一种基于超图 Markov 链松弛的聚类学习方法

郭 鹏[1,2] 李仁发[1] 胡 慧[3]

(湖南大学信息科学与工程学院 长沙 410082)[1] (湖南工程学院计算机与通信学院 湖南 湘潭 411104)[2]

(湖南工程学院电气信息学院 湖南 湘潭 411104)[3]

摘 要 将车联网中高维的时空特征嵌入到低维的特征语义词袋是一种典型的聚类问题。谱聚类因其计算简单且有全局最优解的特点而备受关注,但是关于其聚类数目的研究工作相对较少。针对传统 eigengap 启发式方法无法适应于多噪声点和边界模糊数据集,导致聚簇过度分割的问题,提出了一种基于超图 Markov 链松弛的聚类学习方法(HS-MR 算法)。该算法的基本思想是用 Markov 过程形式化描述超图并开始随机游走。在超图 Markov 链松弛过程中,通过随机转移矩阵 P 的 t 次幂和扩散映射找到数据集有意义的几何分布,然后提出基于互信息的目标函数进行聚类数目的自动收敛。实验结果表明,该算法在准确率上优于简单图谱聚类算法和标准超图谱聚类算法。

关键词 超图 Laplacian,Markov 链松弛,扩散映射,互信息

中图法分类号 TP391 **文献标识码** A

Clustering Method Based on Hypergraph Morkov Relaxation

GUO Peng[1,2] LI Ren-fa[1] HU Hui[3]

(College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China)[1]

(College of Computer and Communication, Hunan Institute of Engineering, Xiangtan, Hunan 411104, China)[2]

(School of Electrical and Information, Hunan Institue of Engineering, Xiangtan, Hunan 411104, China)[3]

Abstract How to embed high dimention spatial-temporal feature into low dimention semantic feature word bag is a typic clustering problem in the Internet of vehicle . Spectral clustering algorithm is recently focused because of its simple computing and global optimal solution, however, the research about the numbers of clusters is relatively little. Trandi-tional eigengap heuristic method works well if the clusters in the data are very well pronounced. However, the more noisy or overlapping the clusters are, the less effective this heuristic is. This paper proposed a clustering method based on hypergraph markov relaxation (HS-MR method). The basic idea of this algorithm is using the Markov process to formally describe hypergraph and start random walk. In the relaxation process of hypergraph Markov chain, meaningful geometric distribution of data set is found through tth power of random transfer matrix P and diffusion mapping. Then, the objective function based on mutual information is proposed to automatically converge the clustering number. Finally, the experimental results show that the algorithm is superior to simple graph spectral clustering algorithm and hyperg-raph spectral clustering algorithm in accuracy rate.

Keywords Hypergraph laplacian, Markov relaxation, Diffusion map, Mutual informaion

1 引言

与物联网不同,车联网的数据蕴含丰富的时间和空间特征,随着车联网数据呈指数级的增长,对有效地提取并语义表示这些特征提出了一个挑战。运用类似文本分类词袋法的机制,建立时空特征与有限语义特征词袋之间的联系成为了一个可行的方案。文献[1-2]已经证明了该方法不仅计算简单而且性能较好。因此,车联网时空语义特征的学习问题很自然地被归约成了机器无监督学习的聚类问题,即如何将高维的时空特征嵌入到低维的语义特征词袋。

对聚类算法和流形学习的研究已经做了大量的工作,其可分为线性聚类和非线性聚类,文献[3-4]中经典的 K-means 聚类算法是一种基于欧拉距离的线性划分算法,适合凸球形的样本空间。文献[5]中的 DBSCAN 算法是对 K-means 算法的改进,考虑了数据密度对聚类的影响,在处理空间数据时具有快速、有效处理噪声点和发现任意形状的优点,但需要较多的内存和 I/O 开销。文献[6-8]提出了一种基于图 Laplacian 的谱聚类算法,通过计算相似度矩阵的特征值和特征向量,K 个分区最小归一化割可由非零的 K 个特征值获得,该算法的复杂度低,并能收敛于全局最优。文献[9]提出的超图谱聚类归一化割思想,是超图在谱聚类算法上的扩展。文献[10-12]给出了超图 Laplacian 在半监督学习中的应

