}_i=\boldsymbol{D}^{1/2}\boldsymbol{y}_i$.

4 多数据流的谱聚类算法 SCAM

本节设计了多数据流之间的谱聚类算法 SCAM(spectral clustering algorithm of multi-streams).算法输入为 n-数据流集合 $\boldsymbol{X}^n$、时间区间 I 和聚簇数 k,输出聚类模型$\boldsymbol{\Delta}_k(I)$.算法 1 给出了主要步骤.

算法 1. 多数据流的谱聚类算法 $SCAM(X^n,I,k)$.

输入:n-数据流集合 $\boldsymbol{X}^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$,时间区间 I,聚簇数 k;

输出:聚类模型模型$\boldsymbol{\Delta}_k(I)$.

(1) For each pair of $\boldsymbol{x}_i,\boldsymbol{x}_j\in\boldsymbol{X}^n$, calculate $\omega_{ij}(I)$ according to definition 1;

(2) Construct the matrix $\boldsymbol{\Omega}(I)=(\omega_{ij}(I))_{i,j=1,\ldots,n}$;

(3) Compute the matrix $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{\Omega}\boldsymbol{D}^{-1/2}$, $\boldsymbol{D}=diag(d_1,d_2,\ldots,d_n)$;

(4) Compute the first k smallest eigenvectors $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$ of $\boldsymbol{L}$ by employing Lanczos algorithm[27];

(5) Let $\boldsymbol{Y}\in\mathbb{R}^{n\times k}$ be the matrix containing the vectors $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$ as columns;

(6) For each $i=1,\ldots,n$, let $\boldsymbol{y}_i\in\mathbb{R}^k$ be the vector corresponding to the i-th row of $\boldsymbol{Y}$;

(7) Cluster the $points(\boldsymbol{y}_i)_{i=1,\ldots,n}$ with the k-Means into k clusters $\{\boldsymbol{A}_1,\boldsymbol{A}_2,\ldots,\boldsymbol{A}_k\}$;

(8) Return $\boldsymbol{\Delta}_k(I)=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$, where $\boldsymbol{C}_j=\{\boldsymbol{x}_i|\boldsymbol{y}_i\in\boldsymbol{A}_j\}$, $i=1,\ldots,n$, $j=1,\ldots,k$;

算法 SCAM 使用耦合度来衡量两个数据流之间的相似程度(第(1)行),并以此构造耦合矩阵$\boldsymbol{\Omega}$,然后调用 QR 算法对$\boldsymbol{\Omega}$对应的 Laplacian 矩阵进行谱特征分解(第(4)行),最后在谱映射空间 $\mathbb{R}^k$ 中采用 k-Means 算法将 n-数据流集合 $\boldsymbol{X}^n$ 划分为 k 个簇,从而得到聚类模型$\boldsymbol{\Delta}_k$(第(5)行~第(8)行).注意,算法 1 的主要时间消耗是第(4)行的特征分解,由于耦合矩阵是实对称矩阵,所以采用 QR 算法,这时特征分解的时间复杂度从 $O(n^3)$降为 $O(n)$[27].整个算法的时间复杂度有如下命题:

命题 1. 算法 SCAM 的时间复杂度为 $O(n^2/2)$.

证明:参见附录. □

注意,SCAM 算法需要输入聚簇数 k 作为参数,k 的大小依赖于多数据流耦合关系随时间的演化.如何确定 k,将在第 6 节中加以研究.

5 多数据流的完美聚类模型

本节提出了多数据流的完美聚类模型,并应用矩阵扰动理论证明了达到完美聚类模型的条件.

定义 6(分段常数向量)[22]. 设有 n 维向量 $\boldsymbol{v}$,n-数据流集合为 $\boldsymbol{X}^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$及其聚类模型$\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$,如果当 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 属于同一个簇时有 $v_i=v_j$,v_i,v_j 分别是向量 $\boldsymbol{v}$ 的第 i 个和第 j 个分量,则称 $\boldsymbol{v}$ 对于$\boldsymbol{\Delta}_k$ 是分段常数向量.

定义 7(块随机矩阵)[22]. 设$\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$是耦合图 $\boldsymbol{G}$ 上的一个聚类模型.如果满足以下条件,则称矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$对于$\boldsymbol{\Delta}_k$ 是块随机矩阵:(1) 对任意的 $m,l=1,\ldots,k$,从 $\boldsymbol{C}_l$ 中的任意一点转移到聚簇 $\boldsymbol{C}_m$ 的概率是相等的,即对所有的 $i\in\boldsymbol{C}_l$, $P_{im}=\sum_{j\in\boldsymbol{C}_m}p_{ij}$ 都相等;(2) 矩阵 $\boldsymbol{R}=(r_{ml})_{m,l=1,\ldots,k}$ 是非奇异的,其中, $r_{ml}=\sum_{i\in\boldsymbol{C}_m,j\in\boldsymbol{C}_l}p_{ij}$ 是从聚簇 $\boldsymbol{C}_m$ 转移到 $\boldsymbol{C}_l$ 的概率.

注意,根据随机游走的观点,定义 7 中的 p_{ij} 正是从结点 i 转移到结点 j 的概率,而 r_{ml} 是从聚簇 $\boldsymbol{C}_m$ 转移到聚簇 $\boldsymbol{C}_l$ 的概率.关于块随机矩阵和分段常数向量有下述引理和命题:

引理 2[22]. 如果$\{\lambda_i\}_{i=1,\ldots,k}$和$\{\gamma_i\}_{i=1,\ldots,k}$分别是定义 7 中矩阵 $\boldsymbol{P}$ 和 $\boldsymbol{R}$ 的前 k 个最大特征值,则对所有的 $i=1,\ldots,k$,有$\lambda_i=\gamma_i$.

引理 3[22]. Laplacian 矩阵 $\boldsymbol{L}$ 的前 k 个最小特征值所对应的特征向量对于聚类模型$\boldsymbol{\Delta}_k$ 是分段常数向量当且仅当是 $\boldsymbol{P}$ 对于$\boldsymbol{\Delta}_k$ 为块随机矩阵.

命题 2. 如果 n 维向量集合 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$ 对于聚类$\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$是分段常数向量,则定义 5 给出的谱映射 S 将同一个类中的所有点映射为$\mathbb{R}^k$ 中的同一个点.

证明:参见附录. □

根据引理 3 和命题 2,当 $\boldsymbol{P}$ 对于聚类模型$\boldsymbol{\Delta}_k$ 是块随机矩阵时,此时$\boldsymbol{\Delta}_k$ 达到理想聚类效果:同一聚簇中的点被映射成谱映射空间$\mathbb{R}^k$ 中的单个点.由此可以得到多数据流的完美聚类模型的定义.

定义 8(多数据流的完美聚类模型). 设 n-数据流集合 $\boldsymbol{X}^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$,由 SCAM 得到聚类模型$\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$,对于任意 $\boldsymbol{x}_i\in\boldsymbol{C}_u,\boldsymbol{x}_j\in\boldsymbol{C}_v,i,j=1,\ldots,n,u,v=1,\ldots,k$,如果当 $u=v$ 时谱映射 $S(i)=S(j)$,则称$\boldsymbol{\Delta}_k$ 为 $\boldsymbol{X}^n$ 的完美聚类模型,记为 $\boldsymbol{\Delta}_k^*$,对应的块随机矩阵记为 $\boldsymbol{P}^*$.

图 2 是一个完美聚类模型的示例,图中 2 维数据空间$\mathbb{R}^2$ 中的点被划分为 3 个簇,分别被映射到了谱映射空间$\mathbb{R}^3$ 中的 3 个点.

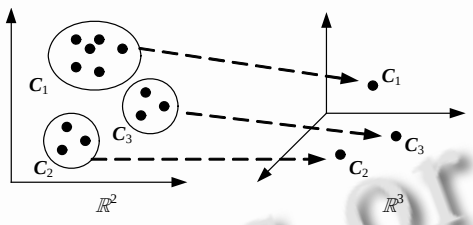


Fig.2 Perfect clustering model

图 2 完美聚类模型

命题 3. 算法 SCAM 求出的聚类模型$\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$是完美聚类模型当且仅当矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$对于$\boldsymbol{\Delta}_k$ 是块随机矩阵.

证明:参见附录. □

由于存在噪声,$\boldsymbol{P}$ 对于$\boldsymbol{\Delta}_k$ 不一定是块随机矩阵 $\boldsymbol{P}^*$.此时,算法 SCAM 求出的聚类模型$\boldsymbol{\Delta}_k$ 不一定是完美的.根据矩阵扰动理论[28],可以认为 $\boldsymbol{P}=\boldsymbol{P}^*+\boldsymbol{H}$,其中,$\boldsymbol{H}$ 为 $\boldsymbol{P}^*$的扰动.首先引入特征间隙和矩阵距离的定义.

