

一种小规模数据集下的贝叶斯网络学习方法及其应用

李亚飞^{1,3} 吕强² 苏伟峰^{1,3} 刘轶³

(北京师范大学-香港浸会大学联合国际学院 珠海 519085)¹

(苏州大学计算机科学与技术学院 苏州 215006)²

(深港产学研基地智能媒体和语音重点实验室 深圳 518057)³

摘要 提出了一种小规模数据集下学习贝叶斯网络的有效算法——FCLBN。FCLBN利用bootstrap方法在给定的小样本数据集上进行重抽样,然后在抽样后数据集上学到的贝叶斯网络来估计原数据集上的贝叶斯网络的高置信度的特征,并用这些特征来指导在原数据集上的贝叶斯网络搜索。用标准的数据集验证了FCLBN的有效性,并将FCLBN应用于酵母菌细胞中蛋白质的定位预测。实验结果表明,FCLBN能够在小规模数据集上学到较好的网络模型。

关键词 学习贝叶斯网络,小规模数据集,特征置信

Learning Bayesian Network from Small Scale Dataset and Application

LI Ya-fei^{1,3} LU Qiang² SU Wei-feng^{1,3} LIU Yi³

(Program of Computer Science and Technology, BUN-HKBU United International College, Zhuhai 519085, China)¹

(School of Computer Science and Technology, Soochow University, Suzhou 215006, China)²

(Shenzhen Key Lab of Intelligence Media and Speech, PKU-HKUST Shenzhen Hong Kong Institution, Shenzhen 518057, China)³

Abstract An efficient algorithm FCLBN for learning Bayesian network from small scale dataset was proposed. FCLBN uses the method of bootstrap to re-sample from the small scale dataset, and estimates the high confidence features of the source small scale dataset from the Bayesian networks learned from the re-sampling small datasets. The high confidence features are taken to guide the search of the best Bayesian network on the source dataset. After being evaluated on the standard benchmark dataset, FCLBN is applied to predict yeast protein localization. The result of the experiments indicates that the FCLBN algorithm can learn relatively accurate network from small scale dataset.

Keywords Learning bayesian network, Small scale dataset, Features confidence

1 引言

贝叶斯网络是将先验知识与样本信息相结合、依赖关系与概率表示相结合的数学模型,其坚实的理论基础、知识结构的自然表述方式、强大的推理能力使其成为数据分析时强有力的建模工具^[1]。

在数据分析的某些应用领域,例如在生物、化学和物理等学科中,由于实验代价、实验难度比较大,可能导致获得的珍贵实验数据比较少(例如本文第5节所描述的例子)。在这种情况下,如果沿用经典的贝叶斯网络学习方法来建模,计算效果可能会因为训练数据量较小而不可靠。通常的贝叶斯网络学习方法有两种:基于打分搜索的方法和基于条件独立性测试的方法,且基于打分搜索的方法一般有较好的结果。本文利用“特征置信+打分搜索”的方法来处理小数据集下贝叶斯网络学习。在对网络结构特征的特征置信分析后,利用置信度高的特征来指导搜索学习过程,以便能学出与潜在模型拟合得

较好的贝叶斯网络。

文献[2]提到了利用bootstrap来解决小样本数据集下贝叶斯网络的学习方案,但文献[2]中只是通过高置信的特征来估计贝叶斯网络的结构,并没有进一步利用好这些高置信的特征;并且这些高置信度的特征也只是通过简单学习而获得。本文提出的FCLBN (Feature Confidence-based Learning Bayesian Network)算法将分析得到的高置信的特征回归到原始数据集的贝叶斯网络的学习过程中,指导贝叶斯网络的搜索,以便能够得更好的网络模型。由于打分方法独立于搜索方法,因此本文中贝叶斯网络的打分方法全部使用MDL打分^[3]。

2 特征置信指导下的贝叶斯网络学习

2.1 bootstrap 抽样估计

bootstrap抽样是一种计算机模拟的方法^[4]。在统计学里,该方法常用于处理小样本数据问题,它基于给定的数据,

到稿日期:2010-08-02 返修日期:2010-11-11 本文受国家自然科学基金项目(61073017),北京师范大学-香港浸会大学联合国际学院校内项目(R201109, UIC2010-S-01.8)资助。

