

基于 BIC 测度和混合遗传算法的 BNC 结构学习

蒋望东^{1,2}, 林士敏², 鲁明羽³

(1. 湖南财经高等专科学校 信息管理系, 湖南 长沙 410205;

2. 广西师范大学 计算机科学系, 广西 桂林 541004;

3. 清华大学 智能技术与系统国家重点实验室, 北京 100084)

摘要:贝叶斯网络分类器(BNC)结构学习是一个 NP 难题。贪婪搜索(GS)算法是一种有效且准确性较高的结构学习算法,但贪婪搜索算法很容易陷入局部最优。标准遗传算法是一种全局搜索优化算法,它通过模拟生物种群的进化过程,得到全局最优解。但其个体而言,个体局部解的质量无法保证,不具备局部寻优的能力。提出了将两种算法相结合,以贝叶斯信息标准(BIC)测度为评价函数,得到一种混合遗传算法,实现了它们的优势互补。实验表明:该算法优于单独利用 GS 算法进行 Bayesian 网络结构学习,从而说明该算法的正确性和有效性。

关键词:贝叶斯网络;结构学习;贪婪搜索算法;遗传算法

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2007)03-0084-04

Structure Learning of BNC Based on BIC and Hybrid Genetic Algorithms

JIANG Wang-dong^{1,2}, LIN Shi-min², LU Ming-yu³

(1. Department of Information Management, Hunan Financial & Economic College, Changsha 410205, China;

2. Computer Science Department, Guangxi Normal University, Guilin 541004, China;

3. State Key Laboratory for Intelligent Technology & System, Tsinghua University, Beijing 100084, China)

Abstract: Structure learning of Bayesian networks classification is an NP hard problem. Greedy search algorithm is an effective and high veracity method, but it is easy to get into the local best. Standard genetic algorithm is a global search optimal algorithm, which simulates the proceeding of natural evolution and can get the global best. But its individual can't provide guarantee of getting the local best. An algorithm is proposed to combine these two algorithms with BIC as evaluation function, which can get better effect. Experimental result shows that this algorithm is better than using GS algorithm only, it is accurate and effective.

Key words: Bayesian networks; structure learning; GS algorithm; genetic algorithm

1 贝叶斯网络和贝叶斯分类器

贝叶斯网络是在不确定性环境下有效的知识表示和概率推理模型,是一种流行的图形化决策分析工具。贝叶斯分类器是基于贝叶斯网络学习方法的分类器^[1-4]。设有变量集 $U = \{A_1, A_2, \dots, A_n, C\}$, 其中 A_1, A_2, \dots, A_n 是实例的 n 个属性变量,实例可用向量 $x_i = (a_1, a_2, \dots, a_n)$ 表示,其中, a_i 是 A_i 的值,令 C 为类变量, c 表示 C 的值。应用贝叶斯定理,实例 x_i 属于

类 c_j 的概率为:

$$p(c_j | a_1, a_2, \dots, a_n) = \frac{p(a_1, a_2, \dots, a_n | c_j) p(c_j)}{p(a_1, a_2, \dots, a_n)} = \alpha p(a_1, a_2, \dots, a_n | c_j) p(c_j) \quad (1)$$

其中, α 是正则化因子, $p(c_j)$ 是类 c_j 的先验概率, $p(c_j | a_1, a_2, \dots, a_n)$ 是类 c_j 的后验概率,先验概率独立于训练数据集,而后验概率反映了样本数据对类 c_j 的影响。

依据概率的链式规则,式(1)可以表示为:

$$p(c_j | a_1, a_2, \dots, a_n) = \alpha p(c_j) \prod_{i=1}^n p(a_i | a_1, a_2, \dots, a_{i-1}, c_j)$$

实例 $e = (a_1, a_2, \dots, a_n)$ 为 c 类的概率为 $p(c | e) = \frac{p(a_1, a_2, \dots, a_n | c) p(c)}{p(a_1, a_2, \dots, a_n)}$ 。实例 e 被分到 c 的最大

收稿日期:2006-05-31

基金项目:国家自然科学基金项目(60473115)