定义 9(特征间隙和矩阵距离)[28]. 设 $\boldsymbol{A},\boldsymbol{H}\in\mathbb{R}^{n\times n}$ 是对称矩阵,$\|\cdot\|$为矩阵的 2-范数.考虑矩阵 $\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{H}$,其中,$\boldsymbol{H}$ 为 $\boldsymbol{A}$ 的扰动.设 $W\subset\mathbb{R}$为一实数区间,包含在 W 中的 $\boldsymbol{A}$ 和 $\tilde{\boldsymbol{A}}$ 特征值集合分别记为$\sigma_W(\boldsymbol{A})$和$\sigma_W(\tilde{\boldsymbol{A}})$,二者对应的特征向量生成的特征空间分别记为 $\boldsymbol{E}$ 和 $\tilde{\boldsymbol{E}}$,则:(1) 矩阵 $\boldsymbol{A}$ 的特征间隙定义为$\delta(\boldsymbol{A})=\min\{|\lambda-s|;\lambda,s$ 是 $\boldsymbol{A}$ 的特征值,

且$\lambda\in W, s\notin W$};(2) $\boldsymbol{A}$ 和 $\tilde{\boldsymbol{A}}$ 的距离定义为 $d(\boldsymbol{A},\tilde{\boldsymbol{A}})=\|\sin\Theta(\boldsymbol{E},\tilde{\boldsymbol{E}})\|$,其中,$\Theta(\boldsymbol{E},\tilde{\boldsymbol{E}})$为 $\boldsymbol{E}^T\tilde{\boldsymbol{E}}$ 的奇异值矩阵.

关于矩阵距离和特征间隙的关系有如下引理和命题:

引理 4[28]. $d(\boldsymbol{A},\tilde{\boldsymbol{A}})\leqslant\|\boldsymbol{H}\|/\delta(\boldsymbol{A})$.

命题 4. 设$\{\lambda_i\}_{i=1,\ldots,k+1}$是矩阵 $\boldsymbol{P}$ 的前 k+1 个最大特征值且$\lambda_i>\lambda_{i+1}$,$\boldsymbol{P}=\boldsymbol{P}^*+\boldsymbol{H}$,如果扰动 $\boldsymbol{H}$ 一定时,则当$\lambda_k-\lambda_{k+1}$越大时,$\boldsymbol{P}$ 在渐进意义上越接近块随机矩阵 $\boldsymbol{P}^*$.

证明:参见附录. □

当聚簇数为 k 时,矩阵 $\boldsymbol{P}$ 的特征间隙记为$\delta_k(\boldsymbol{P})$.根据命题 4,$\delta_k(\boldsymbol{P})=\lambda_k-\lambda_{k+1}$,如果$\delta_k(\boldsymbol{P})$越大,则$\Delta_k$越接近完美.

6 聚类模型质量

本文采用文献[19]提出的 2-路规范化割作为簇间相似度的度量,则聚类模型Δ_k的规范化割和凝聚度分别定义如下:

定义 10(聚类模型的规范化割). 聚类模型$\Delta_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$的规范化割定义为

$$CMNCut(\Delta_k)=\sum_{i=1}^{k}[(\sum_{j=1}^{i-1}Cut(\boldsymbol{C}_i,\boldsymbol{C}_j)+\sum_{j=i+1}^{k}Cut(\boldsymbol{C}_i,\boldsymbol{C}_j))/vol(\boldsymbol{C}_i)],$$

其中,$Cut(\boldsymbol{C}_i,\boldsymbol{C}_j)=\sum_{s\in\boldsymbol{C}_i}\sum_{t\in\boldsymbol{C}_j}\omega_{st}$,$vol(\boldsymbol{C}_i)=\sum_{i'=1}^{k}Cut(\boldsymbol{C}_i,\boldsymbol{C}_{i'})=\sum_{j\in\boldsymbol{C}_i}d_j$.

定义 11(凝聚度). 聚类模型$\Delta_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$的凝聚度定义为 $Coh(\Delta_k)=1/CMNCut(\Delta_k)$.

显然,聚类模型越完美,簇间相似度越小,凝聚度越大.命题 5 给出了完美聚类模型的凝聚度.

命题 5. 如果经算法 SCAM 求出的聚类模型是完美聚类模型 Δ_k^*,则其凝聚度为 $Coh(\Delta_k^*)=1\Big/\sum_{i=1}^{k}\mu_i$.其中,$\mu_1<\mu_2<\ldots<\mu_k$是 Laplacian 矩阵 $\boldsymbol{L}$ 的前 k 个最小特征值.

证明:参见附录. □

容易想到,给定 Laplacian 矩阵 $\boldsymbol{L}$ 和聚簇数 k,如果存在完美聚类模型 Δ_k^*,那么 Δ_k^* 的凝聚度是所有可能的聚类模型Δ_k的凝聚度的上界.为了证明这一结论,首先证明如下引理:

引理 5. 设矩阵 $\boldsymbol{L}$ 是实对称矩阵,其特征值从小到大为$\mu_1<\mu_2<\ldots<\mu_n$,对应的标准化特征向量为 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$,如果向量 $\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_k$ 是 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$ 的线性组合,且 $\boldsymbol{y}_i^T\boldsymbol{y}_i=1,\boldsymbol{y}_i^T\boldsymbol{y}_j=0,i\neq j$,则 $\min(\sum_{i=1}^{k}\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i)=\sum_{i=1}^{k}\mu_i$.

证明:参见附录. □

命题 6. 已知 n-数据流集合 $\boldsymbol{X}^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$的耦合图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$,$\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$,Laplacian 矩阵 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{\Omega}\boldsymbol{D}^{-1/2}$ 和聚簇数 k,$\mu_1<\mu_2<\ldots<\mu_n$是 $\boldsymbol{L}$ 的前 k 个最小特征值,且对应的标准化特征向量为 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$,由 SCAM 算法得到聚类模型$\Delta_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$,则 $Coh(\Delta_k)\leqslant Coh(\Delta_k^*)$.

证明: 参见附录. □

定义 12(凝聚度间距). 聚类模型Δ_k的凝聚度间距 $CohGap(\Delta_k)$定义为

$$CohGap(\Delta_k)=[Coh(\Delta_k^*)-Coh(\Delta_k)]/Coh(\Delta_k^*).$$

命题 6 和定义 12 提示我们,可以用聚类模型的凝聚度间距来衡量聚类模型的质量,凝聚度间距越小,聚类模型越完美;另一方面,根据命题 4 可知,矩阵 $\boldsymbol{P}$ 的特征间隙越大,聚类模型质量越接近完美.综合考虑这两方面的因素,定义聚类模型质量如下:

定义 13(聚类模型质量). 给定矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$和聚簇数 k,$\{\lambda_i\}_{i=1,\ldots,k}$是矩阵 $\boldsymbol{P}$ 的前 k 个最大特征值,由 SCAM 算法求得的聚类模型Δ_k的质量定义为 $Quality(\Delta_k)=\delta_k(\boldsymbol{P})/CohGap(\Delta_k)$.

7 挖掘演化事件

随着时间的推移,数据流之间的耦合关系将发生变化,从而引起聚类模型的演化.本节实现的算法 EEMA (evolutinary events mining algorithm)的基本思想是,在每个时间步确定聚簇数 k,并调用 SCAM 算法求得当前的聚类模型,然后与上一时间步的聚类模型进行比较.如果发生变化,则说明发生了一个演化事件.

确定聚簇数 k 是算法 EEMA 需要解决的主要问题.根据上一节的结论,k 的确定可以转化为从一系列可能的聚类模型 $\Delta_1,\ldots,\Delta_n$ 中选择质量最好的聚类模型,即

$$k=\arg\max_{k=1,\ldots,n} Quality(\Delta_k) \tag{1}$$

算法 2 给出了 EEMA 算法的主要步骤.

算法 2. *EEMA*(X^n,s,e,w).

输入:n-数据流集合 $\boldsymbol{X}^n$,开始时间 s,结束时间 e,时间步长 w;

输出:全局事件划分 $\boldsymbol{T}=\{t_1,t_2,\ldots,t_m,t_{m+1}\}$.