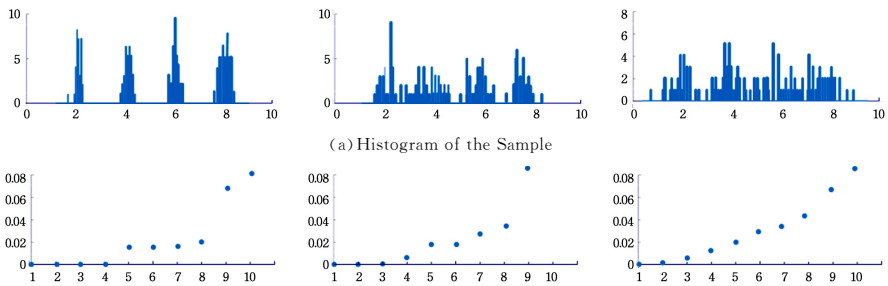
本文受湘潭市科技计划一般项目(GXY-YB 20171004)资助。

郭 鹏(1978—),男,博士生,副教授,主要研究方向为机器学习、边缘计算;李仁发(1957—),男,教授,主要研究方向为嵌入式系统、并行计算,E-mail:da_peng219@126.com(通信作者);胡 慧(1979—),女,博士,副教授,主要研究方向为计算机控制、模糊控制、神经网络控制和非线性系统控制等。

用。文献[13]和文献[14]分别在语义特征降维和人体行为识别中运用了 Markov 随机游走模型和扩散映射。但是，在上述研究中，聚类算法的聚类数目或是人工指定或是基于 eigengap 启发方法[15]，如何自动确定聚类数目成为需要解决的一个关键问题。

2　问题描述

特征语义词袋的粒度对特征提取的性能有着关键性的影响[13]，因为在高维时空特征到低维语义特征的非线性降维过程中，将伴随着信息的损失和噪声干扰，在谱聚类算法中普遍采用 eigengap 启发式的方法来确定聚类数目，该方法对界限明显的数据集的效果较好，但无法适应模糊聚类的情况，例如图 1 中，由左边图的数据集和特征值可知，第 4 和第 5 个特征值之间有明显的 gap，因此最小的 4 个非零特征值以及对应的特征向量能很容易地将数据集划分成 4 个分离的簇，而在右图中，数据集边界模糊，特征值无明显的 gap，从而无法确定其最优的聚类数目。本文提出的解决方案是将超图相似度矩阵转化为 Markov 随机转移矩阵，在 Markov 链随机游走（称之为 Markov 松弛）过程当中，通过衡量互信息损失，来收敛到最优的聚类数目。



(a) Histogram of the Sample

(b) Eigenvalues

图 1　采样数据分布与特征值的关系

3　HS-MR 算法

3.1　超图的构建

在一个简单图中，边常用于表示成对顶点之间的相似度，无法捕捉高序的信息，超图作为简单图的扩展，超边可以包含任意多个顶点，具有简单图不可比拟的优势，不仅能够表示多个顶点的组信息，而且能够表现多样化属性信息，适应于多传感器的特征提取。在超图的形式化描述中，$G=(V,E,W)$。其中，$V=\{v_1,v_2,\cdots,v_n\}$ 表示 n 个顶点；$E=\{e_1,e_2,\cdots,e_m\}$ 表示 m 条超边，每条超边是一个 V 的顶点子集；$W(e)$ 表示超边 e 的权值。不失一般性，可以运用关联距阵 H 描述超图 G，其中：

$$H(v,e)=\begin{cases}1, & \text{if } dv\in e\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