作者简介:蒋望东(1971-),男,湖南永州人,讲师,硕士,研究方向为人工智能、机器学习;林士敏,教授,硕士研究生导师,研究方向为知识工程、数据采掘;鲁明羽,副教授,博士后,研究方向为数据采掘、网络挖掘。

后验概率的类 C^* 中, $g(e) = \arg \max_c p(c | a_1, a_2, \dots, a_n)$, $g(e)$ 称为贝叶斯分类器。

贝叶斯分类器一般有三类^[3,4]:朴素贝叶斯分类器(NBC, Naive Bayes Classifier)、树扩展朴素贝叶斯分类器(TANC, Tree Augmented Naive Bayes Classifier)和贝叶斯网络分类器(BNC, Bayesian Network Classifier)。NBC 的结构最简单,它基于属性变量条件独立的假设,BNC 最能与领域模型吻合,但学习算法复杂,TANC 介于 NBC 和 BNC 两者之间。

2 学习贝叶斯网络结构的 GS 算法

求解 $p(c | a_1, a_2, \dots, a_n)$ 是一个比较困难的问题,一般是采用启发式搜索算法,在有限的搜索空间中寻优。最简单的启发式搜索算法就是 GS(贪婪搜索)算法^[1,3]。GS 算法的执行很像一个瞎子爬山,由于他看不见山峰的位置,只能通过周围的地形来判断,他认为山峰的位置在最陡的方向,故每一步的方向也就是地形最陡的,在这种指导下,期望最终到达山顶。GS 算法的简单步骤如下:

①任意确定一个贝叶斯网络状态作为初始结构,也可以根据领域专家知识随机选取一个特殊的结构,根据测度函数计算出该结构的分值;

②通过修改初始结构求出初始结构的近邻,修改的方法有添加一条有向边、删除一条有向边、改变一条有向边的方向等,限制条件就是修改后的结构不能有有向环。分别计算出这些近邻结构在测度函数下的分值,求出与初始结构的差值。

③在所有的差值中选择最大的正差值,以该差值所对应的结构作为下一次的初始状态,重复步骤②直到没有正差值为止,此时的结构就是最优的。

3 学习贝叶斯网络结构的遗传算法

贪婪搜索(GS)算法是一种有效且准确性较高的结构学习算法,但贪婪搜索算法很容易陷入局部最优;遗传算法是一种全局搜索优化算法,它通过模拟生物种群的进化过程,得到全局最优解,但就其个体而言,个体局部解的质量无法保证,不具备局部寻优的能力。因此,设计了基于 BIC 测度和混合遗传算法的贝叶斯网络分类器(BNC)结构学习算法,将两种算法相结合,以贝叶斯信息标准(BIC)测度为评价函数,得到一种混合遗传算法,实现了它们的优势互补^[5,6]。

