MFASSC:基于间隔 Fisher 分析的半监督聚类方法*

李 森[1a,2], 刘希玉[1b]

(1. 山东师范大学 a. 信息科学与工程学院; b. 管理科学与工程学院, 济南 250014; 2. 山东省分布式计算机软件新技术重点实验室, 济南 250014)

摘 要: 针对高维数据的聚类问题,提出一种基于间隔 Fisher 分析(MFA)的半监督聚类算法。该算法首先使用已标记样本进行 MFA 映射,得到投影矩阵 W 后,再利用求得的投影方法对未标记样本进行降维;然后在低维空间引入基于约束的球形 K-means(PCSKM)算法对降维后的数据进行半监督聚类,根据第一次的聚类结果,交替进行降维与聚类操作,直到算法收敛为止。该算法利用监督信息有效地集成了数据降维和半监督聚类。实验结果表明,该方法能够有效处理高维数据,同时能提高聚类性能。

关键词: 半监督聚类; 成对约束; 间隔 Fisher 分析; 数据降维

中图分类号: TP311 **文献标志码**: A **文章编号**: 1001-3695(2012)11-4093-04

doi:10.3969/j.issn.1001-3695.2012.11.023

MFASSC: semi-supervised clustering approach based marginal Fisher analysis

LI Sen[1a,2], LIU Xi-yu[1b]

(1. *a. School of Information Science & Engineering, b. School of Management Science & Engineering, Shandong Normal University, Jinan* 250014, *China*; 2. *Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan* 250014, *China*)

Abstract: For the problem of clustering with high dimensional data, this paper presented a novel semi-supervised clustering approach based marginal fisher analysis(MFASSC). All the data were first projected to a low-dimensional space using marginal Fisher analysis (MFA) and then clustered by PCSKM in the projected space. The algorithm effectively utilized supervised information to integrate dimensionality reduction and semi-supervised clustering. According to the clustering results above, it conducted dimensionality reduction operations and clustering analysis alternately until convergence. Experimental results show MFASSC can effectively deal with the high-dimensional data and simultaneously improve the clustering performance.

Key words: semi-supervised clustering; pairwise constraint ; marginal Fisher analysis; dimensionality reduction

0 引言

聚类分析是根据样本间的相似性将数据分类到不同的类或簇中的过程,使得同一类内数据具有较高的相似性,不同类间的数据具有较大的差异性。近年来在机器学习和数据挖掘领域中,同时利用监督信息和无监督信息来进行聚类的半监督聚类算法受到了广泛的关注。相比于无监督聚类,半监督聚类的目的是通过监督信息来指导无监督信息进行聚类。在如信息检索、计算生物学和图像处理等许多应用领域中,获得类标记的代价是很困难的。因此,可以通过获得较少的标签数据或者成对约束来指导无监督聚类,即半监督聚类。

现有的半监督聚类算法很多是在传统聚类算法的基础上发展而来的。按照监督信息的使用方式不同,半监督聚类算法可分为三类:

a)基于约束的半监督聚类算法。这类算法使用已知的先验信息来约束聚类的搜索过程,减少搜索的盲目性。该类算法的典型算法有 Basu 等人[1]在 K-means 算法的基础上提出的 Seeded-K-means 和 Constrained-K-means 方法,Wagstaff 等人[2]提出的在算法迭代中强制满足约束条件的 COP-K-means 算法。

b)基于距离的半监督聚类算法。这类算法使用先验信息来得到一个新的距离度量函数,从而改变各样本之间的距离,使其有利于聚类[3];在谱聚类算法中利用约束信息调整样本间的距离,并将聚类问题转换成求图的拉普拉斯矩阵的谱分解[4,5]。

c)基于约束和距离集成的半监督聚类算法,该算法将前两种算法结合在一起使用。经典算法有 Bilenko 等人[6]提出的基于 K-means 的半监督算法。该算法引入成对约束作为监督信息,在目标函数上增加了不满足约束的惩罚项。在此算法的基础上,一系列的改进算法被提出[7,8]。