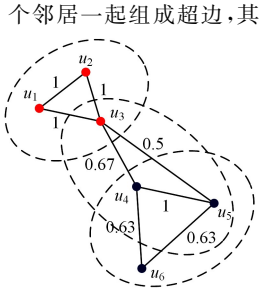
对于 $\boldsymbol{V}$ 中的顶点（$v\in V$），v 的度为：

$$d(v)=\sum_{e\in E}w(e)H(v,e) \tag{2}$$

超边 $e\in E$ 的度为：

$$\delta(e)=\sum_{v\in V}H(v,e) \tag{3}$$

本文用 D_v，D_e，W 分别表示顶点度矩阵、超边度矩阵和超边权值的对角矩阵。在超图构建时，把每一个传感器节点作为超图的一个顶点，然后分别以每个顶点为中心，使用高斯测度函数比较顶点间的相似度值，其 K 个最近邻居（K-NNs）形成一条超边。如图 2 所示，v_1 到 v_6 这 6 个顶点分别和最近 2 个邻居一起组成超边，其中超边 e_1，e_2，e_3 是相同的。



	e_1	e_2	e_3	e_4	e_5	e_6
u_1	1	1	1	0	0	0
u_2	1	1	1	0	0	0
u_3	1	1	1	0	0	0
u_4	0	0	0	1	1	1
u_5	0	0	0	1	1	1
u_6	0	0	0	1	1	1

图 2　由 6 个顶点组成的超图和关联矩阵

3.2　超图 Laplacian

超图 Laplacian 揭示了图的谱属性，即相似度矩阵的特征值和特征向量属性，谱聚类算法从中选择合适的特征向量进行聚类，从而将聚类问题转化为图的最优划分问题。超图 Laplacian 的计算可以分为两类，第一类是将超图扩展成一个简单图（Clique Expansion），然后计算简单图 Laplacian 算子的特征值和特征向量，另一类是定义一个类似的超图 Laplacian 算子。两者在本质上是一致的，本文使用后者，超图 Laplacian 算子的定义如下：

$$\Delta=I-D_v^{-\frac{1}{2}}HWD_e^{-1}H^{\mathrm{T}}D_{v^2}^{-1} \tag{4}$$

式(4)是一个半正定矩阵，最小特征值为 0，对应的特征向量为$\sqrt{d}$。可以证明，超图 Laplacian 与图 Laplacian 是一致的[16]，当 $\delta(e)=2$，$D_e=2I$ 时，超图 Laplacian 就是简单图的标准 Laplacian。

$$\begin{aligned}\Delta&=I-D_v^{-\frac{1}{2}}HWD_e^{-1}H^{\mathrm{T}}D_{v^2}^{-1}\\&=I-\frac{1}{2}D_v^{-\frac{1}{2}}(D_v+A)D_{v^2}^{-1}\\&=\frac{1}{2}(I-D_v^{-\frac{1}{2}}AD_{v^2}^{-1})\end{aligned}$$

3.3　归一化割和 Markov 随机矩阵

$$VolS=\sum_{v\in S}d(v)$$

$$Vol\partial S=\sum_{e\in\partial S}w(e)\frac{|e\cap S||e\cap\partial S|}{\delta(e)}$$

$$NCut(S,S^c)=Vol\partial S(1/VolS+1/VolS^c) \tag{5}$$

归一化割的思想基于这样一个基础：K 个不相交的聚簇之间连接的边的相似度最小，而同一聚簇内部连接的边的相似度最大。式(5)给出了这个目标的形式化描述，V 的一个顶点子集为 $S(S\in V)$，S^c 是 S 的补集（$S\cup S^c=V$），超图 G 的一个割就是将 V 分割成 S 和 S^c 的超边集合，将这个超边集定义为$\partial S=\{e\in E\mid e\cap S\neq\varphi,e\cap S^c\neq\varphi\}$，同时定义一个卷积 $VolS$，表示 S 中所有顶点的度的和，那么归一化割的目标是找到一个最小超边集，最小化式(5)是一个 NP 难问题，可以

将它松弛为一个实数值最优化问题，在 $\|f\|=1,\langle f,\sqrt{d}\rangle\geqslant 0$ 的约束条件下，其最近似最优解就是式(4)的次小特征值和特征向量，而超图的多路分割最优解就是 Δ 的 K 个非零特征值和特征向量。同时，如果相似度矩阵用随机转移矩阵 $P=D_v^{-1}HWD_e^{-1}H^{\mathrm{T}}$ 表示，$p(u,v)$ 表示从顶点 u 到顶点 v 的一步转移概率，从而谱聚类方法给出了一个概率的解释。下面从 Markov 随机游走的角度出发来解释归一化割。