3.1 贝叶斯网络结构的编码表示

贝叶斯网络结构的编码表示方法一般有:矩阵表示法、邻接表链表示法、定长的邻接表链表示法,如图

1 所示,其结构的三种表示形式分别如图 2、图 3、图 4 所示。

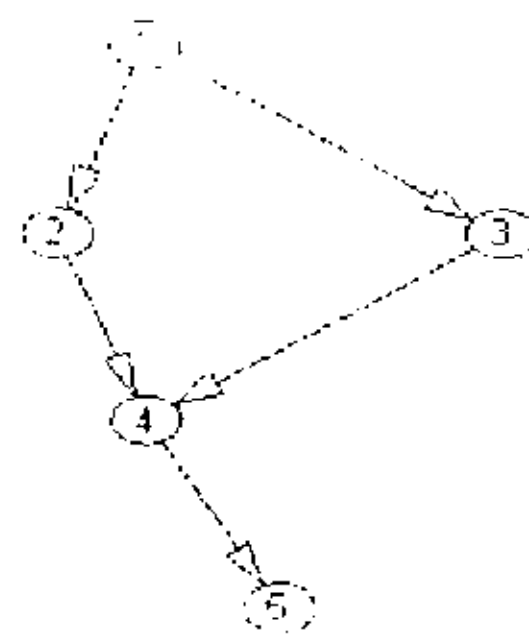


图 1 Bayesian 网

0	1	1	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	0	1
0	0	0	0	0

图 2 矩阵表示法

①
② ①
③ ①
④ ②③
⑤ ④

② ①①
③ ①①
④ ②③
⑤ ④④

图 3 邻接表链表示法 图 4 定长邻接表链表示法

三种表示方法都很直观,且各有优势,邻接表链表示法适合于以指针表示数据结构的编程语言,文中选择的编程工具 Matlab 的基本数据单元是一个维数不加限制的矩阵,用户无需考虑大量的有关矩阵的运算该采用何种算法等低层问题,更不必深入了解相应算法的具体细节,因而对用户算法语言方面的要求十分宽松,用它编写遗传算法程序,比用 C 等其它高级语言要简单、灵活、快捷,程序篇幅也将缩小许多。所以选用第一种表示方法。

3.2 染色体的选择

染色体的选择方法一般有:

(1)适应度比例法,又称轮盘赌选择法。即根据各染色体的适应度值大小进行随机选择。这种方法虽可避免陷入局部极小,但容易引起过早收敛。受寻优条件的限制,一般只能得到全局范围内的次优解,很难得到最优解。

(2)最佳个体保留法。即把群体中适应度最高的个体不经过配对交叉直接复制到下一代中,但这种方法的全局搜索能力差,只适用于只有一个峰值的搜索空间,对于具有多个峰值的搜索空间不适用。

(3)竞争选择法。上述两种选择方法各有所长,也各有缺点,差异性较大,依据选择性集成思想^[7],表现好的个体学习器越精确、差异越大,集成后可以获得的结果越好。因此,文中先采用适应度比例法进行选择,交叉后产生下一代,再利用最佳个体保留法将上一代的最佳个体直接保存下来,然后从新群体中淘汰一个适应度最差的个体。这种方法集成了上述两种方法的优点并克服了它们的缺点,称之为竞争选择法。

3.3 染色体交叉的实现

在遗传算法中如果只使用一种交叉方法,因交叉方式单一,容易引起过早收敛。同样依据选择性集成

思想^[7],可等概率使用以下两种差异性较大的交叉方法,扩大遗传算法的搜索范围,避免过早收敛:

(1)单点交叉。即选中个体对应网络中的某一个结点,将两个个体对应结点的进出弧全部交换,也可以只交换进弧,还可以只交换出弧。

(2)多点交叉。即选中多个结点,将两个个体对应的多个结点分别进行间单点交叉,如图 5 所示。

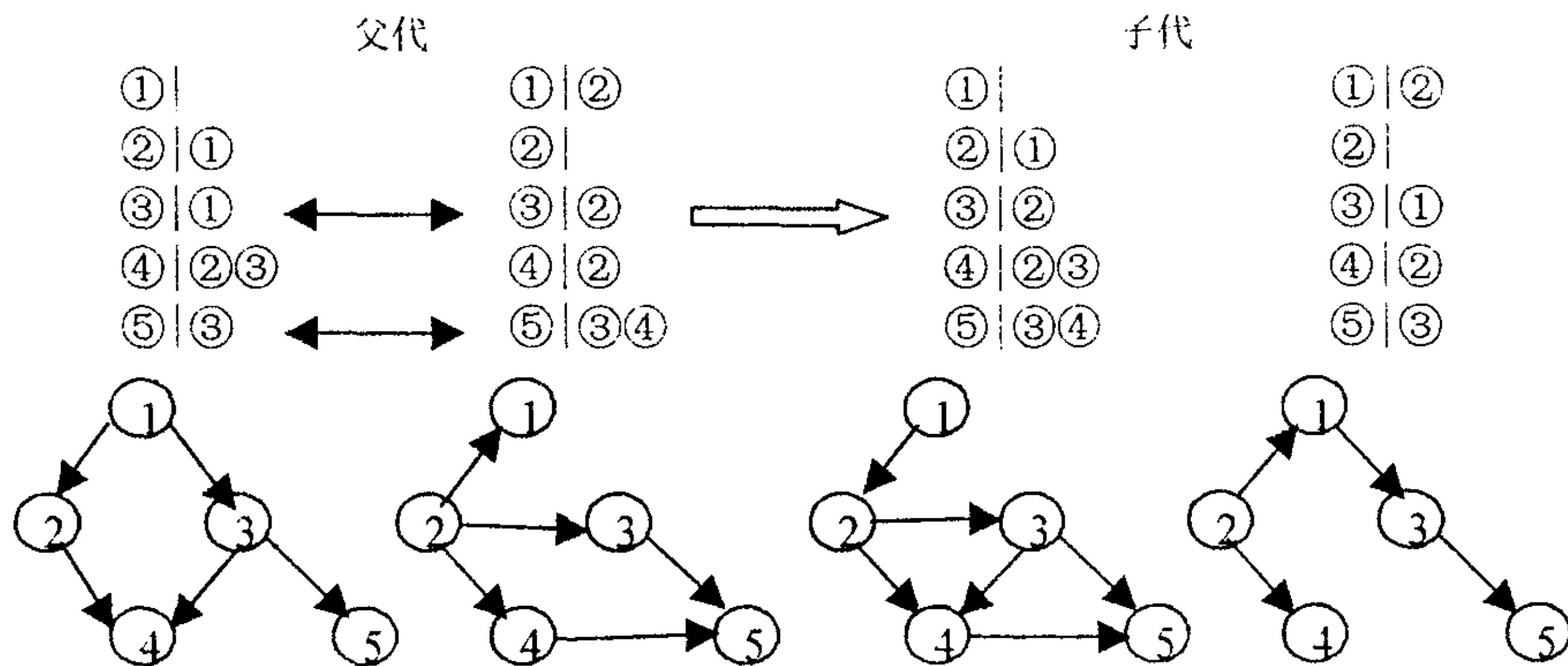


图 5 网络结构的交叉操作

3.4 染色体的变异

与交叉方法一样,如果只使用一种变异方法,同样可能会引起“早熟”。因此,依据选择性集成思想^[7],选择以下三种个性好且差异性较大的变异方法,等概率使用以扩大搜索范围,避免过早收敛:

(1)增加弧。即随机选取两个结点 i 和 j ,若两点间没有弧,则添加一条 $i \rightarrow j$ 或 $j \rightarrow i$ 的弧。

(2)删去弧。即随机选取两个结点 i 和 j ,若两点间存在 $i \rightarrow j$ 或 $j \rightarrow i$ 的弧,则删去此弧。

(3)弧反向。即随机选取两个结点 i 和 j ,若两点间存在 $i \rightarrow j$ 或 $j \rightarrow i$ 的弧,则将此弧反向。

3.5 交叉和变异时异常情况的处理和优化

贝叶斯网络结构是有向无环的,但在执行交叉和变异操作后,往往会产生有回环的网或非贝叶斯网。对此等概率采用以下两种方法:

(1)用上代种群中最好的贝叶斯网代替;

(2)随机产生一个贝叶斯网代替。

为避免固定的交叉、变异概率在迭代过程中使种群中的个体适应度趋向一致,出现“近亲繁殖”,对群体的进化会产生不利的影响。文中采用自适应交叉、变异概率,交叉概率和变异概率随进化的进行自适应调整,种群中个体性能提高时交叉概率提高,反之则变异概率增加。

3.6 测度函数与适应度值的计算

1978 年 Schwarz^[2,3]提出了贝叶斯信息标准测度(BIC, Bayesian information criterion)来评价贝叶斯网络结构。其公式为: $Q_{BIC}(B, D) = LL(B | D) - 1/2 \log N * \text{Dim}(B)$,其中, B 表示所学得的贝叶斯网络结

构, D 是训练数据集, $LL(B | D) = \sum_{i=1}^n \log p(B | D)$ 是基于概率分布描述 D 所需要的比特数的度量, $1/2 \log N$ 表示每一个参数使用的比特数,贝叶斯网络的维