李亚飞(1983—),男,硕士,助教,主要研究方向为数据挖掘, E-mail: lyf1111@gmail.com; 吕强(1965—),男,博士,教授,博士生导师,主要研究方向为自然语言处理、生物信息学、贝叶斯网络等; 苏伟峰(1974—),男,博士,助理教授,主要研究方向为数据挖掘、机器学习等; 刘轶(1972—),男,博士,副教授,主要研究方向为语音识别、机器学习等。

对统计量的分布情况进行仿真处理和统计推断。

应用到贝叶斯网络的学习过程中,该方法主要是对给定的小样本数据集 D 进行随机、有放回地重抽样,可得到 n 个样本子集 $D^* = \{D_1, D_2, \dots, D_n\}$ 。然后对应地在这 n 个抽样后的样本子集上学习出网络结构 $G^* = \{G_1, G_2, \dots, G_n\}$,通过 G^* 来估计原样本数据集中所蕴含的网络结构。很显然,对于结构的估计只能通过网络结构的一些基本组成特征来进行,比如边、结点间顺序、结点邻域等。本文主要使用较为直接的边特征来估计原小样本数据集上的网络结构。

2.2 贝叶斯网络等价类

在贝叶斯网络学习的过程中,对于同一批数据而言,符合同一种模型分布的贝叶斯网络结构可能不止一个。对于同一打分值,可能有不同的网络结构与它相对应。本文要用一个等价类来代表打分值最高的一批网络^[5]。用部分有向无环图 PDAG 来表示这个等价类。在 PDAG 图中有些边是有向的,有些边是无向的,其中无向边不管取哪一种方向,得到的 DAG 图的打分值都是一样的。因此,需要将 2.1 节中学到的贝叶斯网络 DAG 集合 $G^* = \{G_1, G_2, \dots, G_n\}$ 转化为一个 PDAG 图的集合 $G^{*'} = \{G_1', G_2', \dots, G_n'\}$ 。此步是保证对打分值最高的多个网络结构进行统计。文献[5-7]证明了对于属于同一等价类的贝叶斯网络的打分值是相等的,这一点对于 bootstrap 抽样中的边特征置信来说是至关重要的。

图 1(a)和图 1(b)两个网络结构图都属于同一等价类,可以用图 1(c)同时代表它们,这样在统计边时,就可以既对边 $y \rightarrow u$ 统计,又可以对边 $u \rightarrow y$ 统计。这样就可以防止因为只单一地统计属于同一 PDAG 的一个 DAG 图导致的统计缺失。图 1(c)是图 1(a)和图 1(b)对应的 PDAG 图。

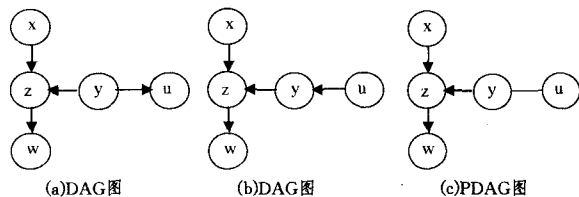


图 1

2.3 搜索算法

贝叶斯网络学习问题是一个 NP 完全问题^[8]。文献[2]中采用的是“hillclimbing (HC) + beamsearch (BS)”算法来进行网络的搜索,这两种算法完全是借助于贪婪的搜索机制,而贪婪所导致的结果往往会造成解的局部最优,不利于最优解的搜索。本文利用了搜索网络质量较好的 PACOB 算法^[9]。PACOB 算法是一种并行化的 ACOB 算法^[10],与 ACOB 算法相比,PACOB 除了继承了 ACOB 的优点(使用了 ACO 元启发指导搜索过程,有效地利用了问题域的启发信息,同时很好地发挥了蚁群之间和同代蚂蚁之间有效的协作),还在时间性能方面有很大的优势,多个并行的蚁群之间共享了信息素矩阵^[11],重叠了大量的工作量,使得搜索最好解的速度大大加快。文献[9,10]验证了 PACOB 是一种在完备数据集上搜索贝叶斯网络的较好的算法。FCLBN 的贝叶斯网络搜索过程中采用的基本搜索方法就是 PACOB。