$$p(u,v)=\sum_{e\in E}w(e)\frac{H(u,e)H(v,e)}{d(u)\delta(e)}\pi(v)=\frac{d(v)}{VolV}$$

$$\begin{aligned}&\sum_{u\in V}\pi(u)p(u,v)\\&=\sum_{u\in V}\frac{d(u)}{VolV}\sum_{e\in E}\frac{w(e)H(u,e)H(v,e)}{d(u)\delta(e)}\\&=\frac{1}{VolV}\sum_{u\in V}\sum_{e\in E}\frac{w(e)H(u,e)H(v,e)}{\delta(e)}\\&=\frac{1}{VolV}\sum_{e\in E}w(e)\sum_{u\in V}\frac{H(u,e)H(v,e)}{\delta(e)}\\&=\frac{1}{VolV}\sum_{e\in E}w(e)H(v,e)\\&=\frac{d(v)}{VolV}=\pi(v)\end{aligned}\tag{6}$$

$$NCut(S,S^c)=\frac{Vol\partial S}{VolV}\left(\frac{1}{VolS/VolV}+\frac{1}{VolS^c/VolV}\right)\tag{7}$$

$$\frac{VolS}{VolV}=\sum_{v\in S}\frac{d(v)}{VolV}=\sum_{v\in S}\pi(v)\tag{8}$$

$$\begin{aligned}\frac{Vol\partial S}{VolV}&=\sum_{e\in\partial S}\frac{w(e)}{VolV}\frac{|e\bigcap S||e\bigcap\partial S|}{\delta(e)}\\&=\sum_{e\in\partial S}\sum_{u\in e\bigcap S}\sum_{v\in e\bigcap\partial S}\frac{w(e)d(u)}{VolV}\frac{H(u,e)H(v,e)}{d(u)\delta(e)}\\&=\sum_{u\in S}\sum_{v\in S^c}\frac{d(u)}{VolV}\sum_{e\in\partial S}w(e)\frac{H(u,e)H(v,e)}{d(u)\delta(e)}\\&=\sum_{u\in S}\sum_{v\in S^c}\pi(u)p(u,v)\end{aligned}\tag{9}$$

式(6)－式(9)证明了从随机游走出发的最小割就是，使得不同聚簇之间顶点的转移概率很低，而同一聚簇内的顶点趋向于同一个亚稳态分布的一个分区。超图随机转移矩阵 P 与标准 Laplacian 算子 Δ 是一致的，如果 λ 和 V 分别是 Δ 的特征值和特征向量，那么 $1-\lambda$ 和 V 就分别是 P 的特征值和特征向量[17]。

3.4 Markov 松弛和扩散映射

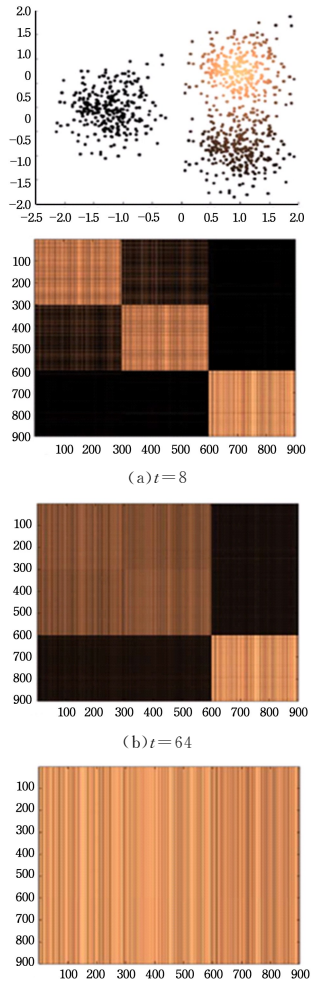
超图随机转移矩阵 P 的 t 次迭代称为 Markov 随机游走或 Markov 松弛过程。Markov 松弛的重要属性之一就是传导性(conductance)，当 Markov 链向前游走 t 步时，$p_{i,j}^t$ 表示从 i 点经过 t 步到 j 点的概率，初始分布中原来没有连接的两点经过很多短路径连接在了一起，即 i 点可以从其他点间接转移到 j 点。相比直接对随机转移矩阵进行谱聚类，Markov 松弛过程的谱聚类能够保持几何结构的全局属性。