(1) $\boldsymbol{T}=\{s\}$; $I=[s,s+w]$;
(2) $\Delta_{old}=OptimalClustering(\boldsymbol{X}^n,I)$; //初始化第 1 个时间周期的最优聚类模型
(3) $t=s$;
(4) while ($t\leqslant e$) {
(5) $I=[t,t+w]$;
(6) $\Delta_{new}=OptimalClustering(\boldsymbol{X}^n,I)$; //当前时刻的最优聚类模型
(7) if ($\Delta_{new}\neq\Delta_{old}$) $\boldsymbol{T}=\boldsymbol{T}\cup\{t\}$; //如果聚类模型发生变化,则加入一个事件点
(8) $t=t+w$;
(9) }
(10) if ($e\notin\boldsymbol{T}$) $\boldsymbol{T}=\boldsymbol{T}\cup\{e\}$;
(11) return $\boldsymbol{T}$;

算法 2 在每一个时间步调用 *OptimalClustering* 得到当前最新的聚类模型(第(6)行),然后与上一个时间步的聚类模型比较,如果发生改变,则把当前的时刻作为事件点加入到结果集合中(第(7)行).

算法 3 给出了函数 *OptimalClustering* 的主要步骤.

算法 3. *OptimalClustering*(X^n,I).

输入:n-数据流集合 $\boldsymbol{X}^n$,时间区间 I;

输出:最优的聚类模型 $\Delta_{opt}(I)$.

(1) $\Delta_{opt}(I)=\Delta_1=SCAM(\boldsymbol{X}^n,I,1)$; //初始化最优聚类模型
(2) $P_{\max}=Quality(\Delta_1)$; //根据定义 13 初始化聚类模型质量
(3) $k=2$;
(4) while ($k<n=$ {
(5) $\Delta_k(I)=SCAM(\boldsymbol{X}^n,I,k)$; //调用算法 SCAM 求得具有 k 个聚簇的聚类模型
(6) $P_k=Quality(\Delta_k)$; //根据定义 13 计算当前聚类模型质量
(7) if ($P_k>P_{\max}$) { //寻找质量最大的聚类模型
(8) $P_{\max}=P_k$;
(9) $\Delta_{opt}(I)=\Delta_k(I)$;
(10) }
(11) $k=k+1$;
(12) }
(13) return $\Delta_{opt}(I)$;

注意,在实现算法 3 和 SCAM 算法时可做出优化:在循环开始前而不是在 SCAM 算法中计算 $\boldsymbol{\Omega}$,从而只需计算 1 次 $\boldsymbol{\Omega}$.优化后的算法称为 O-EEMA.关于 EEMA 和 O-EEMA 的时间复杂度有下述命题:

命题 7. 算法 EMMA 的时间复杂度为 $O(n^3/2)$,算法 O-EMMA 的时间复杂度为 $O(cn^2/2)$,其中,c 为常数 $\lceil(e-s)/w\rceil$.

证明:参见附录. □

8 实验和分析

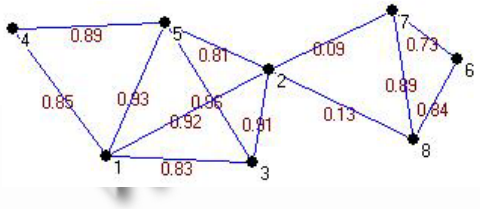
实验有 4 个内容:(1) 基于合成数据集验证 SCAM 算法的有效性;(2) 基于真实数据集验证 EEMA 算法的有效性;(3) 基于真实数据集比较 EEMA 算法和其他算法的聚类效果;(4) 测试 EEMA 算法的性能和规模可伸缩性.此外,为更加客观地评价聚类效果,本文采用文献[29]提出的 Silhouette 值作为第三方评价标准,与本文得出的聚类模型质量相互印证.下面首先简要介绍 Silhouette 值的概念.

8.1 Silhouette值

结点 i 的 Silhouette 值定义为 $b(i)$和 $a(i)$的标准差,其中,$a(i)$是结点到同簇中结点的平均距离,而 $b(i)$是结点到其他簇中所有结点的平均距离.Silhouette 值的取值范围为[−1,1].如果 Silhouette 值接近 1,那么结点 i 离自己的簇比离其他邻近的簇要近,所以是分类良好的;反之,如果接近−1,则是被错分的[29].

8.2 基于合成数据验证SCAM算法的有效性

本节实验的目的是在合成数据上:(1) 验证第 6 节中给出的聚类模型质量定义(定义 13)的有效性;(2) 验证以第 7 节中给出的公式(1)作为优化目标、能够优化地确定聚簇数 k 并由 SCAM 算法生成最优聚类模型$\boldsymbol{\Delta}_k$.合成数据集为 8-数据流集合 $\boldsymbol{X}^8=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_8\}$,其耦合图和耦合矩阵如图 3 所示.直观地,期望得到的最优聚类模型是$\boldsymbol{\Delta}_2=\{\boldsymbol{C}_1,\boldsymbol{C}_2\}$,$\boldsymbol{C}_1=\{1,2,3,4,5\}$,$\boldsymbol{C}_2=\{6,7,8\}$.



(a) Coupling graph of $\boldsymbol{X}^8$
(a) $\boldsymbol{X}^8$ 的耦合图

$$\begin{bmatrix} 1.0000 & 0.9900 & 0.8300 & 0.8500 & 0.9300 & 0 & 0 & 0 \\ 0.9900 & 1.0000 & 0.9100 & 0 & 0.8100 & 0 & 0.0900 & 0.1300 \\ 0.8300 & 0.9100 & 1.0000 & 0 & 0.9600 & 0 & 0 & 0 \\ 0.8500 & 0 & 0 & 1.0000 & 0.8900 & 0 & 0 & 0 \\ 0.9300 & 0.8100 & 0.9600 & 0.8900 & 1.0000 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.0000 & 0.7300 & 0.8400 \\ 0 & 0.0900 & 0 & 0 & 0 & 0.7300 & 1.0000 & 0.8900 \\ 0 & 0.1300 & 0 & 0 & 0 & 0.8400 & 0.8900 & 1.0000 \end{bmatrix}$$

(b) Coupling matrix of $\boldsymbol{X}^8$
(b) $\boldsymbol{X}^8$ 的耦合矩阵

Fig.3 Synthetic data of $\boldsymbol{X}^8$
图 3 合成数据集 $\boldsymbol{X}^8$

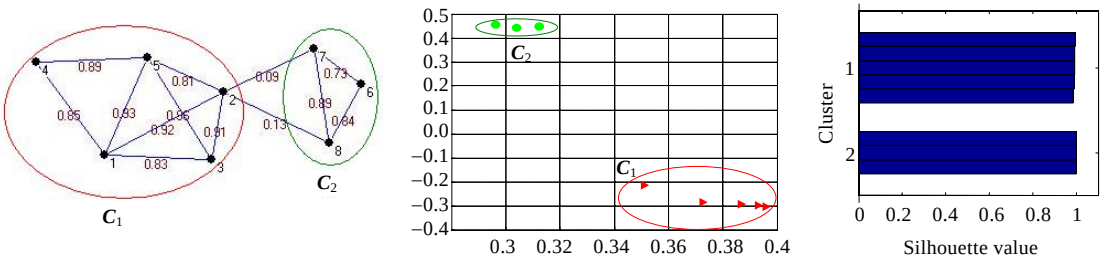
实验结果如图 4 所示.图 4(a)、图 4(d)是分别选取 k=2 和 k=3 时的聚类模型$\boldsymbol{\Delta}_2$ 和$\boldsymbol{\Delta}_3$,图 4(b)、图 4(e)分别是图 4(a)、图 4(d)的聚类模型在谱映射空间中的数据分布.从图中可以看出:(1) $CohGap(\boldsymbol{\Delta}_2)$=0.38,$\delta_2(\boldsymbol{P})$=0.6,$Quality(\boldsymbol{\Delta}_2)$=1.58.由于 k=2,所以谱映射空间是 2 维空间$\mathbb{R}^2$;(2) $CohGap(\boldsymbol{\Delta}_3)$=0.48,$\delta_3(\boldsymbol{P})$=0.3,$Quality(\boldsymbol{\Delta}_3)$=0.63.此时 k=3,所以谱映射空间是 3 维空间$\mathbb{R}^3$;(3) $\boldsymbol{\Delta}_2$ 在谱映射空间中簇间结点距离较$\boldsymbol{\Delta}_3$的要大,簇内结点距离较$\boldsymbol{\Delta}_3$的要小.

上述实验结果说明,$\boldsymbol{\Delta}_2$ 更接近完美聚类,这是因为聚类模型$\boldsymbol{\Delta}_2$ 的凝聚度间距较$\boldsymbol{\Delta}_3$ 要小,$\delta_2(\boldsymbol{P})$较$\delta_3(\boldsymbol{P})$要大,所以,$\boldsymbol{\Delta}_2$的聚类模型质量好于$\boldsymbol{\Delta}_3$($Quality(\boldsymbol{\Delta}_2)$=1.58>$Quality(\boldsymbol{\Delta}_2)$=0.63),因此根据公式(1),确定最优聚簇数为 k=2,最优聚类模型为$\boldsymbol{\Delta}_2$,实验结果符合预期.

