数据缺失下学习贝叶斯网的一种混合启发方法*)

廖学清[1] 吕 强[1,2]
(苏州大学计算机科学与技术学院 苏州 215006)[1]
(江苏省计算机信息处理技术重点实验室 苏州 215006)[2]

摘 要 建立了具有数据缺失训练集下学习贝叶斯网的一种混合启发方法:SGS-EM-PACOB算法。它基于打分-搜索方法,利用GS和EM数据补全策略分别得到学习所需要的统计因子,并将两者联合起来作为PACOB算法的启发因子。实验证明,SGS-EM-PACOB算法充分保留GS和EM两者的优点,促使算法能够平稳地收敛到理想结果。相对于只具有单一数据补全策略的算法,该算法不仅在度量数据拟合程度的Logloss值上保持稳定,而且在学习到的贝叶斯网络结构上也有改进。

关键词 学习贝叶斯网,数据补全策略,混合启发

Hybrid Heuristic for Learning Bayesian Network with Missing Values

LIAO Xue-qing[1] LU Qiang[1,2]
(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)[1]
(Jiangsu Provincial Key Lab for Computer Information Processing Technology, Suzhou 215006, China)[2]

Abstract Presented an efficient hybrid heuristic SGS-EM-PACOB algorithm for learning Bayesian network with missing values. It is based on scoring and searching method by using GS and EM data completion policies to attain statistic information, which is essential in learning Bayesian network. SGS-EM-PACOB algorithm combines these two policies for PACOB, an excellent parallel ant colony heuristic for learning bayesian network with complete dataset. The experiments showed SGS-EM-PACOB algorithm fully out-performed both GS and EM, and made the algorithm converge to ideal results smoothly. Comparing with those algorithms having only one data completion policy, SGS-EM-PACOB algorithm not only achieves a stable Logloss value, which measures how well the dataset matches the learned network, but also makes improvements on the learned bayesian network structure.

Keywords Learning bayesian network, Data completion policy, Hybrid heuristic

1 引言

由于诸多内外因素,学习贝叶斯网所需要的现实训练数据集存在着不同程度的数据缺失现象[1]。由于数据非随机缺失问题可以通过引入隐藏变量转化为随机缺失问题[2],因此本文只讨论随机数据缺失的情况。目前,关于数据缺失情况下学习贝叶斯网的问题,国内外学者提出了诸多算法。有代表性的研究有1987年Tanner和Wong等人提出的DA[3]算法,Friedman于1997年提出的SEM算法[4]和1998年提出的Bayesian-SEM算法[5],王双成等人于2004年提出的BN-GS算法[6],Riggelsen等人于2005年提出的eMC[7]和后来的MBP[8]算法等。不难发现,上述算法共同特点都是利用相应补全策略补全不完备数据集,得到学习所需要的统计因子,使数据缺失的学习问题转化为完备数据集下学习贝叶斯网的问题。比如,SEM就是通过EM算法得到期望统计因子,BN-GS则通过GS策略[9]来修正缺失数据集,而MBP利用马尔可夫链局部近似估计方法来估计缺失数据集。但是,这些算法都面临着一个重要的挑战——如何避免以较高概率地收敛到局部最优结构。

本文鉴于此,建立了数据缺失情况下一种基于GS和EM两种数据补全策略混合启发的学习贝叶斯网算法:SGS-EM-PACOB算法。它基于打分-搜索方法[10],利用GS和EM数据补全策略分别得到各自学习所需的统计因子,并将两者联合起来作为PACOB算法[11-13]的启发因子。实验证明,SGS-EM-PACOB算法充分保留了GS和EM两种数据补全策略的优点,使得两者相互作用,共同启发,相互制约,促使算法能够平稳地收敛到理想结果。相对于只具有单一数据补全策略(GS或EM)的算法,SGS-EM-PACOB算法,不仅在度量数据拟合程度的Logloss值上保持稳定,而且在学习所得到的贝叶斯网络结构上也有改进。