Markov 松弛过程中的谱属性可以通过扩散映射获得[18-19]。作为一种流形降维工具，经过随机矩阵迭代，扩散映射能发现不同尺度 t 的特征值和特征向量，还能发现不同尺度下相关的几何结构，图 3 给出了 3 个聚簇构成的 900 个数据点，当 $t=8$ 时，数据集很明显是由 3 个独立的聚簇组成，当 $t=64$ 时，两个较近的聚簇合并，数据集分成了两个独立的聚簇，最后，当 $t=1024$ 时，所有的聚簇合并成了一个。这是由于当 Markov 松弛时，数据点局部向高密度概率点汇集，随后在一定的 t 内，p^t 的某些行具有近似的条件概率分布，数据点保持在一个相对稳定的亚稳状态。扩散映射的另一个贡献是给出了将非线性高维向量映射到低维的欧拉空间的方法，在低维欧拉空间中，点 y_1 和 y_2 之间的距离用扩散距离 $D_t(y_1,y_2)$ 表示。

$$S(\sigma_1,t)=\max\{1\in N,|\lambda_l|^t>\sigma_1|\lambda_1|^t\}$$

$$D_t(y_1,y_2)=\left(\sum_{l\geqslant 1}^{s(\sigma_1,t)}\lambda_l^{2t}(\phi_l(y_1)-\phi_l(y_2))^2\right)^{1/2}\tag{10}$$

式(10)给出了扩散距离 D_t 的计算公式，λ_l 和 φ_l 分别是第 l 维的特征值和特征向量，特征值 $1=\lambda_1\geqslant\lambda_2>\cdots>0$，因此给定一个阈值 $1>\sigma_1>0$，可以确定 t 步时的维度 l，σ_1 是粒度参数，σ_1 值越小，聚簇的粒度越细。



(a)$t=8$

(b)$t=64$

(c)$t=1024$

图 3 不同时间尺度 t 时的亚稳态

3.5 目标函数

由图 3 和式(10)可以得知，当 $t\to\infty$ 时，低特征值对应的特征向量将逐步丢失，最终所有的点将扩散到一个稳态分布，这是由于信息伴随着 Markov 松弛过程而损失，下面将定义一个目标函数，让 t 收敛到一个有意义的几何结构。

互信 $I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log(p(x,y)/p(x)p(y))$ 常用于描述概率分布之间的相关度测量[20]。设初始状态时顶点变量为 $X(0)=\{X_j(0)\}$，Markov 松弛到 t 步时的顶点变量为 $X(t)=\{X_i(t)\}$，那么 $X(0)$ 和 $X(t)$ 的互信息为：

$$p_i^t=\sum_j p_{i,j}^t p_j p(X_i(t)|X_j(0))=P_{i,j}^t$$

$$I(X(0);X(t))=\sum_j p_j\sum_i P_{i,j}^t\log\frac{P_{i,j}^t}{p_i^t}=\sum_j p_j D_{KL}[P_{i,j}^t|p_i^t]\tag{11}$$

$$-\frac{\mathrm{d}I(t)}{\mathrm{d}t}=-\frac{\mathrm{d}I(X(0);X(t))}{\mathrm{d}t}$$

$$=-\sum_j p_j\sum_i\left[tP_{i,j}^t\log\frac{P_{i,j}^t}{p_i^t}+\frac{t}{\ln 2}P_{i,j}^t p_i^t\right] \quad (12)$$