图 4(c)、图 4(f)分别显示了$\boldsymbol{\Delta}_2$ 和$\boldsymbol{\Delta}_3$ 的 Silhouette 值.其中,纵坐标是聚簇号,横坐标是 Silhouette 值.从图 4(c)、图 4(f)中可以看出:$\boldsymbol{\Delta}_2$ 中每个聚簇中结点的 Silhouette 值都近似为 1,轮廓线较为整齐,聚类质量较好;而$\boldsymbol{\Delta}_3$ 中聚簇 $\boldsymbol{C}_1$ 中的结点的 Silhouette 值差异较大,聚类质量较差.这一结果与采用定义 13 的聚类模型质量评价方法得到的评价相一致.

实验结果表明:(1) 第 6 节中提出的聚类模型质量定义(定义 13)由于既考虑了矩阵 $\boldsymbol{P}$ 的特征间隙对聚类质量的影响,又考虑了簇间相似度(通过凝聚度)对聚类质量的影响,因而能够对聚类结果给出全面而有效的评价;(2) 由于第 6 节中给出的公式(1)以最大化聚类模型质量作为优化目标,通过求解公式(1)能够优化地确定聚

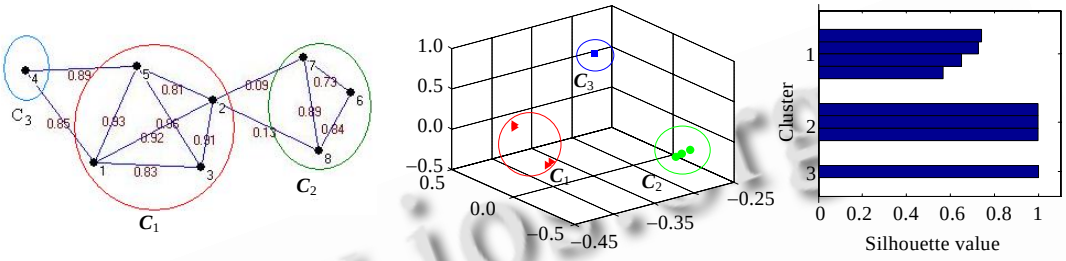
簇数 k,并通过 SCAM 算法找到最优的聚类模型.



(a) Optimal clustering model Δ_2 (b) Δ_2 in spectral mapping space (c) Silhouette values of Δ_2

(a) 最优聚类模型Δ_2 (b) 谱映射空间中的Δ_2 (c) Δ_2 的 Silhouette 值

$\Delta_2=\{C_1,C_2\}$, $C_1=\{1,2,3,4,5\}$, $C_2=\{6,7,8\}$, $CohGap(\Delta_2)=0.38$, $\delta_2(P)=0.6$, $Quality(\Delta_2)=1.58$



(d) Incurrately clustering model Δ_3 (e) Δ_3 in spectral mapping space (f) Silhouette values of Δ_3

(d) 不准确的聚类模型Δ_3 (e) 谱映射空间中的Δ_3 (f) Δ_3 的 Silhouette 值

$\Delta_3=\{C_1,C_2,C_3\}$, $C_1=\{1,2,3,5\}$, $C_2=\{6,7,8\}$, $C_3=\{4\}$, $CohGap(\Delta_3)=0.48$, $\delta_3(P)=0.3$, $Quality(\Delta_3)=0.63$

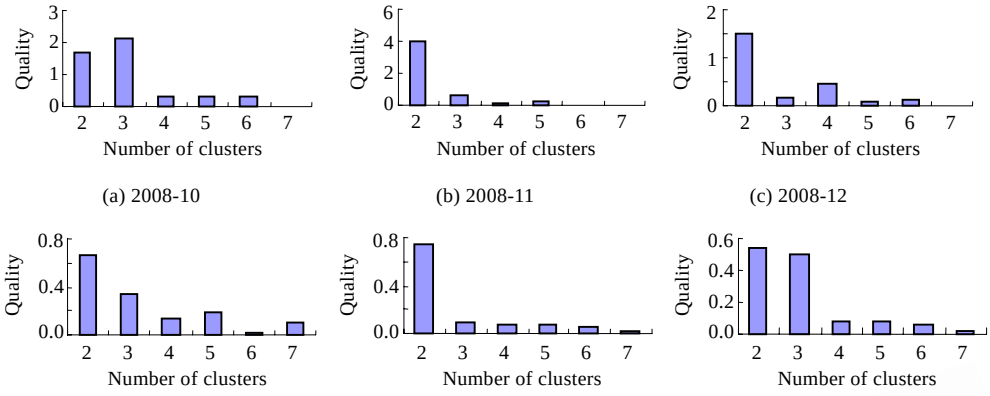
Fig.4 Effectiveness of SCAM

图 4 SCAM 的有效性

8.3 基于真实数据验证算法有效性

本节在真实数据集上验证 EEMA 算法的有效性,选取的数据集是上海证券交易所 2008 年 10 月~2009 年 3 月共 6 个月的数据.随机地从银行、保险和地产 3 个板块中选取 8 支股票,标记为 $x_i,i=1,\ldots,8$.其中,x_1,x_2,x_4,x_5 属于银行板块,x_3,x_7 属于保险板块,x_6,x_8 属于地产板块.实验预期为属于同一板块或相近板块(银行和保险板块)的数据流应当表现出耦合关系,同时,由于这 3 个板块都是相关板块,所以可能出现一个板块的某些数据流与另一板块表现出耦合关系.实验结果显示在图 5 和图 6 中.

图 5 显示了各月不同聚簇数 k 对应的聚类质量.可以看到,2008 年 10 月 k=3 时的聚类模型质量最优.此后,2008 年 11 月~2009 年 3 月最优聚簇数变为 k=2.同时,保险板块和银行板块的股票因为是金融相关的板块,所以表现出了较强的耦合关系.图 6 显示了各月的最优聚类模型在谱映射空间内的分布.可以看到:(1) 2008 年 10 月数据的最优聚簇数为 k=3,最优耦合聚类模型为Δ_3(2008-10)=$\{C_1=\{x_1,x_2,x_4,x_5\},C_2=\{x_3,x_7\},C_3=\{x_6,x_8\}\}$,符合实验数据分属 3 个不同板块的实验预期;(2) 2008 年 11 月~2009 年 3 月的最优聚簇数都是 k=2,但是最优耦合聚类模型在不断演化;(3) 2009 年 2 月和 3 月的最优耦合聚类模型与 2009 年 1 月保持一致;(4) 因为在 2008 年 11 月、2008 年 12 月和 2009 年 1 月聚类模型都较上一个月发生了变化,所以最终得到的演化事件点集合为 T={2008-11,2008-12,2009-1}.