度 $\text{Dim}(B) = \sum_{i=1}^n (r_i - 1) q_i = \sum_{i=1}^n (r_i - 1) \prod_{X_j \in \text{pa}(X_i)} r_j$

是指明随机变量 X 的联合概率分布所需要的自由参数的数目。在文中算法中,先求所有个体 BIC 测度总

和,再求各个体 BIC 测度值与 BIC 测度总和之比,即为每个个体的适应度值,然后进行遗传算法的操作。

4 基于 BIC 测度和混合遗传算法的 BNC 结构学习算法及实验结果分析

4.1 实验步骤

实验步骤分为以下几步:

(1)初始化,根据数据集大小生成第一代随机的贝叶斯网络种群,种群规模视数据集而定,一般取 30~100。

(2)根据每个个体对应的结点顺序,使用 GS 算法学习拟合当前数据集训练集的贝叶斯网络结构。

(3)对生成后的贝叶斯网络结构进行 BIC 测度打分,并将 BIC 测度值转换为相应的适应度值。

(4)循环对上一代种群进行选择、交叉、变异,初始的交叉、变异概率分别取 0.9 和 0.05,循环代数取 200 代。选择算子选用竞争选择算子,两种交叉和三种变异算子均等概率使用。每循环一次,交叉概率增加 1/10000,变异概率减少 1/10000,中止条件是直到 200 代或 20 代内最佳 BIC 测度无变化。

(5)输出最佳个体对应的贝叶斯网络结构,进行下一步参数学习、推理和构建分类器进行分类。

4.2 实验结果分析

实验是在 MBNC(Bayesian Classifier using Matlab)实验平台^[8,9]上进行的,实验数据集取自 UCI^[10](University of California in Irvine)。数据集的选取与分类器准确性评估方法与文献[2]一致(见表 1)。

不同的数据集是在完全同等的环境下进行运算的,第 3 列是文献[2]中 NBC 实验结果,第 5 列是文献[2]中 TANC 实验结果。空格表示文献[2]中没有列出的数据。

实验结果如表 2 所示。所有算法的参数学习均采用 BDeu 先验,先验值 priors 的值取 1,推理算法采用全局联合树推理算法。

表 1 数据集的概况

数据集	属性	类别	训练	测试	数据集	属性	类别	训练	测试
Australian	14	2	690	CV5	Glass	9	7	214	CV5
Breaks	9	2	683	CV5	Glass2	9	2	163	CV5
Car	6	4	1880	CV5	Heart	13	2	270	CV5
Cleve	10	2	296	CV5	Hepatitis	19	2	80	CV5
Corral	6	2	128	CV5	Iris	4	3	150	CV5
Crx	15	2	653	CV5	Mofn3-7-10	10	2	300	1024
Diabetes	8	2	768	CV5	Nursery	8	2	11025	CV5
Flare	10	2	1066	CV5	Pima	5	2	768	CV5

表 2 实验结果数据

数据集	NBC	文献[2] NBC	TANC	文献[2] TANC	BNC- GS	GAGS- BIC
Australian	86.94	86.23	84.93	81.30	86.66	85.65
Breaks	97.65	97.36	96.91	95.75	97.06	96.91
Car	87.39		91.28		91.28	92.55
Cleve	83.39	82.76	80.34	79.06	83.05	83.39
Corral	86.4	85.88	99.2	95.32	100	100
Crx	87.54	86.22	85.39	83.77	87.07	86.92
Diabetes	77.78	74.48	76.99	75.13	77.78	77.78
Flare	80.56	79.46	83.1	82.74	82.54	82.44
Glass	75.71		68.57		73.7	73.33
Glass2	85		83.75		70.48	82.50
Heart	83.33	81.48	83.7	82.96	75.63	84.44
Hepatitis	90	91.25	86.25	85.00	83.33	83.75
Iris	94		95.33		86.25	95.33
Mofn3-7-10	86.62	86.42	90.14	91.70	89.63	93.56
Nursery	88.44		90.75		91.26	91.63
Pima	79.22	75.51	76.73	75.13	77.51	78.04
平均值	85.62		85.84		84.58	86.76