本文接下来的安排为:第2部分介绍GS和EM数据补全策略与PACOB的结合;第3部分详细阐述SGS-EM-PACOB算法;第4部分列举出有关实验数据并进行分析;最后为总结及展望。

2 SGS-PACOB和SEM-PACOB算法

利用GS和EM两种数据补全策略分别与PACOB算法联合构成相应SGS-PACOB和SEM-PACOB算法,进行了有关实验。ACOB[13]算法是完备数据集下学习贝叶斯网络的出色算法。4个ACOB并行策略构成的PACOB算法,可以大

*)基金项目:本文获国家教育部博士点基金(20060285008),江苏省自然科学基金(BK2003030)资助。**廖学清** 工学学士,硕士研究生,研究方向为智能信息处理;**吕 强** 硕士生导师,教授,主要研究方向是计算机操作系统、分布式计算、计算语言学、自然语言处理等。

范围地、有启发地加大搜索空间，能更好地考察候选网络，为基于打分-搜索机制的类似算法提供良好支持。SGS-PACOB详细算法如下(SEM-PACOB算法只需把下面算法中的GS换成EM)：

Procedure SEM-PACOB()

Choose M^0 and Θ^0 using some tactics;

Loop for n = 0,1,…,until convergence

Completing Dataset using GS Completing Tatics with M^n and Θ^n;

Using PACOB find the M^{n+1} with the highest MDL score;

Learning EM Parameters Θ^{n+1} for M^{n+1};

Finish Loop

Return M^{n+1} and Θ^{n+1}

实验中PACOB采用4个ACOB并行的策略，每个ACOB涉及到的参数与文献[12]相同，而GS数据补全策略采用的补全方案与文献[6]相同。本文应用到的ALARM训练缺失数据集和测试数据集是根据 http://www.norsys.com 提供的ALARM网概率分布图生成的，生成了10000个例子的测试数据集和2000个例子的训练数据集。在2000个例子的训练数据集中又分别随机生成具有10%，20%和30%数据率的缺失训练数据集。

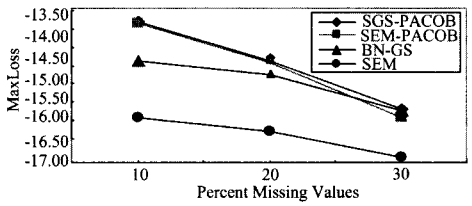

图1 SGS-PACOB，SEM-PACOB，BN-GS和SEM算法MaxLoss的比较

图1显示出由于PACOB提供了较优候选网络的因素，SGS-PACOB算法相对于BN-GS算法[6]有了较大的改进，而SEM-PACOB算法则比文献[6]中的SEM算法有重大提高。但是，在多次的实验中，我们却发现SGS-PACOB和SEM-PACOB算法同样存在着陷入局部最优网络的可能性，而且，这种陷入局部最优的可能性相当地高，从而使得Logloss值所跨越的区间相当大。表1中的实验数据是两个算法在每个缺失训练数据集上执行100次所得的统计数据。该表可以例证算法较高概率地收敛到局部最优结构的情况(其中Maxloss，Minloss，Aveloss和StdLoss分别为100次运行后Logloss值的最好、最差、均值和标准差；Interval为相应Maxloss，Minloss之差)。

表1 SGS-PACOB和SEM-PACOB算法100次执行的统计数据

		10%	20%	30%
SGS-PACOB	MaxLoss	−13.8202	−14.6910	−15.7780
	MinLoss	−14.1887	−15.0722	−16.5322
	AveLoss	−14.0435	−14.8812	−16.0485
	StdLoss	0.0900	0.1061	0.1522
	Interval	0.3685	0.3812	0.7542
SEM-PACOB	MaxLoss	−13.8361	−14.7262	−15.9815
	MinLoss	−14.2996	−15.1558	−16.6429
	AveLoss	−14.0785	−14.9417	−16.2341
	StdLoss	0.0807	0.0921	0.1223
	Interval	0.4635	0.4296	0.6614