D_{KL} 是 Kulback-Liebler 散度，既然一个可遍历的 Markov 链存在一个唯一的稳态分布 $\pi_i=d(y)/\sum_{z\in X}d(z)$，对于所有的 j 满足 $\lim_{t\to\infty}p(X_i(t)\mid X_j(0))=\pi_i$，也就是说 $P_{i,j}^t$ 的所有行都松弛到 π_i，与初始分布概率相比，信息将完全损失（$D_{KL}=0$）。互信息 $I(X(0);X(t))$ 随着 t 的增长将指数渐近地衰减到 0，目标函数(12)描述了 Markov 松弛时 $I(X(0);X(t))$ 的信息损失率，值得注意的是，当 Markov 松弛处于某个亚稳态时，信息损失率变化很小，对于模糊边界的聚类，将出现多个亚稳态，信息呈梯形下降，由于特征值之间的 gap 随着 Markov 松弛将会放大，原来相距较远的会更远，原来相距较近的将更近，因此 $-\frac{\mathrm{d}I(t)}{\mathrm{d}t}$ 初始的下降速度很慢，随着 t 的增长下降速度呈指数增加。给定阈值 σ_2，$\frac{\mathrm{d}I(t)}{\mathrm{d}t}>\sigma_2\frac{\mathrm{d}I(0)}{\mathrm{d}t}$ 即可让 Markov 松弛收敛到一个信息损失量较小的有意义的几何结构。

3.6　HS-MR 算法描述

输入：时空特征数据集 $X(x_1,x_1,\cdots,x_n)$（相似度用高斯核函数计算）

输出：聚类数目 m，特征函数 $(\lambda_1^t\phi_1(x),\lambda_2^t\phi_2(x),\cdots,\lambda_m^t\phi_m(x))$

1. 构造超图 G，对于每个顶点 x_i，用 K-NNS(K=2)初始化超边。

2. 计算关联矩阵 H 及随机转移矩阵 P。

3. 初始化 t=0，m=0，特征值 $\lambda_1,\lambda_2,\cdots,\lambda_n$ 和特征向量 $\varphi_1,\varphi_2,\cdots,\varphi_n$。

4. Markov 松弛，t=t+1，计算 λ^t，$S(\sigma_1,t)$。

5. 如果 m=0，m=$S(\sigma_1,t)$；如果 m=$S(\sigma_1,t)$，转步骤 4。

6. 如果目标函数 $\frac{\mathrm{d}I(t)}{\mathrm{d}t}>\sigma_2\frac{\mathrm{d}I(0)}{\mathrm{d}t}$，m=$S(\sigma_1,t)$，转步骤 4。

7. 输出结果。

算法只会在前亚稳态发生变化时才比较互信息损失率，计算互信息损失率的时间复杂度为 $O(n^3)$，超图相似度矩阵的计算复杂度为 $O(C_n^2)$，超图构建时间复杂度为 $O(n^2)$，整个算法的时间复杂度为 $O(n^4)$。

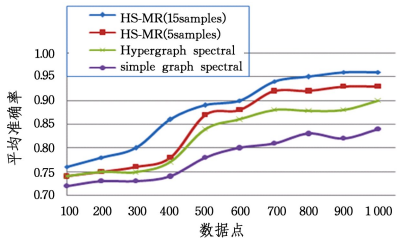
4　实验结果及分析

4.1　数据集

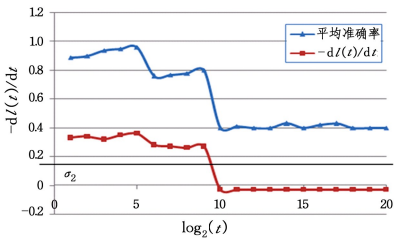
实验原始数据集来自于 UCI 的 Robot Execution Failures，它记录了机器进行故障检测时，连续 315 ms 时间窗口内力和力矩传感器 15 次采样的数据测量值，为了提高聚类准确率，使用离散傅里叶变换策略对数据进行预处理，采样特征数据的形式化描述是向量（FX,FY,FZ,TX,TY,TZ），特征语义词袋词包含（正常，前碰撞，后碰撞，左碰撞，右碰撞，移动…），实验中收集了 5 组类型数据 $LP1-LP5$，一共有 $(88+47+47+117+164)\times15=6945$ 个特征数据。每个数据作为超图中的一个顶点，初始化时每个顶点和最近两个邻居构成一条超边。