2.4 边特征置信指导下的贝叶斯网络学习

对于贝叶斯网络的特征置信方法有很多种,主要包括边置信、结点邻域置信、结点顺序置信。本文主要采用较为直接

的边置信。所谓的边置信就是:计算 G^* 中的边 e_{ij} 在 2.1 节和 2.2 节中学到的 PDAG 图集合 $G^{*'}$ 中出现的概率,设定一个阈值 k ,当边 e_{ij} 在 $G^{*'}$ 中出现的概率高于 k 时,就认为边 e_{ij} 极可能存在于潜在分布的结构中。当遍历完网络结构集 G^* 中所有的边后,就得到一个高置信边的集合 E 。文献[2]中只是通过 bootstrap 抽样估计得到了高置信度的边集 E ,就直接利用这些高置信的边来构造贝叶斯网络。本文利用这个高置信边的集合 E 指导 PACOB 再次在原始小规模数据集 D 上学习贝叶斯网络:用 E 中的高置信边修补 PACOB 中每一代蚂蚁学出的网络结构,以便对下一代蚂蚁的搜索过程进行较好的指导,最终学出更好的网络模型。

3 FCLBN 算法设计

综上所述,给出本文的 FCLBN 算法,如下所示:

Input 数据集 $D, n=100, k \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$

Output 贝叶斯网络

- ① 读入初始数据集 D ;
- ② 在数据集 D 上随机有放回地抽样出 n 个数据集 $D' = \{D_1, D_2, \dots, D_n\}$
- ③ 调用 PACOB 算法在抽样后的数据集 D_i 上学习出一个最优的贝叶斯网络 $G_i, G^* = \{G_1, G_2, \dots, G_n\}$
- ④ 将 G^* 中的所有 DAG 图转换为 PDAG 图 $G^{*'} = \{G_1', G_2', \dots, G_n'\}$
- ⑤ For G^* 中所有可能出现的边
- ⑥ For $G^{*'}$ 中所有的 PDAG 图
- ⑦
$$p(e) = \frac{1}{n} \sum_{G_i' \in G^{*'}} 1\{e \in G_i'\}$$
 并把所有 $p(e)$ 大于阈值 k 的边记录到 E 中
- ⑧ End for
- ⑨ End for
- ⑩ 用 E 指导 PACOB 在原数据集 D 上进行贝叶斯网络学习
- ⑪ 输出最终的网络结构 B 和参数 θ

其中公式 $p(e) = \frac{1}{n} \sum_{G_i' \in G^{*'}} 1\{e \in G_i'\}$ 是用来计算每条边的置信度, $1\{e \in G_i'\}$ 表示如果边 $e \in G_i'$, 记为 1, 否则记为 0。在 FCLBN 算法中把置信度大于阈值 k 的边保存在 E 中,然后用 E 来指导在原数据集上的贝叶斯网络的搜索。本算法主要包括两个主要步骤:第一步是第①—⑨行所进行的 bootstrap 抽样估计,得到一些高置信度的边 E ;第二步是第⑩行所进行的原数据集上高置信度边指导下的贝叶斯网络学习。这里的边置信指导指的是:用 PACOB 中每一代蚂蚁学出的网络结构与第一步中得到的高置信度边做比较,若高置信度的边没有出现在学到的网络结构中,则把它添加进去。加入后如果出现了环,那么就从这个环中找到一条置信度最低的边,把它删除掉,用新得到的网络结构继续指导下一代蚂蚁的搜索。

FCLBN 的重要贡献在于两点:(1)利用了较好的 PACOB 搜索算法替换文献[2]中的“HC+BS”算法,使搜索到的网络更为精确,为最终的边置信和网络结构学习提供了保障;(2)用高置信度的边指导原小规模数据集的贝叶斯网络学习,为搜索到最优的网络结构提供了更加鲁棒的支持。

FCLBN 的重要贡献在于两点:(1)利用了较好的 PACOB 搜索算法替换文献[2]中的“HC+BS”算法,使搜索到的网络更为精确,为最终的边置信和网络结构学习提供了保障;(2)用高置信度的边指导原小规模数据集的贝叶斯网络学习,为搜索到最优的网络结构提供了更加鲁棒的支持。

FCLBN 算法的时间消耗主要包含两个部分:(1)PACOB 搜索网络的过程,这部分的时间复杂度为 $O(ma^4r)^{[9]}$,其中, m 是所使用数据集的行数, a 为数据集中属性的个数, r 为属性值最多的那个属性的属性值个数;(2)高置信度边的搜索过程,这部分时间复杂度为 $O(ne^2)$,其中 n 是 G^* 中贝叶斯网络