表1所示SEM-PACOB和SGS-PACOB算法无论在Logloss值的跨越区间Interval上还是其标准差Stdloss上，都存在着很大的波动性，特别在面对30%缺失率的训练数据集时，SGS-PACOB算法的Interval值和Stdloss值分别高达0.7542和0.1522。由此可见，SGS-PACOB和SEM-PACOB算法虽然较之单纯GS-BN和SEM有改进，但同样面临陷入不被预知的多个局部最优。

3 SGS-EM-PACOB算法的设计

针对SGS-PACOB和SEM-PACOB算法较高概率收敛到局部最优的情况，下面提出了SGS-EM-PACOB算法。它是基于打分-搜索方法，利用GS和EM两种数据补全策略分别得到学习所期望的因子，并将两者联合起来作为PACOB的启发因子。

3.1 初始网络的选择

随机初始化数据，利用K2SN[14](本文用MDL[15]打分代替K2[16]打分)算法先学到贝叶斯网，然后以此网络作为初始网络。该算法简单，但能一般性地刻画原始训练数据集所蕴含的信息，又可以不过拟合于训练数据集。本文简称这样的初始网络为K_SN。K_SN作为一个拟合训练数据集的初始网络，对于提高算法效率，改善对较高缺失率规模又不大的训练数据集下解的质量，促进算法平稳收敛到理想结果是有很大帮助的[17]。

3.2 GS和EM策略共同启发

学习贝叶斯网最关键的是能够从训练数据集中学习到真实的因果依赖关系和切实的网络结构。所以，能不能够找到较优候选网络，特别是对于不完备数据集下学习贝叶斯网，显得尤为重要。在Logloss值上SGS-PACOB胜于BN-GS和SEM-PACOB远胜于SEM的事实，其主要因素在于PACOB有启发、较大规模的搜索空间为算法提供了良好的候选网络。按照文献[11-13]，ACOB选择下一条边有0.8的概率是选择拥有 $\tau_{ij}^{\alpha}\eta_{ij}^{\beta}(i\neq j)$ 最大值的边，其中 τ_{ij} 为信息素，η_{ij} 启发因子(公式1)，α 和 β 为相应权值。

$$\eta_{ij}=f(X_i,Pa(X_i)\cup\{X_j\})-f(X_i,Pa(X_i))(i\neq j) \quad (1)$$

f 为打分函数，本文采用MDL打分。η_{ij} 为 X_i 结点变量在其原有父亲 $Pa(X_i)$ 加入新结点变量 X_j 后的分值增益。

我们在实验中发现SGS-PACOB和SEM-PACOB两种算法之间存在着一方对某些边很敏感，另一方却表现出不感兴趣，或者说是某些边出现次数两算法有较大悬殊的情况。表2是关于此在2000个例子具有数据缺失率10%的训练数据集下运行100次后得到最优贝叶斯网边的统计数据(列举的边均为标准网络中的边)。

表2 SGS-PACOB和SEM-PACOB两算法运行后边出现次数悬殊情况

	7 ↓ 9	5 ↓ 13	14 ↓ 5	6 ↓ 28	11 ↓ 30	32 ↓ 30	36 ↓ 33
SGS-PACOB	81	90	91	49	16	71	26
SEM-PACOB	92	70	69	59	31	87	65

从表2可以得出结论：不同的启发策略将导致选择不同的边。为了能够保留SGS-PACOB和SEM-PACOB两种算法，准确地来说，是两种算法各自数据补全策略在网络结构上的优良特性，我们采用了混合启发策略并构建了SGS-EM-

PACOB算法。该算法中ACOB选择下一条边加入到网络结构的启发因子修改为：

$$\eta_{GSij}^{\beta_{GS}} * \eta_{EMij}^{\beta_{EM}} (i \neq j) \quad (2)$$

其中 η_{GSij} 是GS补全策略的启发因子。相应地，η_{EMij} 是EM补全策略的启发因子，分别表示各自策略下结点变量 X_i 在其原有父亲 $Pa(X_i)$ 上加入新结点 X_j 后的分值增益，而 β_{GS} 和 β_{EM} 为两者在混合启发算法中的相应权值。事实证明，采用这样的联合启发因子，能够更好地选出更优网络结构。