以上三类算法主要用于对低维数据进行聚类,当遇到高维数据时,其聚类效果会比较差。然而在许多应用领域中,如基因表达、图像检索和人脸识别等方面,所处理的数据具有高维的特点,如果不对数据进行有效的降维处理,则很容易出现所谓的维数灾难问题。针对高维数据聚类问题,Tang 等人[9]提出了一种基于特征投影的半监督聚类算法,该算法采用 must-link 和 cannot-link 成对约束得到投影矩阵,忽略了大量无标号

收稿日期: 2012-03-04; **修回日期**: 2012-04-25 **基金项目**: 国家自然科学基金资助项目(61170038);山东省自然科学基金资助项目(ZR2011FM001);山东省软科学重大项目(2010RKMA2005)

作者简介:李森(1988-),女,山东济宁人,硕士研究生,主要研究方向为数据挖掘、人工智能(lisen0222@163.com);刘希玉(1964-),男,"泰山学者"特聘教授,博导,主要研究方向为数据挖掘、计算智能.

样本数据,因此在一定程度上限制了半监督聚类性能的提高。Ding 等人[10]提出了一种自适应无监督降维迭代算法,该算法先使用 K-means 方法来聚类产生数据的类标号,然后用线性判别分析方法对高维数据降维,再使用 K-means 方法对数据聚类。该算法解决了数据高维问题,但是利用 K-means 方法首先在高维数据上进行聚类,聚类的误差性干扰了降维效果。

本文通过引入 MFA 降维思想和 PCSKM 聚类算法,提出一种基于间隔 Fisher 分析的半监督聚类算法来解决高维数据的聚类问题。本算法使用 MFA 对样本进行降维,然后在低维空间中对数据进行半监督聚类,再利用聚类结果指导降维,重复执行降维和聚类过程,直到算法收敛为止。实验结果表明,该方法能够有效处理高维数据,同时充分利用有监督信息,有效地提高了聚类性能。

1 基于间隔 Fisher 分析的半监督聚类方法

1.1 MFA 简介

LDA 等降维方法要求数据服从高斯分布假设,而在实际应用问题中这种假设并不一定总是成立。为了克服 LDA 等方法的不足,Yan 等人[11]利用图框架思想提出一种间隔 Fisher 判定分析(MFA)算法。MFA 提出构建类内紧凑图 G^l 和类间分离图 G^p 来分别描述同类样例之间的连接性和异类样例之间的分离性,如图 1 所示。

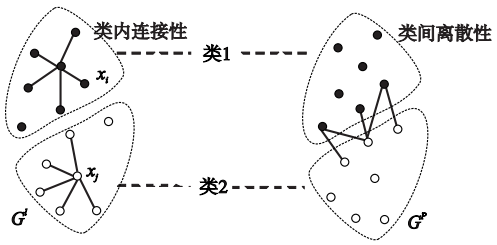

图1 类内连接性与类间离散性示意

在类内紧凑图中,S_c 用来描述两点之间的连接性,定义为

$$S_c = \sum_i \sum_{\text{or}} \| w^{\mathrm{T}} x_i - w^{\mathrm{T}} x_j \|^2 = 2w^{\mathrm{T}} X(D - W) X^{\mathrm{T}} w \tag{1}$$

$$W_{ij} = \begin{cases} 1 & \text{if} \quad i \in N_{k_1}^+(j) \quad \text{or} \quad j \in N_{k_1}^+(i) \\ 0 & \text{else} \end{cases}$$

其中:$N_{k_1}^+(i)$ 表示样例 x_i 的同类集合中其 k_1 近邻的索引集合。

在类间分离图中,S_p 用来描述样例之间的分散性,定义为

$$S_p = \sum_i \sum_{\text{or}} \| w^{\mathrm{T}} x_i - w^{\mathrm{T}} x_j \|^2 = 2w^{\mathrm{T}} X(D^p - W^p) X^{\mathrm{T}} w$$

$$W_{ij}^p = \begin{cases} 1 & \text{if} \quad (i,j) \in P_{k_2}(c_i) \quad \text{or} \ (i,j) \in P_{k_2}(c_j) \\ 0 & \text{else} \end{cases} \tag{2}$$

其中:$P_{k_2}(c)$ 表示样例 x_i 和 x_j 是 k_2 最近邻居且它们属于不同的类别。

遵循线性化的图嵌入框架,最小化目标函数为

$$w^* = \arg\min_w \frac{w^{\mathrm{T}} X(D - W) X^{\mathrm{T}} w}{w^{\mathrm{T}} X(D^p - W^p) X^{\mathrm{T}} w} \tag{3}$$