的个数, e 为 G^* 所有可能出现的边的条数。所以整体上 FCLBN 的时间复杂度为 $O(ma^4r) + O(na^4)$ 。

4 FCLBN 的验证

本节用学习贝叶斯网络的经典 ALARM 数据集来验证 FCLBN 的有效性。

4.1 实验数据集及实验设计

本实验程序运行平台为 IBM 的 P550 并行机,其硬件配置为 4 个双核 CPU,主频为 1.5G 赫兹,内存大小为 6GB,操作系统为 Linux。程序的编译和运行环境为 Linux 下的 gcc。本节用到的 ALARM 数据集^[12]是根据 <http://www.norsys.com> 提供的 ALARM 网概率分布图(包括 37 个结点变量,46 条边)生成的,包括一个 1000 条数据的训练集和一个 1000 条数据的测试集。对于 ALARM 网络,一般来说当数据集的条数达到 3000 以上时才能较好地学出该网络结构,所以本实验取 1000 条数据来模拟小规模数据是合理的。本实验训练结果都是程序在含有 1000 条数据的训练集上运行 50 次的平均结果,其中 FCLBN 中 n 的初始值设置为 100。

本节的实验设计如下,首先,通过实验考察不同置信度下判定边的准确性;其次,用不同置信度下的 E 来指导在原数据集上的贝叶斯网络学习,通过与标准图比较考察加置信指导和不加置信指导的贝叶斯网络学习哪一种更为准确;最后,从学出网络与测试数据集拟合程度的角度,来比较加置信指导和不加置信指导的贝叶斯网络学习哪一种学出的网络模型的泛化性较好。这样就验证了 FCLBN 的有效性。

4.2 置信度阈值的选择

首先,运行算法 FCLBN 的①—⑨行,得到高质量的边集 E 。通过与标准图比较,判定置信度 k 取不同值时判定边存在与否的准确性。其中当 $p(e)$ 大于阈值 k 时,判定此边在标准图中存在,反之,就判定此边在标准图中不存在。 tp 表示在置信度为 k 时判定存在、而且实际标准图中也存在的边的条数; fp 表示在置信度为 k 时判定存在、而实际标准图中不存在的边的条数; fn 表示在置信度为 k 时判定不存在、而实际标准图中存在的边的条数。从表 1 可以看出,随着边置信度 k 值从 0.6 提高到 1.0, tp 由 36.2 减小到 20.5, fp 由 5.3 减小到 0, fn 由 10.3 增加到 25.5。与文献[2]相比,本文判定边的准确性要更高一些,原因主要是使用了搜索性能较好的 PACOB 算法,提高了搜索边的准确性,也降低了判定边存在的错误率。而文献[2]中仅仅通过“HC+BS”等贪婪搜索的方法,很容易使搜索过程陷入局部极值。

表 1 不同置信度下边的估计情况

置信度(k)	tp	fp	fn
1.0	20.5	0.0	25.5
0.9	28.1	0.9	16.9
0.8	31.5	2.1	14.3
0.7	33.4	2.5	10.1
0.6	36.2	5.3	10.3

4.3 置信指导下的结构学习分析

用上面找到的高置信度的边指导在原数据集上进行贝叶斯网络搜索,在搜索到的网络结构方面,从用边置信指导和不用边置信指导这两个方面来进行比较,实验结果如表 2 所列,用两个大的指标来衡量正确边和错误边。为了更进一步地了解错误边的具体情况,又从 3 个方面来考虑错误边:新边、缺边与反转边。新边表示学习到的网络中存在而标准图中没有

