

一种基于遗传算法的受限制的分类器学习算法

董立岩¹,苑森森¹,刘光远²,李永丽³,关伟洲³

(1. 吉林大学 计算机科学与技术学院, 长春 130012; 2. 吉林大学 通信工程学院, 长春 130022; 3. 东北师范大学 计算机学院, 长春 130024)

摘要:提出了一种基于遗传算法的受限制 BAN 分类器算法—GBAN(genetic algorithm based BAN)。新算法采用了遗传算法进行网络结构的学习,限制了所学习的 BAN 分类器结构的复杂度。同时对 TAN 分类器的结构进行了扩展,得到了一种受限制的 BAN 分类器。针对这种分类器的结构学习,设计了结合对数似然的适应度函数及相应的遗传算子,并给出了网络结构的编码方案,使得该算法能够收敛到全局最优的结构。实验结果表明,当数据集属性之间关系相对复杂的时候,GBAN 比 TAN 的分类准确率高,分类效果较好。

关键词:人工智能;贝叶斯网络分类器;遗传算法;对数似然

中图分类号:TP182 **文献标识码:**A **文章编号:**1671-5497(2007)03-0595-05

Constrained classifier learning algorithm based on genetic algorithm

Dong Li-yan¹, Yuan Sen-miao¹, Liu Guang-yuan², Li Yong-li³, Guan Wei-zhou³

(1. College of Computer Science and Technology, Jilin University, Changchun 130012, China; 2. College of Communication Engineering, Jilin University, Changchun 130022, China; 3. College of Computer, Northeast Normal University, Changchun 130024, China)

Abstract: A restricted BAN classifier learning algorithm — GBAN based on genetic algorithm is proposed. Genetic algorithm was used to study the network structure. The structure of TAN classifier was extended by restricting the complexity of the structure of BAN classifier. And then a restricted BAN classifier is obtained. As far as this classifier's structure studying, the fitness function based on logarithm likelihood was designed. The code scheme of network structure, and the corresponding genetic operators are designed. As a result, the algorithm converges on the overall optimal structure. The experimental result indicated that GBAN algorithm has good classifying effect and is more accurate than TAN classifier when the relationship between attributes of a data set is relatively complicated.

Key words: artificial intelligence; bayesian network classifier; genetic algorithm; logarithm likelihood

朴素贝叶斯分类器是目前公认的一种简单有效的概率分类方法,在独立性假设下表现出很好的性能^[1,2]。所谓“独立性假设”是指给定一个实例的类标签,实例中每个属性的出现独立于实例

收稿日期:2006-12-21.

基金项目:国家自然科学基金资助项目(60275026).

作者简介:董立岩(1966-),男,副教授,博士研究生.研究方向:数据库,数据挖掘. E-mail:dongly@jlu.edu.cn

通讯联系人:苑森森(1943-),男,教授,博士生导师.研究方向:数据库,人工智能及计算机网络系统.

E-mail:yuansenmiao@126.com

中其他属性的出现。该独立性假设使得朴素贝叶斯方法相对于其他分类方法有一个独特之处,即它不需要搜索,只需简单地计算训练数据集中各个属性值发生的频率,就可得出每个属性的参数。然而,在大多数现实世界中,这个独立性假设很难成立。TAN 分类器(Tree Augmented Naive Bayes)^[3]使用 Chow-Liu 的最大权重跨度树算法学习属性节点之间的结构,得到了最大似然树,在放松了 Naive Bayes 对属性节点间条件独立性假设的同时,保持了结构的简单性和优越性。然而对于属性之间的关系更加复杂的数据集,TAN 分类器的准确率有很大的下降。BAN 分类器(Bayes Network Augmented Naive-Bayes)^[4]对朴素贝叶斯网络的条件独立性假设做了进一步的放松,在对于数据集的各个属性具有较复杂的关系时,BAN 分类器有着更优秀的学习能力。然而,BAN 分类器的学习过程实质上是受限制的贝叶斯网络学习的过程,其学习算法存在搜索空间较大,计算复杂,而且容易陷入局部最优解的问题,影响了 BAN 分类器的分类效率。作者结合贝叶斯理论的基本原理及遗传算法在最优搜索方面的优点提出了一种改进的 TAN 分类器算法即 GBAN(genetic algorithm based BAN)分类器算法。

1 GBAN 算法学习得到的分类器

GBAN 算法是一种基于遗传算法的 BAN 分类器学习算法。GBAN 对学习得到的网络结构进行了以下的限制:设 A_1, \dots, A_n 是属性节点, C 是分类节点,GBAN 学习得到这样一类网络结构:①分类节点 C 是所有属性节点的父节点。②属性节点之间构成贝叶斯网络。③对于任意属性节点 $A_i, i=1, \dots, n$,限制 A_i 除了分类节点 C 是其父节点外,只能有不超过 m 个父节点(一般取 $m \leq 4$ ^[6])。

由此可见,GBAN 学习所得到的分类器是一种受限制的 BAN 分类器,也可以看作是对 TAN 分类器的一种扩展,当限制 $m=1$ 并且属性节点 A_1, \dots, A_n 构成最大权重跨度树时,GBAN 算法学习得到的是 TAN 分类器。

GBAN 学习过程本质上是一种受限制的贝叶斯网络学习过程。贝叶斯网络学习包括结构学习和参数学习两部分,由结构和数据集可确定参数,因此贝叶斯网络结构学习是贝叶斯网络学习

的核心。对于给定的数据集 D ,结构学习是为找到与 D 匹配程度最高的贝叶斯网络。通常的做法是定义一个评分函数,评判某个具体结构反映的独立关系和样本的匹配程度,选择适宜的搜索算法搜索分值最高的网络模型。