其中:W 是类内紧凑图 G 中的加权矩阵;D 为对角矩阵,其定义为 $D_{ii} = \sum_j W_{ij}$;W^p 是类间分离图 G^p 中的相似矩阵;D^p 为对角矩阵,其定义为 $D_{ii}^p = \sum_j W_{ij}^p$。通过求解上述目标函数式(3),输出最终优化的线性投影方向 w^*。

1.2 基于成对约束的球形 K-means

半监督学习中监督信息包括类属信息和成对约束信息两类。a)类属信息,其直接标记样本的类别。b)成对约束信息,Wagstaff 等人引入两种类型的成对约束(must-link 和 cannot-link)来辅助聚类搜索。Must-link 限制规定两个样本必须在同一聚类中;cannot-link 限制规定两个样本不能在同一聚类中。基于约束的先验信息比类属信息更为一般化,可以从类属信息获得等价的成对约束信息,反之则不然。本文中用带有类标记的先验信息来指导降维,在聚类中类属信息可以转换为成对约束信息来指导聚类。由于降维中用到的监督信息为带类标的数据,聚类中用到的监督信息为成对约束信息,所以数据要经过类标信息转换为约束信息的过程。转换规则为:

$$\begin{aligned} &x_i \in A \ \&\& \ x_y \in A \Rightarrow \\ &(x_i, x_y) \in \text{must-link} \\ &x_i \in A \ \&\& \ x_y \in B \Rightarrow \\ &(x_i, x_y) \in \text{cannot-link} \end{aligned} \tag{4}$$

其中:x_i、x_y 是带有类标的数据;A、B 为类簇。

经过类标信息转换后,受文献[9]的启发,本文引入基于成对约束的球形 K-means 聚类算法(pairwise constraint spherical K-means, PCSKM)进行聚类。如果成对约束完全正确,很容易从 must-link 约束中推导出一个等价关系。如图 2 所示,实线代表 must-link 约束,虚线代表 cannot-link 约束;可以用均值样本代替 must-link 约束的传递闭包;白点表示原始数据,黑点表示传递闭包中的均值样本;数据集 $\{a_1, a_2, a_3\}$、$\{b_1, b_2, b_3, b_4, b_5\}$ 和 $\{c_1, c_2, c_3\}$ 分别表示不同的传递闭包;a、b、c 分别代替数据集的均值样本。这样 must-link 约束就可以用 a、b、c 来表示。在转换后的数据集中,每个传递闭包的大小表示均值样本的权重。

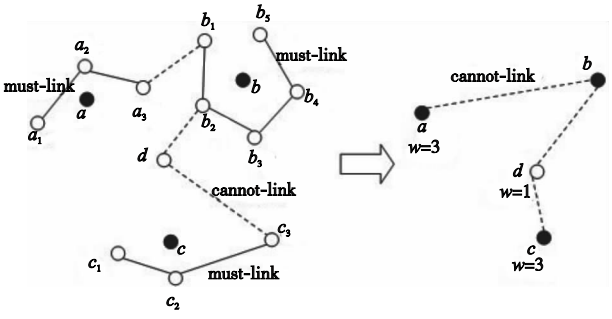

图2 Must-link约束的合并示意

经过上面步骤的操作,数据 X 可简化为 $X' = \{x'_1, x'_2, \cdots, x'_{n'}\}$ $(n' < n)$,并且相应的权值为 $W = \{w_1, \cdots, w_{n'}\}$。Cannot-link 成对约束的集合为 $C_{CL} = \{(x'_i, x'_j)\}$,对 cannot-link 约束要找到两个不同的聚类中心 $\mu_{x'_i}$ 和 $\mu_{x'_j}$,使 $w_i \cdot x_i^{\mathrm{T}} \mu_{x'_i} + w_j \cdot x_j^{\mathrm{T}} \mu_{x'_j}$ 最大化,就分别分配 x'_i、x'_j 到 $\mu_{x'_i}$ 和 $\mu_{x'_j}$ 聚类中。PCSKM 算法见算法 1。