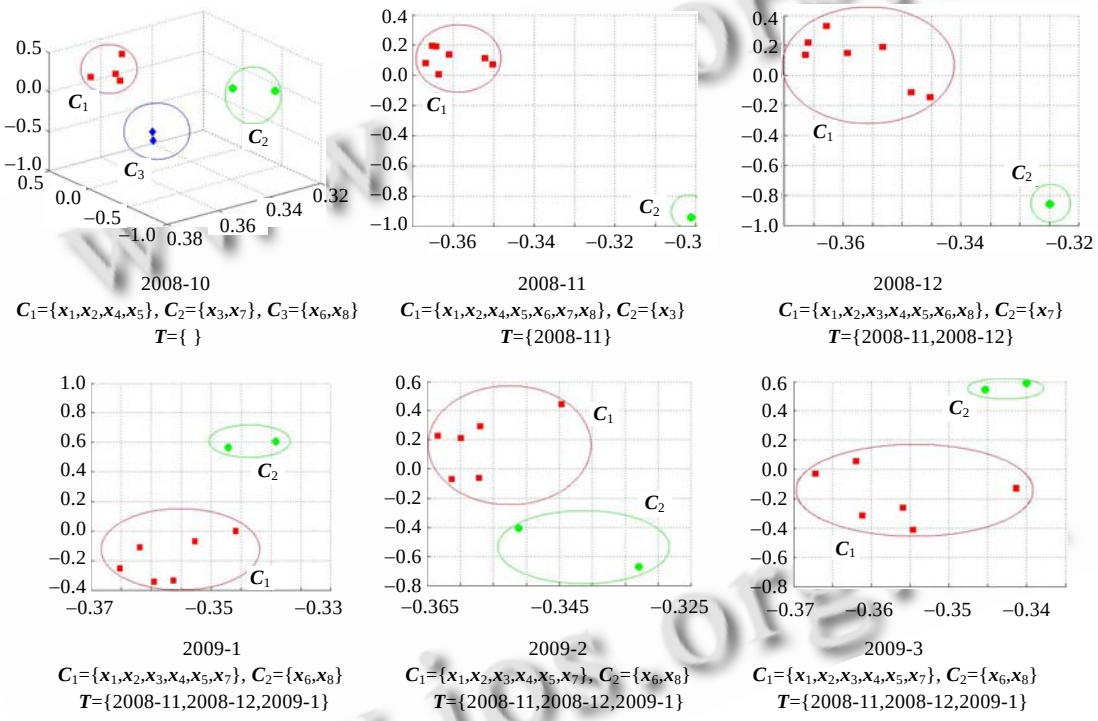


(a) 2008-10 (b) 2008-11 (c) 2008-12

(d) 2009-1 (e) 2009-2 (f) 2009-2

Fig.5 Clustering quality of different number of clusters

图 5 不同聚簇数时的聚类质量



2008-10
$C_1=\{x_1,x_2,x_4,x_5\}$, $C_2=\{x_3,x_7\}$, $C_3=\{x_6,x_8\}$
$T=\{\ \}$

2008-11
$C_1=\{x_1,x_2,x_4,x_5,x_6,x_7,x_8\}$, $C_2=\{x_3\}$
$T=\{2008\text{-}11\}$

2008-12
$C_1=\{x_1,x_2,x_3,x_4,x_5,x_6,x_8\}$, $C_2=\{x_7\}$
$T=\{2008\text{-}11,2008\text{-}12\}$

2009-1
$C_1=\{x_1,x_2,x_3,x_4,x_5,x_7\}$, $C_2=\{x_6,x_8\}$
$T=\{2008\text{-}11,2008\text{-}12,2009\text{-}1\}$

2009-2
$C_1=\{x_1,x_2,x_3,x_4,x_5,x_7\}$, $C_2=\{x_6,x_8\}$
$T=\{2008\text{-}11,2008\text{-}12,2009\text{-}1\}$

2009-3
$C_1=\{x_1,x_2,x_3,x_4,x_5,x_7\}$, $C_2=\{x_6,x_8\}$
$T=\{2008\text{-}11,2008\text{-}12,2009\text{-}1\}$

Fig.6 Data distribution in spectral mapping space

图 6 数据在谱映射空间内的分布

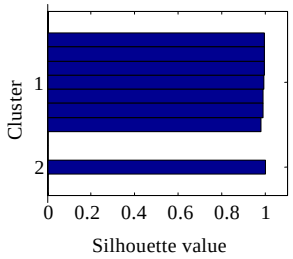
注意,虽然 2009 年 1 月到 3 月的最优聚类模型是一致的,但是由于数据流之间的耦合度在不断变化,所以每月的耦合矩阵Ω并不相同,因此在谱映射空间中的分布也不相同.实验结果表明,由于以最大化聚类模型质量为优化目标(公式(1)),EEMA 能够自动地确定每个月的最优聚簇数 k,并调用 SCAM 生成最优聚类模型Δ_k;同时,EEMA 能够根据聚类模型的变化,发现演化事件点.

8.4 与其他算法比较聚类效果

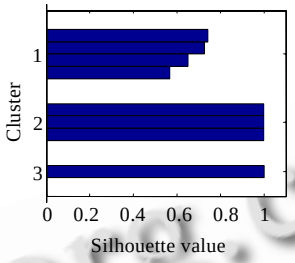
文献[20]提出了以最大化矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$ 特征间隙为优化目标确定聚簇数 k 的算法(下文称为 Eig);文献[22]

提出了以最小化多路规范化割为优化目标确定聚簇数 k 的算法(下文称为 MNCut).本节实验的目的是比较 EEMA 算法和上述两种算法的优化效果.

图 7 显示了 Eig 和 EEMA 在 2008 年 11 月的交易数据集上得到的最优聚类模型的 Silhouette 值.从图 7 可以看到,在 2008 年 11 月的数据上,Eig 得到的最优聚类模型Δ_3中聚簇 1 的结点的 Silhouette 值彼此差异较大,而 EEMA 算法得到的最优聚类模型Δ_2中簇内的结点的值都比较接近,表明Δ_2的质量优于Δ_3.其原因是,Eig 只考虑了特征间隙最大化($\delta_3(\boldsymbol{P})$=0.0673>$\delta_2(\boldsymbol{P})$=0.0526),所以得到的最优聚簇数为 k=3.但是 k=3 时的规范化割远大于 k=2 时的规范化割($CMNCut(\Delta_3)$=112.4348>$CMNCut(\Delta_2)$=14.1811),EEMA 算法综合考虑了两方面因素,得到的最优聚簇数为 k=2.根据定义 13,$Quality(\Delta_2)$=3.9470>$Quality(\Delta_3)$=0.6082,这与 Silhouette 值的评价结果相一致.



(a) Optimal clustering model Δ_2 derived from EEMA
(a) EEMA 得到的优化聚类模型Δ_2
$\delta_2(\boldsymbol{P})$=0.0526, $CMNCut(\Delta_2)$=14.1811, $Quality(\Delta_2)$=3.9470

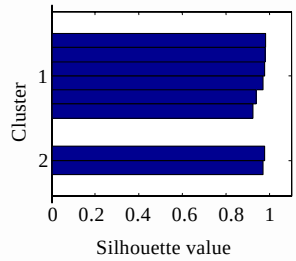


(b) Optimal clustering model Δ_3 derived from Eig
(b) Eig 得到的优化聚类模型Δ_3
$\delta_3(\boldsymbol{P})$=0.0673, $CMNCut(\Delta_3)$=112.4348, $Quality(\Delta_3)$=0.6082

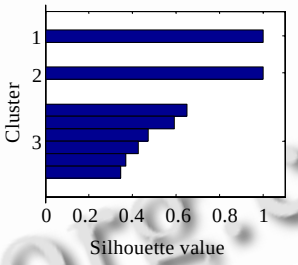
Fig.7 Effectiveness comparison between EEMA and Eig running on data of 2008-11

图 7 在 2008-11 数据集上比较 EEMA 和 Eig 算法的有效性

图 8 显示了 MNCut 和 EEMA 在 2009 年 3 月交易数据集上得到的最优聚类模型的 Silhouette 值.从图 8 可以看到,MNCut 算法得到的最优聚类模型Δ_3的质量明显劣于 EEMA 得到的Δ_2.这是因为 MNCut 只考虑规范化割的最小化($CMNCut(\Delta_3)$=26.8830<$CMNCut(\Delta_2)$=44.2303),没有考虑特征间隙的影响($\delta_2(\boldsymbol{P})$=0.0924>$\delta_3(\boldsymbol{P})$=0.0432).根据定义 13,$Quality(\Delta_2)$=2.1276>$Quality(\Delta_3)$=1.7230,这与 Silhouette 值的评价结果相一致.



(a) Optimal clustering model Δ_2 derived from EEMA
(a) EEMA 得到的优化聚类模型Δ_2
$\delta_2(\boldsymbol{P})$=0.0924, $CMNCut(\Delta_2)$=44.2303, $Quality(\Delta_2)$=2.1276



(b) Optimal clustering model Δ_3 derived from MNCut
(b) MNCut 得到的优化聚类模型Δ_3
$\delta_3(\boldsymbol{P})$=0.0432, $CMNCut(\Delta_3)$=26.8830, $Quality(\Delta_3)$=1.7230

Fig.8 Effectiveness comparison between EEMA and MNCut running on data of 2009-3

图 8 在 2009-3 数据集上比较 EEMA 和 MNCut 算法的有效性

上述实验结果表明,由于 EEMA 算法既考虑了矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$的特征间隙对聚类质量的影响,又考虑了规范化割对聚类质量的影响,而 Eig 和 MNCut 都只分别考虑了其中一个因素,因此当两个因素的影响彼此矛盾时,EEMA 算法能够得出质量更优的最优聚类模型.

8.5 性能和规模可伸缩性测试

从股票交易数据集中分别选取 50~1000(每次增加 50)支股票数据进行实验,分别测试 EEMA 和 O-EEMA 算法的 CPU 时间.测试平台的 CPU 为 Pentium Dual E2160,主频 1.8GHz;内存 2GB.测试结果如图 9 所示.