的边的条数;缺边表示标准图中存在,而学习到的网络中没有的边的条数;反转边表示学习到的网络中存在,但是与标准图中边的方向相反的边的条数。从表 2 中可以看出,与没有高置信的边指导的网络学习相比,在具有较高置信度(k 大于 0.7)的边指导下能够找到较好的网络结构,但在最高置信度 1.0 指导下找到的网络结构却不是最好的,找到的正确边的条数为 40.8 条,错误边的条数为 7.1 条;而在置信度为 0.9 时,找到的正确边为 43.2 条,错误边达到了最低,只有 6.0 条。分析原因主要是虽然 1.0 具有最高的置信度, fp 是 0.0,但是 tp 却只有 20.5,也就是说虽然没有错误的边指导,但是正确边的置信指导却相对比较少。相比之下,0.9 的置信指导下学出的网络结构是最好的,虽然引入了 0.9 条错误的边,但是却有了 28.1 条边正确的指导,在搜索的过程中错误的边也有可能得到纠正。不过随着置信度的降低,对应的错误的边的条数也在增加,错误的边有可能导致错误的置信指导,反映到这张表中的网络的结构准确性也在下降。整体来说,合适的置信指导能够找出较好的网络结构,并且从表 3 也可以看到,0.7~1.0 置信度指导下学出的网络结构的结果都要比不加置信指导的效果要好得多。也证明了在数据集较小的情况下,采用置信指导贝叶斯网络的搜索过程是必要的。

表 2 不同置信度下网络结构的估计情况

置信度	正确边	错误边			总计
		新边	丢失边	反转边	
1.0	40.8	4.2	1.8	1.1	7.1
0.9	43.2	3.9	1.1	1.0	6.0
0.8	41.5	4.3	2.1	0.7	7.1
0.7	42.2	5.5	1.2	1.7	8.4
0.6	39.5	5.8	1.4	3.7	10.9
没有置信指导	40.2	5.7	2.2	1.8	9.7

4.4 置信下学习的泛化性能分析

以上是单从网络结构方面来分析实验结果,下面用数据拟合度 $\logloss$ 来衡量置信指导下贝叶斯网络学习算法 FCLBN 的泛化性能。用 4.3 节训练到的贝叶斯网络对 1000 条记录的测试集的计算概率的 $\logloss$ 越高,说明已学到的贝叶斯网络 B 越拟合测试集 D 。 $\logloss$ 的计算公式如下:

$$\logloss = \frac{1}{N} \sum_{i=1}^N \log P(X_i | B)$$

式中, $\log P(X_i | B)$ 表示测试集中第 i 行记录 X_i 在贝叶斯网络模型中出现的概率, N 表示测试集记录的条数。表 3 中的 Min_loss, Max_loss, Average_loss 分别表示 50 次测试中 $\logloss$ 最小值、最大值、平均值,从表中可以看到置信度大于 0.7 时, FCLBN 的 Min_loss, Max_loss, Average_loss 值都比不加置信指导时大,其中当置信度为 0.9 时最好,这说明 FCLBN 算法此时学出的网络模型与数据集拟合程度是最好的。

表 3 不同置信度下学出的网络的 $\logloss$ 值

置信度	Min_loss	Max_loss	Average_loss
1.0	-4.500407	-4.499259	-4.499833
0.9	-4.500374	-4.498350	-4.499362
0.8	-4.500407	-4.499260	-4.499834
0.7	-4.500574	-4.500407	-4.500409
0.6	-4.503610	-4.502610	-4.503110
没有置信指导	-4.502610	-4.501250	-4.501930

5 FCLBN 的一个应用

本节描述 FCLBN 在生物信息学中的一个成功的应用,

试图解决酵母细胞中蛋白质的细胞定位(Cell Compartments Localization)问题。该问题最早由耶鲁大学的 Gerstein 小组研究过^[13], 本节的数据也来自该文献的支持网站 <http://bioinfo.mbb.yale.edu/genome/>。

5.1 问题描述

酵母细胞是在全基因组和全蛋白质组层面上已经被研究得比较完整的代表性细胞, 已知其含有 6043 条基因及其相应的蛋白质。蛋白质定位问题是指将这 6043 条基因相应的蛋白质定位到细胞的各个亚室。文献[13]把细胞核、细胞质、细胞膜、液泡、线粒体和内质网等 10 种不同的亚室归类为 C, M, N, T, E 5 类位置信息。

传统上通过各种生化实验来鉴别蛋白质的定位。文献[13]收集了 30 种针对所有这 6043 条蛋白质的实验数据, 并从中推导出了 19 种实验属性对蛋白质的定位起主导作用。自然, 由于这些生物实验的难度和多样性, 致使获得的实验数据是缺失的, 且可能不一致。同时, 文献从 MIPS^[14]、YPD^[15] 和 SWISS-POT^[16] 3 大权威蛋白质数据库中提取相关的蛋白质定位数据。取这 3 大数据库中对这 6000 多条蛋白质定位信息一致的交集, 作为最权威的训练集, 形成 465 种已“确定”定位的蛋白质, 以及相关的 19 个实验属性值。显然这是一个小规模训练数据集。