4.2　性能比较

实验中，将 HS-MR 算法的平均准确率和简单图谱聚类、标准超图谱聚类进行了比较。聚类的平均准确率取 10 次运行结果的平均值，从图 4(a)可以看出，随着数据集的增加，超图谱聚类的平均准确率整体高于简单图谱聚类，而 HS-MR 算法在每次采集 15 个数据点的情况下平均准确率会有一个明显的提高。这说明当采样数据点聚类分区界限比较清晰时，HS-MR 算法和超图谱聚类的准确率是近似的，但是当出现很多模糊边界数据点时，HS-MR 算法的平均准确率好于超图谱聚类。



(a)数据集和平均准确率



(b)尺度和信息损失率

图 4　平均准确率及参数影响

4.3　参数 σ_1 和 σ_2 的影响

HS-MR 聚类方法是基于一种凝聚的思想：将聚簇集合以某种形式合并，从而形成最终的聚簇结果。参数 σ_1 决定了将原始顶点打碎的粒度，$\sigma_1=0$ 时一个顶点就是一个聚簇。特征值 eigengap 明显时，σ_1 的取值的影响不大，但数据点边界模糊时，σ_1 值越小，聚类的性能越好。

参数 σ_2 是衡量信息损失率的程度，$1>\sigma_2>0$，从图 4(b)可以看出，信息损失率随着随机游走步数 t 的增加而降低，平均准确率也随之降低，当 $t\to\infty$ 时，信息损失率将趋向于 0，即信息完全损失，但是 t 在一定范围内，平均准确率的变化波动很小，这是因为数据分布处于亚稳态，信息损失相对较小。因此以，σ_2 确定了不影响平均准确率的信息损失率的阈值。

结束语　车联网中的数据蕴含大量的时空特征，如何将高维的时空特征嵌入到低维的特征语义袋，可以被归约成一个聚类问题，本文提出了一种自适应超图谱聚类算法（HS-MR 算法）。该算法的主要贡献在于：1）首次使用了超图的 Markov 松弛过程；2）分析并提出了一种衡量 Markov 松弛过程信息损失的目标函数；3）通过迭代学习，自动确定了聚类数目。实验结果表明，该方法对于初始几何分布无明显边界的数据集具有更准确的聚类性能。

参 考 文 献

[1] BLEI D, NG A, JORDAN M. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[2] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis[J]. Machine Learning, 2001, 42(1): 177-196.

[3] SELIM S Z. K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality[J]. IEEE Transactions on Pattern Analysis & Machine, 1984, 6(1): 81-87.

[4] 李永森.空间聚类算法中的 K 值优化问题研究[J].系统仿真学报,2006,18(3):573-576.

[5] 周水庚.基于数据分区的 DBSCAN 算法[J].计算机研究与发展,2000,37(10):1153-1159.

[6] LUXBURG U V. A tutorial on spectral clustering[J]. Statistics and Computing,2007,17(4):395-416.

[7] SHI J,MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(8):888-905.

[8] LUXBURG U V,BOUSQUET O. Limits of spectral clustering [C]// Advances in Neural Information Processing Systems (NIPS). 2005:857-864.

[9] SCHÖLKOPF B. Learning with Hypergraphs:Clustering,Classification,and Embedding[C]// Conference on Advances in Neural Information Pr. 2006:1601-1608.

[10] ZHAN Y. A semi-supervised incremental learning method based on adaptive probabilistic hypergraph for video semantic detection[J]. Multimedia Tools and Applications, 2015, 74(15): 5513-5531.

[11] LIU X. Event-Based Media Enrichment Using an Adaptive Probabilistic Hypergraph Model[J]. IEEE Transactions on Cybernetics,2015,45(11):2461-2471.

[12] MICHOEL T. Alignment and integration of complex networks by hypergraph-based spectral clustering[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics,2012,86(5 Pt 2): 1188-1205.

[13] LIU J. Learning semantic features for action recognition via diffusion maps[J]. Computer Vision & Image Understanding, 2012,116(3):361-377.

[14] LU L. Recognizing human actions by two-level Beta process hidden Markov model[J]. Multimedia Systems,2015:1.

[15] LANGE T. Stability-based validation of clustering solutions[J]. Neural Computation,2004,16(6):1299-323.

[16] ZHOU D. Beyond pairwise classification and clustering using hypergraphs[C]// Max Plank Institute for Biological Cybernetics. Tübingen,Germany,2005.