3.3 以GS策略为主线

从表1和图1可以看出，SGS-PACOB算法除了在Std-Loss不如SEM-PACOB算法之外，其它各项均表现出相对较好的特性。同时，注意到EM数据补全策略得到一种确切的期望统计量，而这些总计量完全由当前最优网络和训练数据集已观察到的记录决定。而GS策略则不同，虽然它也是根据当前网络得到统计量，可这是一种抽样，存在一定程度的随机性。为了确保算法既能够向前平稳推进，又可以继承先前的优良特性，SGS-EM-PACOB算法采用了以GS数据补全策略为主，EM数据补全策略作为联合启发为辅的策略。实验中 β_{GS} 取1.5，β_{EM} 取1.0。还有就是，利用 *PACOB* 算法选择最佳候选网络结构的MDL打分等均基于GS数据补全策略所得的统计因子。

3.4 SGS-EM-PACOB算法小结

综上，我们可以把SGS-EM-PACOB算法总结为：

Procedure SGS-EM-PACOB()

Randomly initial D to and Choose M^0 and Θ^0 using K_SN by MDL score

Loop for n = 0,1,…,until convergence

Completing Dataset using GS Completing Tactics with M^n andΘ^n;

Completing Dataset using EM Completing Tactics with M^n andΘ^n;

Using PACOB Algorithm find the M^{n+1} with the highest MDL score,heuristically hybrid the GS and EM policy;

Learning EM Parameters Θ^{n+1} for M^{n+1};

Finish Loop

Return M^{n+1} andΘ^{n+1}

4 实验数据及其分析

本节列举的是关于ALARM数据集的结果。以下实验数据均为SGS-EM-PACOB算法在ALARM训练数据集执行100次的统计数据。

4.1 数据拟合度评价

SGS-EM-PACOB算法的数据拟合度与SGS-PACOB和SEM-PACOB算法的比较如图3所示。

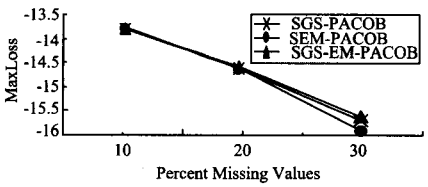

图3 3种算法100次执行后各自MaxLoss值的比较

从图3中可以看出，SGS-EM-PACOB算法在Logloss最大值上略高于SGS-PACOB和SEM-PACOB算法。同时，我们还比较了3种算法在10%，20%，30%缺失率的训练数据集上，执行100次之后Logloss值的标准差及其最大跨越区间的比较(如图4、图5所示)。

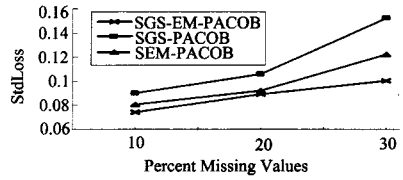

图4 3种算法Logloss标准差图

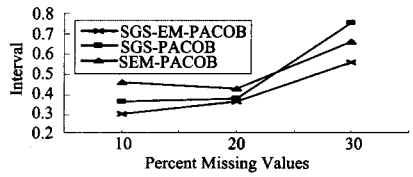

图5 3种算法Logloss最大跨越区间图

图4和图5显示出SGS-EM-PACOB算法在Logloss标准差及其最大跨越区间上都远优于SGS-PACOB和SEM-PACOB算法。各次执行的Logloss值表现得相对稳定，加上Logloss最大值比较理想，可以表明SGS-EM-PACOB算法在数据拟合度上不仅取得了理想结果，而且具有很高的稳定性，特别是对于高缺失数据的情况下，SGS-EM-PACOB算法有了稳定表现。