在本节中希望用 FCLBN 从该训练集学习贝叶斯分类器, 并对酵母细胞的其它蛋白质的定位进行预测。

5.2 实验数据集和实验设计

实验所用软硬件平台与 ALARM 实验相同, 所用训练集是从文献[13]获得的关于酵母菌细胞的 465 种蛋白质定位及其实验数据集。所用的测试集是除了 465 种之外的 517 种蛋白质。把原来除了 465 种之外的 5500 种在实验数据中属性值大片缺失的蛋白质除去, 还余下 1200 种左右。由于文献[13]中网站上给出的是 2000 年时的分类标签, 有些可能不正确, 因此在 2008 年 10 月 16 日 MPIS 数据库中对这 1200 种左右的蛋白质分类标签重新进行了核对, 除去那些有二义性的标签, 余下就是上面提到的包含 517 种蛋白质的测试集。由于文献[13]中网站给出的 19 种实验属性中有一个属性值严重缺失达 93.8%, 因此本节的训练集和测试集都只使用到 18 种重要实验属性值。本文实验所用的数据集就是由酵母菌细胞蛋白质分子的 18 个重要属性和 1 个分类标签属性构成的。文献[13]的结果对定位给了 5 种分类标签, 但在该网站上得到的数据集只有 4 种分类标签, 且测试集属性数据有缺失(数据缺失率为 0.96%, 对这些缺失的数据进行了随机的填补), 所以本节的结果与文献[13]的结果不具有可比性。但本节对 FCLBN 在 0.6, 0.7, 0.8, 0.9, 1.0 置信度下的分类效果分别进行了测试, 并且与基于贝叶斯的分类器 Naive, Tan^[17] 和基于 Svm 的分类器的分类结果进行了比较。其中, Naive, Tan 分类器由 JBNC^[18] 实现, SVM 分类器由 SVM-light^[19] 实现, SVM-light 的运行参数是其默认值。

本节实验用 FCLBN 在训练集上学出的网络结构和参数对测试集进行分类, 从原数据集上抽样的样本个数 $n=100$, 采用交叉验证的方式来验证 FCLBN 算法的训练结果。具体做法是: 对 465 种蛋白质分子中的每一种进行随机 1—7 的标号, 从中选取属于同一标号的蛋白质作为测试集, 余下的蛋白质作为训练集, 一共进行 7 次分类测试, 最后实验结果取 7 次测试结果的平均值。在训练集上进行交叉验证之后, 再用以

训练集上学到的最好的网络和参数, 对本实验的 517 种蛋白质的测试集进行测试。

5.3 训练集学习分类器结果比较

表 4 中第 2—8 列表示所进行的交叉验证中的 7 次测试, 第 9 列表示 7 次测试结果的平均值, 第 2—6 行表示不同置信度下 FCLBN 算法的分类结果, 第 7—9 行表示采用 Naive, Tan 和 Svm 分类器的分类结果。从表中可以看出 FCLBN 在置信度为 0.7, 0.8, 0.9, 1.0 时, 分类的正确率分别为 69.31%, 72.65%, 72.21%, 70.71%, 而用 Naive, Tan, Svm 分类器进行分类的结果分别为 52.11%, 63.83%, 66.45%, 均低于 FCLBN 在置信度大于 0.7 时的分类结果。这也验证了本文提出的 FCLBN 算法处理小数据集分类问题的能力。

表 4 对 FCLBN, Naive, Tan, Svm 进行交叉验证的结果

正确率 分类方法	1	2	3	4	5	6	7	平均值
1.0	71.21%	75.23%	74.35%	73.17%	70.58%	62.78%	67.68%	70.71%
0.9	73.00%	77.20%	76.00%	75.10%	71.00%	63.23%	70.00%	72.21%
0.8	73.01%	77.50%	76.21%	75.40%	71.00%	65.13%	70.30%	72.65%
0.7	70.01%	77.45%	72.30%	75.40%	75.00%	60.00%	65.05%	69.31%
0.6	60.05%	70.55%	64.45%	65.25%	69.44%	55.21%	58.33%	63.35%
Naive	62.60%	56.92%	50.00%	53.62%	43.75%	49.27%	48.64%	52.11%
Tan	59.50%	65.00%	70.37%	69.56%	60.81%	56.62%	65.00%	63.83%
Svm	65.15%	60.60%	71.64%	71.21%	67.16%	60.60%	68.65%	66.45%