[17] NADLER B,LAFON S. Diffusion maps,spectral clustering and reaction coordinates of dynamical systems[J]. Applied & Computational Harmonic Analysis,2005,21(1):113-127.

[18] COIFMAN R R,LAFON S. Diffusion Maps[J]. Submitted to Applied Computational and Harmonic Analysis,2004.

[19] MOUYSSET S,NOAIUES J,RUIZ D,et al. Spectral Clustering:Interpretation and Gaussian Parameter[M]. Data Analysis Machine Learning and Knowledge Discovery,2014.

[20] 夏利民.基于信息瓶颈算法的图像语义标注[J].模式识别与人工智能,2008,21(6):812-818.

(上接第 438 页)

D_b 用来衡量不同厂家生产的智能电表质量的总体差异。从表 3 中的结果可以看出,杭州百富和宁夏隆基两个厂家所生产的智能电表质量差异较小,杭州百富与河南许继两个厂家所生产的智能电表质量差异较大。计算厂家之间的 D_b 值可以得到,D_b(杭州百富和宁夏隆基)=0.707,D_b(杭州百富和河南许继)=15.09,D_b(杭州百富和河南许继)远大于 D_b(杭州百富和宁夏隆基),也充分证实了从表格中得出的结论。

总的来说,由于智能电表运维系统中,数据从安装开始的运行时间短,因此积累的故障数据量少,同时有些数据已经丢失了,分析中智能电表在故障率上的表现要比实际使用时高出少许,这有利于将来积累更多数据后对模型做进一步的改进,从而提高模型分析的准确性。应用多层模型研究智能电表故障率的问题还有很大的空间可以挖掘,可以将该研究方法推广到其他产品领域。

结束语 本文通过建立多层次的混合数据模型对天津电力公司电力研究院供的智能电表运行故障数据进行了纵向分析,评估了不同购入批次的智能电表以及不同厂家生产的智能电表在故障率方面的表现,为智能电表使用单位的采购提供了依据,也为各厂家提供了质量反馈。由于天津电力公司电力研究院院提供的智能电表的故障数据不是非常的完美,因此分析结果难免有偏差,未来随着智能电表运行时间的增长,更多的运行数据被保存,模型将会得到改进和完善。

参 考 文 献

[1] 董力通,周原冰,李蒙.智能电网对智能电表的要求及产业发展建议[J].能源技术经济,2010,22(1):15-17.

[2] 唐涛涛.电能表的误差发生分析与解决办法[J].现代测量与实验室管理,2011(3):13-15.

[3] O'Connor,Kleyner. Guide for selecting and using reliability predictions based on IEEE 1413[EB/OL].[2012-5-15]. http://www.doc88.com/p-90425856345.html.

[4] 王慧芳,杨荷娟,何奔腾,等.输变电设备状态故障率模型改进分析[J].电力系统自动化.2011,35(16):27-31.

[5] 顾伟,褚建新.基于故障统计模型的可修系统维修周期预测法[J].机械强度,2000,22(1):22-27.

[6] GB 17215.911-200×/IEC/TR 62059-11-2002,电量计量设备.可靠性.第 11 部分:一般概念[EB/OL].[2012-5-12]. http://www.tsinfo.js.cn/wine/index/gbtdetails.aspx?A100=IEC/TR%2062059-11-2002.

[7] GB 17215.911-200×/IEC/TR 62059-31-1-2008,电能测量设备--可靠性--第 31-1 部分:加速可靠性测试--提高温度和湿度[EB/OL].[2012-5-12]. http://www.tsinfo.js.cn/inquiry/gbtdetails.aspx? A100=IEC%2062059-31-1-2008.

[8] 徐瑞.多层线性模型在高一物理学习影响因素分析中的应用[D].昆明:云南师范大学,2009.

[9] 刘红云,孟庆茂.纵向数据分析方法[J].心理科学进展,2003,11(5):586-592.

[10] 孙吉贵,刘杰,赵连宇.聚类算法研究[J].软件学报,2008,19(1):48-61.

[11] 李瑞莹,康锐.基于 ARMA 模型的故障率预测方法研究[J].系统工程与电子技术,2008,30(8):1588-1591.