4.2 网络结构评价

在学习到的网络结构方面，SGS-PACOB和SEM-PACOB算法存在着正如表2所示的差异。那么SGS-EM-PACOB的情况呢？可看表3(其中GS偏差值是某边在SGS-EM-PACOB算法中出现次数减去该边在SGS-PACOB算法的次数，EM偏差值类同)。

表3 SGS-EM-PACOB算法主要边出现情况

	7 ↓ 9	5 ↓ 13	14 ↓ 5	6 ↓ 28	11 ↓ 30	32 ↓ 30	36 ↓ 33
SGS-EM-PACOB	90	77	78	48	25	89	44
GS偏差值	+9	−13	−13	−1	+9	+18	+18
EM偏差值	−2	+7	+9	−11	−6	+2	−21

从表3可以看出SGS-EM-PACOB算法除了2条边偏差值全正或全负的情况，其余都是SGS-PACOB和SEM-PACOB两算法的折中。偏差值全正的边固然是好，而全负的边也是靠近SGS-PACOB或SEM-PACOB算法的出现次数。同时本文也统计了已学到网络结构与标准网络结构相同边以及无父亲相同结点的数目总和SA(Same Arcs)。

表4 SGS-EM-PACOB等3种算法100次运行之后SA统计情况

		10%	20%	30%
100次SA总和	SGS-PACOB	2998	2402	1777
	SEM-PACOB	3068	2285	1757
	SGS-EM-PACOB	3059	2480	1859
100次SA标准差	SGS-PACOB	4.1196	3.7295	2.9231
	SEM-PACOB	3.9314	3.8359	3.3653
	SGS-EM-PACOB	2.9722	3.3357	2.8892

表4呈现出在SA总和上SGS-EM-PACOB逐渐超越SEM-PACOB和SGS-PACOB两算法，并且逐渐拉开了距离；在SA标准差上，SGS-EM-PACOB表现得更稳定。这再一次

显示出利用两种补全策略共同启发起到了正面作用。由此可以推测出 GS 和 EM 数据补全策略共同启发的策略起到了良好作用，能够相对促使 SGS-EM-PACOB 算法平稳地得到较优网络结构。

4.3 实验小结

我们还对 INSURANCE 等数据集进行了实验，其结果与本节反映的结论基本一致。限于篇幅，只列举了 ALARM 数据集的情况。

综上可以得出由 GS 和 EM 两种数据补全策略联合的 SGS-EM-PACOB 算法无论在数据拟合度上，还是网络结构上都表现出较好的特性。更重要的是该算法能够平稳地收敛到理想结果，而不会像 SGS-PACOB 和 SEM-PACOB 算法那样较高概率地收敛到局部最优。

结束语 本文建立了具有数据缺失学习贝叶斯网混合启发模型的算法，把 GS 和 EM 两种数据补全策略较好地联合起来共同启发，能够使 SGS-EM-PACOB 算法平稳地收敛到理想结果，相对于单一数据补全策略的算法取得了改进结果。但同时也注意到 EM 补全策略计算量相当庞大，当面对大规模数据缺失或者大规模例子训练数据集时是一个瓶颈。未来的研究中，可以找一计算量相对较少的数据补全策略进行更深更紧的联合。比如 C. Riggelsen 提出的 MBP 算法中补全策略就是一个不错的选择。

参 考 文 献

[1] Heckerman D. Bayesian networks for data mining. Technical Report, MSR-TR-97-02. Microsoft Research, Redmond, 1997

[2] 张连文，郭海鹏. 贝叶斯网络引论. 北京：科学出版社，2007

[3] Tanner M, Wong W. The calculation of posterior distributions by data augmentation. J. of the Am. Stat. Assoc, 1987, 82(398): 528-540

[4] Friedman N. Learning belief networks in the presence of missing values and hidden variables // Proc. of the 14th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997: 125-133

[5] Friedman N. The Bayesian Structural EM Algorithm // Proc. of the 14th Int'l Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1998: 129-138