另外, 本实验还将不同置信度下的 FCLBN 分类的实验结果进行比较, 发现当置信度为 0.8 时, FCLBN 的分类正确率为 72.65%, 高于置信度为 0.6, 0.7, 0.9, 1.0 时的分类结果。这就进一步验证了 FCLBN 算法在 ALARM 数据集上的实验结论(在最高置信度下学习的结果并不是最好的)。提高置信度则意味着置信特征在质的方面提高了, 但在量的方面却降低了, 对网络搜索过程中提供的指导信息也就较少; 降低置信度则意味着置信特征在量的方面提高了, 但是在质的方面却降低了, 这两种情况都不是理想的结果。为此, 要在置信特征的数量和质量上做个折中。在 FCLBN 学习过程中, 要选取什么样的置信度对学习过程进行指导, 这与具体的应用场景有关。两个实验的结果(ALARM 在置信度为 0.9 时实验结果最好, 而在蛋白质分类实验中却是置信度取 0.8 时最好)都证明了这一结论。

5.4 预测实验比较

在 5.3 节中利用交叉验证的方法分析了不同的分类器在 465 条训练集上的学习情况, 在置信度为 0.8 时 FCLBN 的分类效果要高于 Naive, Tan 和 Svm。本节通过利用分类效果较好的 0.8 置信度的 FCLBN 训练出来的模型作为贝叶斯网分类器来对 517 种蛋白质测试集进行了分类测试, 分类结果如表 5 所列。

表 5 对 FCLBN, Naive, Tan, Svm 进行分类预测的结果

分类器	FCLBN(置信度 0.8)	Tan	Naive	Svm
正确率	51.62%	44.32%	38.20%	26.32%

从表 5 中可以看到在对包含 517 种蛋白质的测试集上进行测试后, FCLBN(置信度为 0.8)的分类正确率为 51.62%, 要好于其他的 Tan, Naive, Svm 的分类结果 44.32%, 38.20%, 26.32%。

5.5 结果分析

由于训练集中的 465 条蛋白质分子是 MIPS, YPD 和
(下转第 234 页)

- dings of the First IEEE Conference on Evolutionary Computation. Orlando; IEEE Press, 1994; 63-66
- [4] Rudolph G. Convergence of Evolutionary Algorithms in General Search Space[C]//Proceedings of 3rd IEEE Conference on Evolutionary Computation. Piscataway; IEEE Press, 1996; 50-54
 - [5] Jaegerskuepper J. Algorithmic Analysis of a Basic Evolutionary Algorithm for Continuous Optimization [J]. Theoretical Computer Science, 2007, 379: 329-347
 - [6] 谢承旺, 丁立新. 多目标进化算法中选择策略的研究[J]. 计算机科学, 2009, 36(9): 167-172
 - [7] 阎岭, 蒋静坪. 进化学习策略收敛性和逃逸能力的研究[J]. 自动化学报, 2005, 31(6): 873-880
 - [8] He Jun, Yao Xin. Towards an Analytic Framework for Analyzing the Computation Time of Evolutionary Algorithms[J]. Artificial Intelligence, 2003, 145: 59-97
 - [9] Rudolph G. Finite Markov Chain Results in Evolutionary Computation: A Tour d' Horizon [J]. Fundamenta Informaticae, 1998, 35(1-4): 67-89
 - [10] 钱敏平, 龚光鲁. 应用随机过程[M]. 北京: 北京大学出版社, 1998
 - [11] 严加安. 测度论讲义(第二版)[M]. 北京: 科学出版社, 2004
 - [12] Oliveto P S, He Jun, Yao Xin. Time Complexity of Evolutionary Algorithms for Combinatorial Optimization: A Decade of Results [J]. International Journal of Automation and Computing, 2007, 4(3): 281-293

(上接第 184 页)