算法 1 基于成对约束的球形 K-means 算法

输入:样本数据 X',cannot-link 约束集合 C_{CL},相应权值 W 和聚类数目 K。

输出:K 个不相交的样本分割。

a)初始化 K 个聚类中心 $\{\mu_h\}_{h=1}^K$,置 $t \leftarrow 1$。

b)重复执行以下步骤,直到收敛:

for $i = 1$ to m

(a)对每一个不在任何 cannot-link 成对约束中的样本 x'_i,

找到一个聚类中心满足 $y_n = \arg\max_k x_i'^{\mathrm{T}}\mu_k$;

(b)对于每对属于 cannot-link 约束的点对(x'_i, x'_j),找到两个不同的聚类中心 μ_k 和 $\mu_{k'}$,使得 $w_i \cdot x_i'^{\mathrm{T}}\mu_k + w_j \cdot x_j'^{\mathrm{T}}\mu_{k'}$ 最大化;

(c)对聚类 k,使$\chi'_k = \{x'_i \mid y_i = k\}$,更新聚类中心 $\mu_k = \sum x \in \chi'_k / \| \sum x \in \chi'_k \|$。

c)$t \leftarrow t+1$。

1.3　基于间隔 Fisher 分析的半监督聚类算法

已有的半监督聚类方法在处理高维数据时会降低聚类的性能,同时耗费大量的时间以及空间。DSCA[12]等一些方法虽然一定程度上降低了数据空间的维数,但这些方法都以破坏数据的信息结构为代价,因此降低了聚类准确率。为此,本文提出一种新的基于间隔 Fisher 分析的半监督聚类算法(semi-supervised clustering approach based marginal Fisher analysis, MFASSC)。由式(1)~(3)可知,MFA 在降维的过程中充分利用其类别信息,通过最大化类间离散度和最小化类内连接性来保持高维数据内在的低维流形,使投影后的数据更好地保持数据的信息和分布结构。

首先,对标记样例进行 MFA 降维。如 1.1 节所述,MFA 在降维过程中使用样本的类别信息,因此使用已标记样本进行 MFA 映射,得到投影矩阵 W 之后,再利用上述求得的投影方法对未标记样本进行降维。然后,对所有降维后的样本执行 1.2 节所述的 PCSKM 算法进行半监督聚类。由于 MFA 降维以及聚类方法本身存在一些不确定因素(如参数设置),因此暂时称上述聚类结果得到的类别信息为模糊类标。至此,所有样本均为已标记样本。最后,重复以下步骤,逐步更新未标记样本的类标,使上述模糊类标最大程度上接近其真实类别标记:a)使用已标记样本的类标和未标记样本的模糊类标再次进行 MFA 降维;b)对降维后的样本再次执行 PCSKM 算法。重复进行 a)和 b),直到聚类结果收敛为止,得到每个未标记样本的类标。实验观察到该算法迭代小于 10 次即可收敛。该算法使降维与聚类交替进行,不断更新其模糊类别,提高了聚类准确率,并且 MFA 在降维的同时较好地保持了样本的类别信息,为提高聚类性能奠定了良好的基础。MFASSC 算法见算法 2。

算法 2　基于间隔 Fisher 分析的半监督聚类算法

输入:样本数据 X(labeled data 和 unlabeled data),聚类数目 K。

输出:K 个不相交的样本分割。

a)用 MFA 对样本进行降维(首先处理标记样本,然后处理未标记样本),$t \leftarrow 1$。

b)执行算法 1,对初步降维后的数据集进行聚类,得到 K 个样本分割。

c)重复以下步骤:

(a)利用 K 个样本分割,对样本集进行 MFA 降维;

(b)使用 PCSKM 算法对降维后的数据集再次聚类,得到 K 个样本分割;

(c)$t \leftarrow t+1$,返回 c),直到算法收敛为止。

2　实验及其分析

2.1　实验设置

为了评价新算法的有效性,首先选择三种相关算法来进行对比。这三种算法分别为:

a)基于成对约束的球形 K-means 算法(PCSKM)[9]。该算法在 2.2 节中已详细介绍。

b)自适应无监督降维迭代算法(LDA-Km)[10]。该算法使用 K-means 来产生数据的类标号,然后用线性判别分析方法对高维数据降维,在降维空间中再次进行聚类。

c)特征投影的半监督聚类算法(SCREEN)[9]。该算法是在 2007 年的 KDD 会议上提出的一种半监督聚类算法,其利用成对约束信息得到投影矩阵,在子空间中使用 PCSKM 方法对数据聚类。