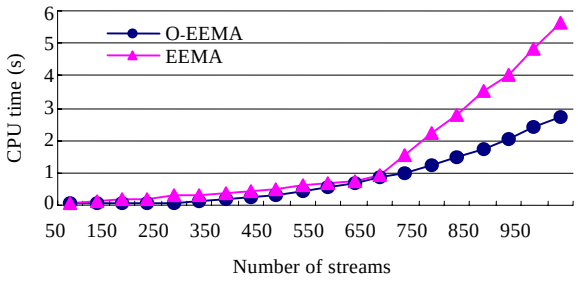


Fig.9 Performance of EEMA on data sets of different sizes

图 9 EEMA 在不同大小的数据集上的性能

从图 9 可以观察到:当数据流个数 n 较小时(n<600),EEMA 和 O-EEMA 算法的 CPU 时间曲线都随 n 的增大而平缓上升;但是当 n>600 时,EEMA 算法的 CPU 时间上升较快,而 O-EEMA 算法则体现出优势,CPU 时间仍然保持平缓上升.其原因正如第 6 节所述,是因为 O-EEMA 算法中对耦合矩阵的计算作了优化,仅在循环外计算 1 次,其时间复杂度为 $O(cn^2/2)$;而 EEMA 算法在 SCAM 算法中计算耦合矩阵,每次循环都要重复计算,其时间复杂度为 $O(n^3)$.

9 结论和展望

本文研究了应用耦合关系在多数据流之间实现聚类划分和挖掘演化事件的问题.在将多数据流转化为耦合图和耦合矩阵模型后,提出了多数据流上基于耦合度的谱聚类算法 SCAM 和演化事件挖掘算法 EEMA 及其优化实现 O-EEMA,时间复杂度为 $O(cn^2/2)$.算法有如下特点:(1) 具有坚实的数学基础,尤其是矩阵理论基础;(2) 自动优化确定聚簇数,不需要聚簇数的先验知识;(3) 聚类模型的生成和评价采用了一致的优化标准.实验结果表明:(1) 算法 EEMA 和 SCAM 能够挖掘多数据流随时间演化的聚类模型,并发现演化事件点;(2) 算法 EEMA 基于聚类模型质量最大化,实现聚类模型的优化和评价(包括自动选择聚簇数 k),与采用第三方度量标准 Silhouette 值的评价结果相一致;(3) 由于本文提出的优化标准既考虑特征间隙的影响,又考虑了规范化割的影响,因而能够比现有算法(Eig,MNCut)更加准确地确定聚簇数 k;(4) 当采用优化实现版本 O-EEMA 时,表现出了较好的规模可伸缩性.

由于新的应用出现了越来越多的在时间上异步、在空间上分散的异构和异质数据流,对于它们之间的演化聚类分析问题,则是进一步的研究方向.

References:

[1] Chuck C, Theodore J, Oliver S, Vladislav S. Gigascope: A stream database for network applications. In: Proc. of the ACM SIGMOD 2003. New York: ACM, 2003. 647−651. http://db.ucsd.edu/sigmodpods03

[2] Johannes G, Samuel M. Query processing in sensor networks. IEEE Pervasive Computing, 2004,3(1):46−55.

[3] Charu CA. A framework for diagnosing changes in evolving data streams. In: Proc. of the ACM SIGMOD 2003. New York: ACM, 2003. 575−586. http://db.ucsd.edu/sigmodpods03

[4] Yunyue Z, Dennis S. StatStream: Statistical monitoring of thousands of data streams in real time. In: Proc. of the 28th Int'l Conf. on Very Large Data Bases (VLDB 2002). New York: VLDB Endowment, 2002. 358−369. http://www.cse.ust.hk/vldb2002

[5] Yunyue Z, Dennis S. Efficient elastic burst detection in data streams. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD 2003). New York: ACM, 2003. 336−345. http://www.sigkdd.org/kdd2003

[6] Golab L, TamerOzsu M. Issues in data stream management. ACM SIGMOD Record, 2003,32(2):5−14. [doi: 10.1145/776985.776986]

[7] Guha S, Mishra N, Motwani R, O'Callaghan L. Clustering data streams: Theory and practice. IEEE Trans. on Knowledge and Data Engineering, 2003,15(3):515−528. [doi: 10.1109/TKDE.2003.1198387]

[8] O'Callaghan L, Mishra N, Meyerson A, Guha S, Motwani R. Streaming-Data algorithms for high-quality clustering. In: Proc. of the 18th Int'l Conf. on Data Engineering (ICDE 2008). Washington: IEEE Computer Society, 2002. 685−694. http://www.icde2008.org

[9] Aggarwal CC, Han J, Wang J, Yu PS. A framework for clustering evolving data streams. In: Johann CF, Peter CL, Serge A, Michael JC, Patricia GS, Andreas H, eds. Proc. of the 29th Int'l Conf. on Very Large Data Bases (VLDB 2003). New York: VLDB Endowment, 2003. 81−92.

[10] Deepayan C, Ravi K, Andrew T. Evolutionary clustering. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2006). New York: ACM, 2006. 554−560. http://www.kdd2006.com

[11] Zhou AY, Cao F, Yan Y, Sha CF, He XF. Distributed data stream clustering: A fast EM-based approach. In: Proc. of the IEEE 23rd Int'l Conf. on Data Engineering (ICDE 2007). Washington: IEEE Computer Society, 2007. 736−745. http: //www.srdc.metu.edu.tr/webpage/icde/index.php

[12] Papadimitriou S, Sun J, Faloutsos C. Streaming pattern discovery in multiple time-series. In: Proc. of the 3lst Int'l Conf. on Very Large Data Bases (VLDB 2005). New York: VLDB Endowment, 2005. 697−708. http://vldb.idi.ntnu.no

[13] Daniel K, Shai BD, Johannes G. Detecting changes in data streams. In: Proc. of the 30th Int'l Conf. on Very Large Data Bases (VLDB 2004). New York: VLDB Endowment, 2004. 180−191. http://www.vldb04.org

[14] Chen AL, Tang CJ, Yuan CA, Peng J, Hu JJ. An anti-noise algorithm for mining asynchronous coincidence pattern in multi-streams. Journal of Software, 2006,17(8):1753−1763 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1753.htm [doi: 10.1360/jos171753]

[15] Chen AL, Tang CJ, Yuan CA, Zhu MF, Duan L. A compression algorithm for multi-streams based on wavelets and coincidence. Journal of Software, 2007,18(2):177−184 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/177.htm [doi: 10.1360/jos180177]

[16] Fan C. Spectral Graph Theory. Providence: American Mathematical Society, 1997.

[17] Ulrike VL. A tutorial on spectral clustering. Statistics and Computing, 2007,17(4):395−416. [doi: 10.1007/s11222-007-9033-z]

[18] Lars H, Andrew BK. New spectral methods for ratio cut partitioning and clustering. IEEE Trans. on Computer-Aided Design, 1992, 11(9):1074−1085. [doi: 10.1109/43.159993]

[19] Jianbo S, Jitendra M. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(8):888−905. [doi: 10.1109/34.868688]

[20] Andrew YN, Micheal IJ, Yair W. On spectral clustering: Analysis and an algorithm. In: Proc. of the Nueral Information Processing System Conf. 2001 (NIPS 2001). Cambridge: MIT Press, 2001. 849−856. http://nips.cc/Conferences/2001

[21] Marina M, Jianbo S. A random walks view of spectral segmentation. In: Proc. of the 8th Int'l Workshop on Artificial Intelligence and Statistics. San Francisco: Morgan Kaufmann Publishers, 2001. 4−7. http://www.gatsby.ucl.ac.uk/aistats/aistats2001/index.html

[22] Marina M, Liang X. Multiway cuts and spectral clustering. Technical Report, 442, University of Washington, 2004.

[23] Chi Y, Song XD, Zhou DY, Hino K, Tseng BL. Evolutionary spectral clustering by incorporating temporal smoothness. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (SIGKDD 2007). New York: ACM, 2007. 153−162. http://www.sigkdd.org/kdd2007

[24] Richard TS. Temporal databases. IEEE Computer, 1986,19(9):35−72.

[25] Von Luxburg U, Belkin M, Bousquet O. Consistency of spectral clustering. The Annals of Statistics, 2008,36(2):555−586. [doi: 10.1214/009053607000000640]

[26] Kamvar SD, Klein D, Manning CD. Spectral learning. In: Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence. MA Cambridge: MIT Press, 2003. 561−566. http://ijcai.org/~ijcai03/1024/index.html

[27] Beresford NP. The QR algorithm. Comupting in Science & Engineering, 2000,2(1):38−42.