SWISS POT 3 大权威蛋白质数据库中提取的相关的蛋白质定位数据, 是 3 大数据库中都是一致的, 而且属性值没有缺失, 因此分类的正确率要高于在 517 条测试集上的实验结果。从 5.3 节中的实验结果也可看出, 分类结果准确率最高的是 FCLBN 在置信度为 0.8 时的分类结果, 正确率达到了 72.65%。而在 5.4 节的预测实验中, 几种分类器的分类准确率都相对较低, 这与 517 条实验测试集数据缺失有关系。但在这种情况下 FCLBN 的分类正确率仍能达到 51.62%, 相对于 Tan, Naïve, Svm 的 44.32%, 38.20%, 26.32% 来说, 已经是大大领先了。同时, 通过这个应用结果也可知, 缺失数据下的贝叶斯网络学习也是十分重要的。

结束语 本文提出的 FCLBN 算法利用 bootstrap 抽样估计得到一些高置信度的边, 并用这些高置信度的边指导在原数据集上的贝叶斯网络学习。同时将搜索网络结构较好的启发式算法 PACOB 引入到贝叶斯网络的学习过程中, 修改了文献[2]中单一使用“HC+BS”搜索算法导致的局部最优的情况, 使边的特征置信更加准确, 这也为在原数据集上的网络学习提供了更高的置信指导。在 ALARM 数据集上的贝叶斯网络学习实验和酵母菌细胞的蛋白质定位预测实验都已经验证: 在小样本数据集下, 采用置信指导学到的贝叶斯网络模型要比不采用置信指导学到的贝叶斯网络模型好。

在小规模的完备数据集上, 基于边特征置信指导 FCLBN 算法表现出了比较好的性能。那么不完备数据集下基于特征置信指导贝叶斯网络学习算法是否仍然如此, 这将是下一步工作的重要内容, 未来工作就是要给 FCLBN 算法加上缺失数据处理。另外, 置信度阈值的选择也是一个值得深入探讨的问题。

参考文献

- [1] 张连文, 郭海鹏. 贝叶斯网引论[M]. 北京: 科学出版社, 2006
- [2] Friedman N, Goldszmidt M, Wyner A. On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks [Z]. AI&STAT VII, 1999
- [3] Lam W, Bacchus F. Learning Bayesian Belief Networks: An Approach Based on the MDL Principle [J]. Computational Intelligence, 1994, 10: 269-294
- [4] Efron B, Tibshirani R J. An Introduction to the Bootstrap[M]. New York: Chapman and Hall, 1993
- [5] Chickering D M. Learning Equivalence Classes of Bayesian-Network Structures [J]. The Journal of Machine Learning Research, 2002, 2(2): 445-498
- [6] Chickering D M. A Transformational Characterization of Equivalent Bayesian Network Structures[C]//UAI'95. 1995, 11: 87-98
- [7] Meek C. Causal inference and causal explanation explanation with background knowledge[C]//Philippe Besnard and Steve Hanks, eds. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Inc, San Mateo, CA: Morgan Kaufmann Publishers, 1995: 403-410
- [8] Chickering D M. Learning Bayesian Networks is NP-Complete [C]//Fisher D, Lenz H-J, eds. Learning from Data: Artificial Intelligence and Statistics V. Springer Verlag, 1996
- [9] De Campos L M, Fernandez-Luna J M, Gamez J A, et al. Ant colony optimization for learning Bayesian networks[J]. Int. J. Approx. Reasoning, 2002, 31(3): 291-311
- [10] 潘吉斯, 吕强, 王红玲. 一种并行蚁群 Bayesian 网络学习的算法[J]. 小型微型计算机系统, 2007(4): 651-655
- [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[J]. 自动化学报, 2007, 33: 418-421
- [12] Beinlich I, Suermondt H J, Chavez R M, et al. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks[C]//Proc. of the Second European Conf. on Artificial Intelligence in Medicine. 1989, 38: 247-256
- [13] Drawid A, Gerstein M. A Bayesian System Integrating Expression Data with Sequence Patterns for Localizing Proteins: Comprehensive Application to the Yeast Genome [OL]. <http://www.idealibrary.com on J. Mol. Biol. 2000, 301: 1059-1075>
- [14] Fusarium graminearum Genome Database[OL]. <http://mips.gsf.de/genre/proj/fusarium>
- [15] Swiss-Prot Database[OL]. <http://www.expasy.ch/sprot/>
- [16] Proteome database[OL]. <http://www.proteome.com/databases/>
- [17] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997, 29: 131-161
- [18] JBNC [OL]. <http://jbnc.sourceforge.net/>
- [19] SVMlight [OL]. <http://svmlight.joachims.org/>