实验环境:Core i3 2.13 GHz,2 GB 内存,MATLABR2009a。本文所用的测试数据集为标准人脸数据库 YaleB、PIE、ORL、Yale 和 JAFFE,数据集信息如表 1 所示。

表 1　数据集特性

数据集	样本数	样本维	聚类数
YaleB	55	1 024	5
PIE	510	1 024	3
ORL	100	4 096	10
Yale	55	1 024	5
JAFFE	213	4 096	10

为了客观地评价算法的性能,文中分别对各算法的聚类准确率和 Rand Index(RI)值随监督信息的百分比的变化状况进行了测试。在已知分类结果的条件下,最常用的评价聚类性能的指标是 RI 指标,该指标视聚类结果为对每一成对点是否来自同一类所作出的判断,从而 RI 指标定义为

$$\mathrm{RI} = N_R / N_P \tag{5}$$

其中:N_R 是正确的判别数目;N_P 是所有的样本组成的样本对数目,其值为 $n(n-1)/2$。

2.2　实验结果

聚类准确率取五次实验的平均值。实验中带标记数据比例取 25%。算法聚类准确率如表 2 所示。

表 2　各算法聚类准确率　　/%

数据集	PCSKM	LDA-Km	SCREEN	MFASSC
YaleB	82.43	83.67	87.65	88.62
PIE	80.03	79.88	84.69	85.65
ORL	79.93	80.65	84.25	84.74
Yale	81.47	82.42	85.31	84.83
JAFFE	79.31	80.35	84.13	84.68

四种算法在各数据集上的聚类准确率如表 2 所示。首先,MFASSC 算法在四个数据集上的准确率均高于其他三种算法,只在一个数据集上略低于 SCREEN,但相对于另外两种算法仍有明显优势,因此相对来说 MFASSC 拥有最高的聚类性能。其次在数据集 ORL 与 JAFFE 上,四种算法的聚类准确率略低于其他三个数据集,这可能是因为 ORL 与 JAFFE 的聚类个数较大,且数据维数也相对较高。

四种算法在五个数据集上的 RI 值如图 3 所示。首先随着标记样例数量的增多,MFASSC、SCREEN 和 PCSKM 的 RI 值均有上升趋势,并且 MFASSC 随标记样例的增多 RI 值平滑升高,而 SCREEN 和 PCSKM 波动较大,说明 MFASSC 性能更稳定。其次,基于降维思想的半监督聚类方法 MFASSC 和 SCREEN 的 RI

值高于 PCSKM 方法,这是因为当数据维数较高时,数据降维过程有效地保持了数据的信息特征,同时忽略了部分冗余信息,提高了聚类的性能。最后,与其他三种方法相比,MFASSC 在五个数据集上取得最高的 RI 值。一方面是因为 MFA 在降维过程中充分利用其类别信息来保持数据的信息特征;另一方面,说明半监督聚类中的监督信息对聚类性能的改善效果显著。