[28] Gilbert WS, Jiguang S. Matrix Perturbation Theory. Boston: Academic Press, 1990.

[29] Derek A, James MK. Recognizing falls from silhouette. In: Proc. of the 28th Annual Int'l Conf. of IEEE EMBS. New York, 2006. 6388−6391. http://embc2006.njit.edu

附中文参考文献:

[14] 陈安龙,唐常杰,元昌安,彭京,胡建军.挖掘多数据流的异步偶合模式的抗噪声算法.软件学报,2006,17(8):1753−1763 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1753.htm [doi: 10.1360/jos171753]

[15] 陈安龙,唐常杰,元昌安,朱明放,段磊.基于小波和偶合特征的多数据流压缩算法.软件学报,2007,18(2):177−184 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/18/177.htm [doi: 10.1360/jos180177]

附录. 命题 1~命题 7 和引理 5 的证明

命题 1. 算法 SCAM 的时间复杂度为 $O(n^2/2)$.

证明:运行时间主要消耗在耦合矩阵的构造(第(2)行)和 Laplacian 矩阵的特征分解(第(4)行)上.由于耦合矩

阵是实对称矩阵,所以:(1) 计算耦合矩阵的时间复杂度为 $O(n^2/2)$;(2) 采用 QR 算法实现特征分解的时间复杂度为 $O(n)$[27].所以,算法 SCAM 总的时间复杂度为 $O(n^2/2)$. □

命题 2. 如果 n 维向量集合 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$ 对于聚类 $\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$ 是分段常数向量,则定义 5 给出的谱映射 S 将同一个类中的所有点映射为 $\mathbb{R}^k$ 中的同一个点.

证明:由定义 5 可知,谱映射将结点 $i(i=1,\ldots,n)$ 映射到空间 $\mathbb{R}^k$ 中的一个点 $(v_{1i},v_{2i},\ldots,v_{ki})$,其中,$v_{ri}$ 是第 r 个向量的第 i 个分量.设 $\boldsymbol{x}_i,\boldsymbol{x}_j$ 属于同一个簇,又已知 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_k$ 对于聚类 $\boldsymbol{\Delta}_k$ 是分段常数向量,则由定义 5 有,$v_{1i}=v_{1j},v_{2i}=v_{2j},\ldots,v_{ki}=v_{kj}$,即 $(v_{1i},v_{2i},\ldots,v_{ki})=(v_{1j},v_{2j},\ldots,v_{kj})$.所以,$\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 被映射到了 $\mathbb{R}^k$ 中的同一个点,命题成立. □

命题 3. 算法 SCAM 求出的聚类模型 $\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$ 是完美聚类模型当且仅当矩阵 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$ 对于 $\boldsymbol{\Delta}_k$ 是块随机矩阵.

证明:由引理 3 和命题 2 立即可得. □

命题 4. 设 $\{\lambda_i\}_{i=1,\ldots,k+1}$ 是矩阵 $\boldsymbol{P}$ 的前 $k+1$ 个最大特征值且 $\lambda_i>\lambda_{i+1},\boldsymbol{P}=\boldsymbol{P}^*+\boldsymbol{H}$,如果扰动 $\boldsymbol{H}$ 一定时,则当 $\lambda_k-\lambda_{k+1}$ 越大时,$\boldsymbol{P}$ 在渐进意义上越接近块随机矩阵 $\boldsymbol{P}^*$.

证明:取实数区间 $W=[0,\lambda_k]$,显然有 $\lambda_k\in W,\lambda_{k+1}\notin W$,由已知 $\lambda_k>\lambda_{k+1}$,所以 $\lambda_k-\lambda_{k+1}$ 满足定义 9 中的条件(1),即 $\delta(\boldsymbol{P})=\lambda_k-\lambda_{k+1}$.由引理 4,$d(\boldsymbol{P},\boldsymbol{P}^*)\leqslant\|\boldsymbol{H}\|/(\lambda_k-\lambda_{k+1})$,即 $\|\boldsymbol{H}\|/(\lambda_k-\lambda_{k+1})$ 是 $d(\boldsymbol{P},\boldsymbol{P}^*)$ 的上界.故当 $\boldsymbol{H}$ 一定时,$\lambda_k-\lambda_{k+1}$ 越大,$d(\boldsymbol{P},\boldsymbol{P}^*)$ 越小,$\boldsymbol{P}$ 在渐进意义上越接近块随机矩阵 $\boldsymbol{P}^*$. □

命题 5. 如果经算法 SCAM 求出的聚类模型是完美聚类模型 $\boldsymbol{\Delta}_k^*$,则其凝聚度为 $Coh(\boldsymbol{\Delta}_k^*)=1\Big/\sum_{i=1}^k\mu_i$,其中,$\mu_1<\mu_2<\ldots<\mu_k$ 是 Laplacian 矩阵 $\boldsymbol{L}$ 的前 k 个最小特征值.

证明:

因为 $vol(\boldsymbol{C}_i)=\sum_{j=1}^k Cut(\boldsymbol{C}_i,\boldsymbol{C}_j)=\sum_{j=1}^{i-1}Cut(\boldsymbol{C}_i,\boldsymbol{C}_j)+\sum_{j=i+1}^k Cut(\boldsymbol{C}_i,\boldsymbol{C}_j)+Cut(\boldsymbol{C}_i,\boldsymbol{C}_i)$,

所以 $CMNCut(\boldsymbol{\Delta}_k)=\sum_{i=1}^k[vol(\boldsymbol{C}_i)-Cut(\boldsymbol{C}_i,\boldsymbol{C}_i)]/vol(\boldsymbol{C}_i)=\sum_{i=1}^k(1-\sum_{u,v\in\boldsymbol{C}_i}\omega_{uv}\Big/\sum_{j\in\boldsymbol{C}_i}d_j)$.

根据定义 7,$r_{ii}=\sum_{u,v\in\boldsymbol{C}_i}\omega_{uv}\Big/\sum_{j\in\boldsymbol{C}_i}d_j$ 正是随机游走停留在聚簇 $\boldsymbol{C}_i$ 中的概率,所以,

$$CMNCut(\boldsymbol{\Delta}_k^*)=\sum_{i=1}^k(1-r_{ii})=k-trace(\boldsymbol{R})=k-\sum_{i=1}^k\gamma_i,$$

其中,$\{\gamma_i\}_{i=1,\ldots,k}$ 是定义 7 中的矩阵 $\boldsymbol{R}$ 的前 k 个最大特征值.设 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$ 的前 k 个最大特征值为 $\{\lambda_i\}_{i=1,\ldots,k}$,则由引理 2,$\lambda_i=\gamma_i$,所以 $CMNCut(\boldsymbol{\Delta}_k^*)=k-\sum_{i=1}^k\lambda_i$.根据引理 1,$\mu_i=1-\lambda_i$,所以 $Coh(\boldsymbol{\Delta}_k^*)=1\Big/\sum_{i=1}^k\mu_i$. □

引理 5. 设矩阵 $\boldsymbol{L}$ 是实对称矩阵,其特征值从小到大为 $\mu_1<\mu_2<\ldots<\mu_n$,对应的标准化特征向量为 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$,如果向量 $\boldsymbol{y}_1,\boldsymbol{y}_2,\ldots,\boldsymbol{y}_k$ 是 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$ 的线性组合,且 $\boldsymbol{y}_i^T\boldsymbol{y}_i=1,\boldsymbol{y}_i^T\boldsymbol{y}_j=0,i\neq j$,则 $\min(\sum_{i=1}^k\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i)=\sum_{i=1}^k\mu_i$.