由表 2 可知,NBC 和 TANC 的正确率均比文献 [2]结果高,BNC 在未与遗传算法结合前,其分类准确率低于 NBC 和 TANC,但与遗传算法结合后,准确率明显提高。实验结果表明基于 MBNC 实验平台上设计的 3 类贝叶斯分类器是有效和正确的。通过多次反复实验,还发现,GAGS-BIC 算法基本上克服了 NBC 和 TANC 以及 BNC-K2 的分类结果的波动性。说明 GAGS-BIC 算法真正收敛到了当前算法下的全局最优。

(上接第 83 页)

布斯效应。实验结果显示,文中提出的图像去噪方法能有效地去除叠加在图像上的高斯白噪声,更好地保护图像细节,显著提高去噪图像的 PSNR 值。

参考文献:

[1] Donoho D L. De-noising by soft-thresholding[J]. IEEE Trans Information Theory, 1995, 41(3): 613-627.
 [2] Do M N. Directional multiresolution image representation[D]. EPFL, Lausanne, Switzerland: [s. n.], 2001.
 [3] Do M N, Vetterli M. Contourlets: A Directional Multiresolu-

5 结束语

贝叶斯网络分类器的结构学习是贝叶斯网络应用的难点之一。在分析了学习贝叶斯网络结构的 GS 算法和遗传算法各自优点和不足的基础上,将这两种算法相结合,提出了基于 BIC 测度的混合遗传算法的 Bayesian 网络分类结构学习算法,实现了它们的优势互补。实验表明:该算法是正确的和有效的,而且具有鲁棒性。

参考文献:

[1] Mitchell T. Machine Learning[M]. [s. l.]: McGraw-Hill Companies Inc, 1997.
 [2] Friedman N, Goldszmidt M. Building classifiers using Bayesian network[C]// In proc. Nation Conference on Artificial Intelligence. Menlo park, CA: AAAI Press, 1996: 1227-1284.
 [3] Cooper G, Herskovits E. A Bayesian method for the induction of probabilistic networks from data[J]. Machine Learning, 1992, 9: 309-347.
 [4] 林士敏, 田凤占, 陆玉昌. 用于数据采掘的贝叶斯分类器研究[J]. 计算机科学, 2000, 27(10): 73-76.
 [5] Larranaga P, Poza M, Yurramendi Y et al. Structure Learning of Bayesian networks by genetic algorithms: A performance analysis of control parameters[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1996, 18(9): 912-925.
 [6] 刘大有, 王 飞, 卢奕南, 等. 基于遗传算法的 Bayesian 网结构学习研究[J]. 计算机研究与发展, 2001, 38(8): 916-922.
 [7] 周志华. 选择性集成(Selective Ensemble)[C]// 第九届中国机器学习会议. 上海: 复旦大学, 2004.
 [8] 程泽凯, 林士敏, 陆玉昌, 等. 基于 Matlab 的贝叶斯分类器实验平台 MBNC[J]. 复旦学报, 2004(5): 729-732.
 [9] 陆小艺, 程泽凯, 林士敏. 用 Matlab 语言建构贝叶斯分类器[J]. 微机发展, 2004, 14(9): 33-35.
 [10] Blake C, Keogh E, Merz C. UCI repository of machine learning database[EB/OL]. 2006-02-23. <http://www.ics.uci.edu/mllearn/MLRepository.html>, 1998.

tion Image Representation[C]// Proc of IEEE International Conference on Image Processing. Rochester, NY: [s. n.], 2002: 357-360.

[4] Donoho D L. Wavelet Thresholding and W. V. D: A 10-minute Tour[C]// Int Conf on Wavelets and Applications. Toulouse, France: [s. n.], 1992.
 [5] Coifman R R, Donoho D L. Translation Invariant Denoising [C]// Wavelets and Statistics, Springer Lecture Notes in Statistics 103. New York: Springer-Verlag, 1995: 125-150.