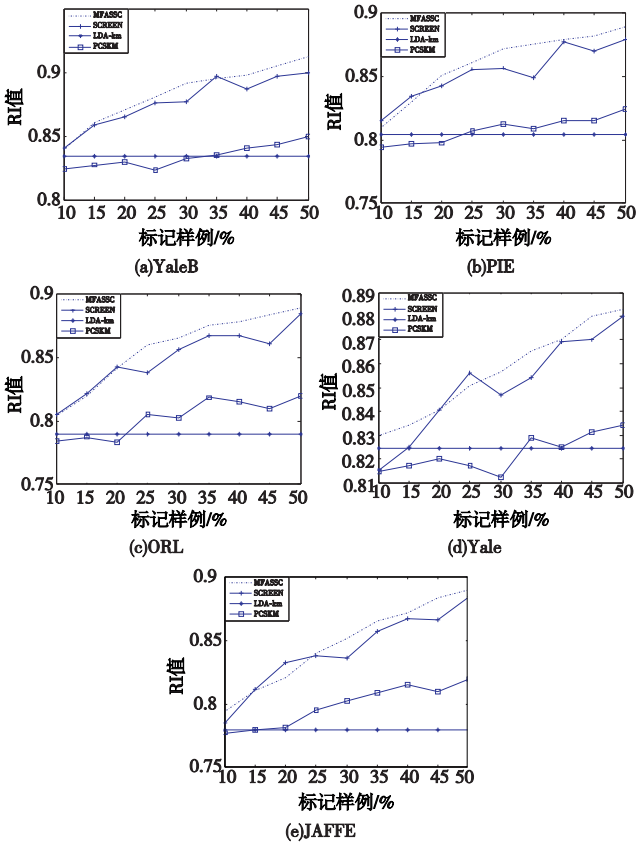

图3 算法在不同数据集上RI值随监督信息量的变化情况

3 结束语

本文提出一种基于间隔 Fisher 分析的半监督聚类算法,算法使用 MFA 对样本进行半监督降维,然后在低维空间中对数据进行半监督聚类,再利用聚类结果指导降维,重复执行降维和聚类过程,直到收敛为止。实验结果表明,该方法能够有效处理高维数据,同时充分利用有监督信息,有效地提高了聚类性能。

在半监督聚类算法中,监督信息越多越有助于提高算法的性能,但获得监督信息的代价较大。如何选择有利的监督信息来提高算法性能成为接下来的研究方向。

参考文献:

[1] BASU S, BANERJEE A, MOONEY R J. Semi-supervised clustering by seeding[C]//Proc of the 19th International Conference on Machine Learning. 2002:19-26.

[2] WAGSTAFF K, CARDIE C, ROGER S, *et al*. Constrained-K-means clustering with background knowledge[C]//Proc of the 18th International Conference on Machine Learning. 2001:577-584.

[3] YEUNG D Y, CHANG Hong. Extending the relevant component analysis algorithm for metric learning using both positive and negative equivalence constraints[J]. Pattern Recognition, 2006, 39(5): 1007-1010.

[4] 王玲,薄列峰,焦李成. 密度敏感的半监督谱聚类[J]. 软件学报, 2007,18(10):2412-2422.

[5] CHEN Wei-fu, FENG Guo-can. Spectral clustering: a semi-supervised approach[J]. Neurocomputing, 2011, 77(1): 229-242.

[6] BILENKO M, BASU S, MOONEY R J. Integrating constraints and metric learning in semi-supervised clustering[C]// Proc of the 21st International Conference on Machine Learning. 2004:81-88.

[7] KULIS B, BASU S, MOONEY R J. Semi-supervised graph clustering: a kernal approach[C]//Proc of the 22nd International Conference on Machine Learning. 2005:457-464.

[8] YAN Bo-jun, DOMENICONI C. An adaptive kernal method for semi-supervised clustering[C]// Proc of the 17th European Conference on Machine Learning. 2006:521-532.

[9] TANG Wei, XIONG Hui, ZHONG Shi, *et al*. Enhancing semi-supervised clustering: a feature projection perspective[C]//Proc of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007:707-716.

[10] DING C H, LI Tao. Adaptive dimension reduction using discriminant analysis and K-means clustering[C]// Proc of the 24th International Conference on Machine Learning. 2007:521-528.

[11] YAN Shui-cheng, XU Dong, ZHANG Ben-yu, *et al*. Graph embedding and extensions: a general framework for dimensionality reduction [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007,29(1):40-51.

[12] 尹学松,胡恩良,陈松灿. 基于成对约束的判别型半监督聚类分析[J]. 软件学报,2008,19(11),2791-2802.

(上接第 4084 页)

[2] 周芳龙,王浩,姚宏亮. 基于粒子滤波的非线性系统静态参数估计方法[J]. 计算机应用研究,2011,28(5):1637-1639,1654.

[3] YU Dong-chuan, RIGHERO M, KOCAREV L. Estimating topology of networks[J]. Physical Review Letters, 2006, 97(18): 188701.

[4] GOUESBET G, LETELLIER C. Global vector-field reconstruction by using a multivariate polynomial L2 approximation on nets[J]. Physical Review E, 1994, 49(6): 4955-4971.

[5] LU Jun-an, LV Jin-hu, XIE Jin. Reconstruction of the lorenz and chen systems with noisy observations[J]. Computers and Mathematics with Applications, 2003, 46(8-9): 1427-1434.