证明:由已知有

$$\boldsymbol{y}_i=\sum_{u=1}^n a_{ui}\boldsymbol{v}_u,i=1,\ldots,k \tag{2}$$

设 $n\times k$ 矩阵 $\boldsymbol{A}=(a_{ui}),u=1,\ldots,n,i=1,\ldots,k$.由已知有,$\boldsymbol{v}_i^T\boldsymbol{v}_j=0,\boldsymbol{v}_i^T\boldsymbol{v}_i=1,i\neq j$,所以,

$$\boldsymbol{y}_i^T\boldsymbol{y}_i=\boldsymbol{a}_{\cdot i}^T\boldsymbol{a}_{\cdot i}=1 \text{ 且 } \boldsymbol{y}_i^T\boldsymbol{y}_j=\sum_{u=1}^n a_{ui}\boldsymbol{v}_u^T\sum_{u=1}^n a_{uj}\boldsymbol{v}_u=\boldsymbol{a}_{\cdot i}^T\boldsymbol{a}_{\cdot j}=0.$$

即,$\boldsymbol{A}$ 的列向量 $\boldsymbol{a}_{\cdot i}$ 是标准正交向量.由拉格朗日乘子法得

$$F=\sum_{i=1}^k\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i+\sum_{i=1}^k\sum_{j=i+1}^k\beta_{ij}\boldsymbol{y}_i^T\boldsymbol{y}_j+\sum_{i=1}^k\beta_{ii}(\boldsymbol{y}_i^T\boldsymbol{y}_i-1) \tag{3}$$

将公式(2)代入公式(3),并注意到 $\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i=\sum_{u=1}^n\mu_u a_{ui}^2$,得

$$F=\sum_{i=1}^k\sum_{u=1}^n\mu_u a_{ui}^2+\sum_{i=1}^k\sum_{j=i+1}^k\beta_{ij}\boldsymbol{a}_{\cdot i}^T\boldsymbol{a}_{\cdot j}+\sum_{i=1}^k\beta_{ii}(\boldsymbol{a}_{\cdot i}^T\boldsymbol{a}_{\cdot i}-1).$$

对 a_{ui} 求偏导,得

$$\frac{\delta F}{\delta a_{ui}}=2\mu_u a_{ui}+\beta_{ui}\sum_{l\neq u}a_{ul}+2\beta_{ii}a_{ui} \tag{4}$$

令公式(4)等于 0,得到 $\frac{\beta_{ui}}{2}\sum_{l\neq u}a_{ul}+\beta_{ii}a_{ui}=-\mu_u a_{ui}$.令 $k\times k$ 矩阵 $\boldsymbol{B}=(b_{li}),l,i=1,\ldots,k$,当 $l\neq i$ 时,$b_{li}=b_{il}=\beta_{il}/2$,当 $l=i$

时,$b_{ii}=\beta_{ii}$.于是 $\boldsymbol{B}\boldsymbol{a}_{u.}=-\mu_u\boldsymbol{a}_{u.}$,$u=1,\ldots,n$,其中,$\boldsymbol{a}_u$ 是 $\boldsymbol{A}$ 的行向量.这说明 $\boldsymbol{a}_u$ 只可能是 $\boldsymbol{B}$ 的特征向量或者零向量.注意,由于矩阵 $\boldsymbol{B}$ 最多有 k 个线性无关的特征向量且 $\boldsymbol{B}$ 是对称矩阵,所以 $\boldsymbol{A}$ 最多有 k 个正交的非零行向量.又因为 $\boldsymbol{A}$ 的 k 个列向量也是正交的,所以,$\boldsymbol{A}$ 至少有 k 个非零正交行向量,所以 $\boldsymbol{A}$ 恰有 k 个非零正交行向量.又因为 $\boldsymbol{A}$ 的列向量是标准正交向量,所以 $\boldsymbol{A}$ 的 k 个非零正交行向量也是标准的,即 $\boldsymbol{a}_{u.}^T\boldsymbol{a}_{u.}=1$.

记 $\boldsymbol{A}$ 的 k 个非零行向量的索引为 $j_1,j_2,\ldots,j_k$,则

$$\sum_{i=1}^{k}\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i=\sum_{i=1}^{k}\sum_{u=1}^{n}\mu_u a_{ui}^2=\sum_{i=1}^{k}\sum_{u=j_1}^{j_k}\mu_u a_{ui}^2=\sum_{u=j_1}^{j_k}\mu_u\boldsymbol{a}_{u.}^T\boldsymbol{a}_{u.}=\sum_{u=j_1}^{j_k}\mu_u.$$

所以,$\sum_{i=1}^{k}\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i$ 的最小值为最小的 k 个 μ_u 之和,即 $\min(\sum_{i=1}^{k}\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i)=\sum_{i=1}^{k}\mu_i$. □

命题 6. 已知 n-数据流集合 $\boldsymbol{X}^n=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$的耦合图 $\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$,$\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{\Omega}$,Laplacian 矩阵 $\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{\Omega}\boldsymbol{D}^{-1/2}$ 和聚簇数 k,$\mu_1<\mu_2<\ldots<\mu_n$ 是 $\boldsymbol{L}$ 的前 k 个最小特征值,且对应的标准化特征向量为 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$,由 SCAM 算法得到聚类格局 $\boldsymbol{\Delta}_k=\{\boldsymbol{C}_1,\boldsymbol{C}_2,\ldots,\boldsymbol{C}_k\}$,则 $Coh(\boldsymbol{\Delta}_k)\leqslant Coh(\boldsymbol{\Delta}_k^*)$.

证明:在命题 5 的证明中已经知道 $CMNCut(\boldsymbol{\Delta}_k)=k-\sum_{i=1}^{k}\sum_{u,v\in\boldsymbol{C}_i}\omega_{uv}\Big/\sum_{j\in\boldsymbol{C}_i}d_j$,对每个聚簇 $\boldsymbol{C}_i$ 引入标准化的 n 维标志向量 $\boldsymbol{\xi}_i=(\xi_{i1},\ldots,\xi_{in})^T$,若 $\boldsymbol{x}_j\in\boldsymbol{C}_i$,则 $\xi_{ij}=1$,否则 $\xi_{ij}=0$,则有

$$\sum_{j\in\boldsymbol{C}_i}d_j=\sum_{j\in\boldsymbol{V}}\xi_{ij}^2d_j,\sum_{u,v\in\boldsymbol{C}_i}\omega_{uv}=\sum_{u,v\in\boldsymbol{V}}\xi_{iu}\xi_{iv}\omega_{uv}=\sum_{j\in\boldsymbol{V}}\xi_{ij}^2d_j-\sum_{e_{uv}\in\boldsymbol{E}}(\xi_{iu}-\xi_{iv})^2\omega_{uv}.$$

所以,$CMNCut(\boldsymbol{\Delta}_k)=\sum_{i=1}^{k}\dfrac{\sum_{e_{uv}\in\boldsymbol{E}}(\xi_{iu}-\xi_{iv})^2\omega_{uv}}{\sum_{j\in\boldsymbol{V}}\xi_{ij}^2d_j}$,令 $\boldsymbol{y}_i=\boldsymbol{D}^{1/2}\boldsymbol{\xi}_i$,则 $CMNCut(\boldsymbol{\Delta}_k)=\sum_{i=1}^{k}\dfrac{\boldsymbol{y}_i^T\boldsymbol{L}\boldsymbol{y}_i}{\boldsymbol{y}_i^T\boldsymbol{y}_i}$.又 $\boldsymbol{\xi}_i$ 是 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$ 的线性组合[17],$\boldsymbol{y}_i$ 也是 $\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n$ 的线性组合,由引理 5 可知,$\min CMNCut(\boldsymbol{\Delta}_k)=\sum_{i=1}^{k}\mu_i$,所以 $Coh(\boldsymbol{\Delta}_k)\leqslant 1\Big/\sum_{i=1}^{k}\mu_i$,由命题 5 可得,$Coh(\boldsymbol{\Delta}_k^*)=1\Big/\sum_{i=1}^{k}\mu_i$,所以 $Coh(\boldsymbol{\Delta}_k)\leqslant Coh(\boldsymbol{\Delta}_k^*)$. □

命题 7. 算法 EMMA 的时间复杂度为 $O(n^3/2)$,算法 O-EMMA 的时间复杂度为 $O(cn^2/2)$,其中,c 为常数 $\lceil(e-s)/w\rceil$.

证明:算法 3 的主要操作为调用算法 SCAM(第(5)行)和计算聚类模型质量(第(6)行).根据定义 9~定义 12 可知,聚类模型质量的计算依赖于耦合矩阵 $\boldsymbol{\Omega}$,而计算 $\boldsymbol{\Omega}$ 的时间复杂度为 $O(n^2/2)$,则不优化时 EEMA 算法的时间复杂度为 $O(n^3/2)$.优化后,由于仅在循环外计算 1 次 $\boldsymbol{\Omega}$,算法 SCAM 的时间复杂度降为 $O(n)$,则优化后的算法 O-EMMA 的时间复杂度为 $O(cn^2/2)$,常量 c 是循环次数,由算法 2 第(4)行可知 $c=\lceil(e-s)/w\rceil$. □

杨宁(1974－),男,四川成都人,博士,讲师,CCF 会员,主要研究领域为机器学习,数据挖掘.

陈瑜(1974－),男,博士,讲师,主要研究领域为数据挖掘,计算智能.

唐常杰(1946－),男,教授,博士生导师,CCF 高级会员,主要研究领域为数据库系统,数据挖掘.

郑皎凌(1981－),女,博士,讲师,主要研究领域为数据库系统,数据挖掘.

王悦(1981－),男,博士,主要研究领域为数据库系统,数据挖掘.