[6] 王双成，苑森淼. 具有丢失数据的贝叶斯网络结构学习研究. 软件学报，2004，15(7)：1042-1048

[7] Riggelsen C , Feelders A . Learning Bayesian network models from incomplete data using importance sampling // Cowell R G, Ghahrammani Z, eds. Proc. of Intelligence and Statistics. 2005: 301-308

[8] Riggelsen C . Learning Bayesian Networks fromm Incomplete Data: An Efficient Method for Generating Approximate Predictive Distributions. Department of Information and Computing Sciences, Universiteit Utrecht, 2007

[9] Geman S, Geman D. Stochastic relaxation, Gibbs distributions and Bayesian restoration of inmages. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984, 6(6): 721-742

[10] Heckerman D. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. MSR-TR-94-09. Microsoft Research, 1995

[11] 吕强，高彦明，钱培德。共享信息素矩阵：一种新的并行 ACO 方法。自动化学报，2007，33(4)：418-421

[12] 潘吉斯，吕强，王红玲. 一种并行蚁群 Bayesian 网络学习的算法. 小型微型机计算机系统，2007，28(4)：651-655

[13] de Campos L M, Fernadez-Luna J M, Gamez J A, et al. Ant colony optimization for learning Bayesian networks. Int. J. Approx. Reasoning, 2002, 31(3): 291-311

[14] Cooper G F, Herskovits E. A Bayesian methmod for the induction of probabilistic networks from data. Machine Learning, 1992, 9(4): 309-348

[15] Bouckaert R R. Probabilistic Network Construction Using the Mininum Description Length Principle // Lecture Notes in Computer Science. 1993

[16] Kruse R, Borgelt C. An empirical investigation of the k2 metric // Proc. 6th European Conf. on Symbolic and Quantitative Approaches to Reasoning and Uncertainty. 2001: 240-251

[17] 廖学清，吕强. 试析数据缺失下学习贝叶斯网中初始网络的选择. 计算机科学专刊录用

(上接第 150 页)

[4] Collobert R, Bengio S. SVMtorch: Support Vector Machines for Large-Scale Regression Problems. Journal of Machine Learning Research, 2001, 1: 143-160

[5] Amari S, Wu S. Improving Support Vector Machine Classifier by Modifying Kernel Function. Neural Networks, 1999, 12(66): 783-789

[6] Crammer J K, Elisseeff A, Shawe-Taylor J. Kernel Design using Boosting // Proceedings of Sixteenth Annual Conference on Neural Information Processing Systems. Combridge, MA, MIT Press, 2002: 537-544,

[7] Charles A M, Pontil M. Learning the Kernel Function via Regularization. Journal of Machine Learning Research, 2005, 6: 1099-1125

[8] Lanckriet G, et al. Learning the Kernel Matrix with Semidefinite Programming. Journal of Machine Learning Research, 2004, 5: 27-72

[9] Cheng S O, Smola A J, Williamson R C. Learning the Kernel with Hyperkernels. Journal of Machine Learning Research, 2005, 6: 1043-1071

[10] Chapelle O , et al . Choosing Multiple Parameters for Support Vector Machines. Machine Learning, 2002, 46(11): 131-159

[11] Chapelle O, Vapnik V N. Model selection for support vector machines // Proceedings of Thirteenth Annual Conference on Neural Information Processing Systems. Combridge, MA, MIT Press, 1999: 230-236

[12] Frauke F, Christian I. Evolutionary tuning of multiple SVM parameters. Neurocomputing, 2005, 64: 107-117

[13] Soares C, Brazdil P B, Kuba P. A Meta-Learning Method to Select the Kernel Width in Support Vector Regression. Machine Learning, 2004, 54: 195-209

[14] Liao S, Jia L. Simultaneous Tuning of Hyperparameter and Parameter for Support Vector Machines // Proceedings of the Eleventh Pacific-Asia Conference on Knowledge Discovery and Data Mining. New York. Springer-Verlag, 2007: 162-172