[6] BEZRUCHKO B P, SMIRNOV D A. Constructing nonautonomous differential equations from experimental time series[J]. Physical Review E, 2000, 63(1): 016207.

[7] WANG Wen-xu, YANG Rui, LAI Ying-cheng, *et al*. Predicting catastrophes in nonlinear dynamical systems by compressive sensing[J]. Physical Review Letters, 2011, 106(15): 154101.

[8] WANG Wen-xu, YANG Rui, LAI Ying-cheng, *et al*. Time-series-based prediction of complex oscillator networks via compressive sensing[J]. Europhysics Letters, 2011, 94(4): 48006.

[9] TIPPING M E. Sparse Bayesian learning and the relevance vector machine[J]. Journal of Machine Learning Research, 2001, 1: 211-244.

[10] NEWMAN M E J, WATTS D J. Renormalization group analysis of the small world network model[J]. Physics Letters A, 1999, 263(4-6): 341-346.

利用输出时间序列估计网络拓扑时,统一多项式的最高幂指数取为 3,式(10)的第一个等式写成统一多项式时共有 320 项,非零项个数很少向量显然是稀疏的,即使是各节点全连接时仍是稀疏的。当各节点的输出时间序列中含有强度为 0.002 的高斯噪声且采样点数为 3 000 时,网络中第 12 个节点 N_{12} 处统一多项式系数估计值如图 2 所示。图 2 中三个较大的值为式(10)的第一个方程的系数估计值,其中 -100.0472 对应 $-100y_{12}$ 项, -33.3579 对应 $-33.33x_{12}^3$ 项。由于 N_{12} 与其他六个节点相连,所以式(10)的第一个方程中如果将 x_{12} 对应的项进行合并则应为 $94x_{12}$,其估计值为 94.225。以同样方式估计式(10)的第二个方程,便得到了节点的动力学方程以及与其耦合的节点。由于系数值与耦合强度数值上差异很大,所以将三个大的系数值置为零后重新显示于图 2 左上方。

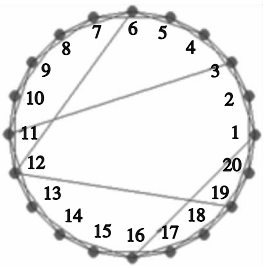

图1 FHN神经元小世界网络拓扑

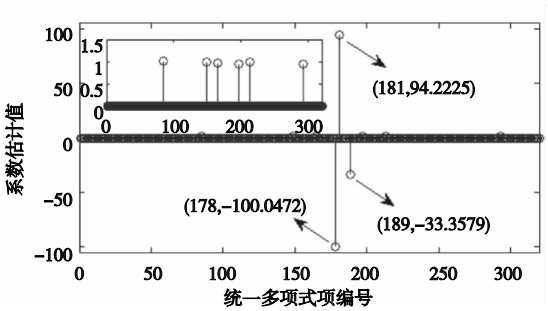

图2 节点N_{12}第一个方程统一多项式系数估计值

虽然式(10)中第二个方程并没有与其他节点耦合,如果仍将其写成第一个方程的统一多项式形式显然也是可以的,并且非零项非常稀疏,这里考虑相对不稀疏的情况,即不考虑耦合的情况,此时统一多项式只有 16 项,分别为

$$x^0y^0 + x^0y^1 + x^0y^2 + x^0y^3 +$$
$$x^1y^0 + x^1y^1 + x^1y^2 + x^1y^3 +$$
$$x^2y^0 + x^2y^1 + x^2y^2 + x^2y^3 +$$
$$x^3y^0 + x^3y^1 + x^3y^2 + x^3y^3$$

其估计结果如图 3 所示。

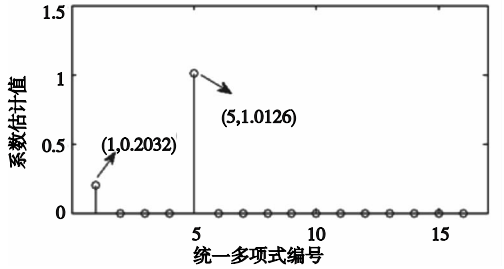

图3 节点N_{12}第二个方程统一多项式系数估计值

图 3 中的第一项对应 x^0y^0 即常数项,第五项对应 x^1y^0 即 x,显然图 2 中也是这么对应的。例如第 178 项为 -100.0472,每个节点都能写成 16 项,显然 178 是第 12 个节点的第 2 项,即对应 x^0y^1。

下面研究噪声对估计误差的影响。对单个节点的估计而言,多项式的非零项或非零项误差有的大于零有的小于零,要比较所有非零项与所有零项的估计误差和时,显然直接用绝对误差会噪声相互抵消,不能充分展示估计误差的真实大小,这里定义噪声的估计误差为绝对误差的绝对值的和,即 $E = \sum_i |e_i|$,其中 e_i 为第 i 个零项或非零项的绝对误差。图 4 为第 12 个节点零项与非零项的估计误差,噪声强度小于 0.004 时估计误差很小,估计得到的动力学方程和其连接的节点很准确,这里的估计误差可看做误差能量,显然误差能量越小,估计精度越高。其余节点都可得到类似结论。

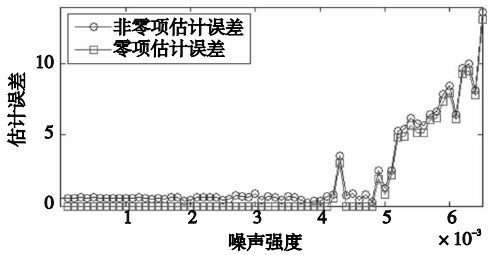

图4 噪声强度对估计值的影响

对每一个节点执行系数估计就得到该网络的拓扑结构,其邻接矩阵如图 5 所示。图中矩阵外的数值代表节点编号,矩阵中的非零值代表估计得到的两节点间的耦合强度,与图 1 描述的拓扑结构一致,并且耦合强度非常接近真实值。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
20	0.989	0.965	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.948	0.954	0	
19	0.973	0	0	0	0	0	0	0	0	0	0	0.972	0	0	0	0	1.016	0.984	0	0.941
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.923	0.955	0	0.952	0.935
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.962	0.97	0	1.015	0.93	0
16	0.953	0	0	0	0	0	0	0	0	0	0	0	0	0.967	0.984	0	0.989	0.968	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0.996	0.935	0	0.975	0.978	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0.97	0.996	0	0.896	0.956	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0.925	0.982	0	0.962	0.959	0	0	0	0	0
12	0	0	0	0	0	1	0	0	0	0.994	0.972	0	0.954	0.981	0	0	0	0	0.942	0
11	0	0	0.986	0	0	0	0	0	0.949	0.951	0	0.985	0.981	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0.986	0.97	0	1.021	0.934	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0.969	0.988	0	0.995	0.937	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0.956	1.024	0	0.947	0.974	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0.953	0.956	0	0.989	0.966	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0.973	0.983	0	1.006	0.953	0	0	0	0.954	0	0	0	0	0	0	0	0
5	0	0	0.934	0.981	0	0.96	0.986	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0.965	0.965	0	0.967	0.998	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0.963	0.983	0	0.975	0.955	0	0	0	0	0	0.945	0	0	0	0	0	0	0	0	0
2	0.965	0	0.975	0.97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.021
1	0	0.972	0.967	0	0	0	0	0	0	0	0	0	0	0	0	0.989	0	0	0.997	0.961

图5 FHN小世界网络邻接矩阵估计值

4 结束语

本文应用相关向量机方法实现了小世界神经元网络的节点动力学和拓扑结构估计。以 FHN 神经元模型为节点、节点间通过膜电压相互电耦合为连边构成的小世界网络为例,通过将耦合方程写成非零项稀疏的统一多项式,利用相关向量机估计得到稀疏解,从而获得节点动力学方程和网络拓扑。结果表明,节点动力学方程结构和网络拓扑均可被准确地估计出来,动力学方程参数和耦合强度也有很高的估计精度,且估计结果对噪声有强鲁棒性。同时,上述结论对网络规模具有鲁棒性,即网络节点数的多少对动力学结构和拓扑估计准确性无影响。本文的研究还表明,该估计策略对 Lorenz 和 Rössler 等混沌系统同样有效。

参考文献:

[1] HUANG De-bin. Synchronization-based estimation of all parameters of chaotic systems from time series [J]. Physical Review E, 2004, 69(6): 067201.

(下转第 4096 页)