变量集 X 上的可能 Bayesian 网络结构有 $n! \cdot 2n$ 种,搜索空间非常巨大,完全搜索是 NP 问题,并且局部最优的网络结构可能有多个。遗传算法具有在复杂空间近似求解全局最优解的能力,很自然地适合解决 Bayesian 网的学习问题,非常适合 BAN 分类器的学习。

2 GBAN 算法

2.1 适应度函数

GBAN 算法适应度函数为

$$\sum_{i, \pi(i) > 0} I_{\hat{P}_D}(A_i; A_{\pi(i)} | C) \tag{1}$$

设 $\hat{P}_D(\cdot)$ 表示事件在 D 中出现的频度的经验分布,即 $\hat{P}_D(A) = \frac{1}{N} \sum_j l_A(u_j)$,对于每一个事件 $A \subseteq Val(U)$,当 $u \in A$ 时, $l_A(u) = 1$ 当 $u \notin A$ 时, $l_A(u) = 0$ 。

A_i 表示属性节点, $A_{\pi(i)}$ 表示属性节点 A_i 除了分类节点 C 以外的父节点集合。

Friedman^[7]证明,常用的 MDL 和贝叶斯打分函数等非限定的打分函数不适合表示良好的贝叶斯分类器。TAN 分类器有着很高的准确率和效率,一个最大的原因就是 TAN 学习得到的网络结构能够最大化对数似然函数,作为 TAN 分类器的一种受限制的扩展,GBAN 算法得到的分类器采用对数似然函数的一个推出式作为评价结构好坏的标准,也就是作为 GBAN 算法的适应度函数,并以最大化该适应度作为进化的目标。

定义 1 设 D 表示大小为 N 、包含属性 C, A_1, \dots, A_n 的数据集,从数据集 D 学习得到的网络结构是 B_S ,则对数似然可定义为

$$LL(B_S | D) = \sum_{i=1}^N \log(P_B(u_i)) \tag{2}$$

$LL(B_S | D)$ 可以分解成

$$LL(B_S | D) = N \sum_{X_i} I_{\hat{P}_D}(X_i; \Pi_{X_i}) + \text{常数项} \tag{3}$$

所以只需要对 $\sum_{X_i} I_{\hat{P}_D}(X_i; \Pi_{X_i})$ 求最大值即可以最大化 $LL(B_S | D)$ 。

GBAN 学习得到的结构可以由 $\pi(\cdot)$ 来定义: 方程 $\pi: \{1, \dots, n\} \rightarrow \{0, \dots, m\}$ 是定义在节点集 $\{C, A_1, \dots, A_N\}$ 上由 $\{1, \dots, n\}$ 到 $\{0, \dots, m\}$ 的映射。 $\pi(i)$ 表示 A_i 节点父节点的个数(不计算分类节点 C), 令 $\pi(C) = 0$ 。不存在这样一个序列, i_1, \dots, i_k , 使得 $\pi(i_j) = i_{j+1}, i \leq j \leq k, \pi(i_k) = i_1$, 即不存在有向环。由于分类节点 C 没有父节点, 所以 $I_{\hat{P}_D}(C; \Pi_C) = 0$ 。

$$I_{\hat{P}_D}(A_i; \Pi_{A_i}) = \begin{cases} I_{\hat{P}_D}(A_i; A_{\pi(i)}, C) & \text{如果 } \pi(i) > 0 \\ I_{\hat{P}_D}(A_i; C) & \text{否则} \end{cases}$$

所以 $\sum_{X_i} I_{\hat{P}_D}(X_i; \Pi_{X_i})$ 在 GBAN 学习的结构中可以表示为

$$\sum_{i, \pi(i) > 0} I_{\hat{P}_D}(A_i; A_{\pi(i)}, C) + \sum_{i, \pi(i) = 0} I_{\hat{P}_D}(A_i; C) \quad (4)$$

根据互信息的链式规则

$$I_P(X; Y, Z) = I_P(X; Z) + I_P(X; Y | Z)$$

由式(4)可以推出

$$\sum_{i, \pi(i) > 0} I_{\hat{P}_D}(A_i; A_{\pi(i)} | C) + \sum_i I_{\hat{P}_D}(A_i; C) \quad (5)$$

由于式(5)中的第二项 $\sum_i I_{\hat{P}_D}(A_i; C)$ 的值与各个节点的父节点的选择 $\pi(i)$ 无关, 即与生成的网络结构无关, 所以只要生成的网络结构能够最大化, 式(5)的第一项 $\sum_{i, \pi(i) > 0} I_{\hat{P}_D}(A_i; A_{\pi(i)} | C)$ 就能够达到最大化 $LL(B_S | D)$ 的目的。

由于 GBAN 算法使用了遗传算法, 并采用式(5)的第一项 $\sum_{i, \pi(i) > 0} I_{\hat{P}_D}(A_i; A_{\pi(i)} | C)$ 作为适应度函数, 根据遗传算法的收敛特性, 算法 GBAN 会以概率为 1 收敛到所有 GBAN 所限制的结构中最大化对数似然 $LL(B_S | D)$ 的结构。

2.2 编码方法

GBAN 算法中个体对应着贝叶斯分类器的网络结构 S , 用邻接矩阵 $C = (C_{ij})$ (其中 $i, j = 1, \dots, n$) 对网络结构 S 进行编码, 编码方法为

$$C_{ij} = \begin{cases} 1, & \text{如果存在边 } i \rightarrow j \\ 0, & \text{否则} \end{cases} \quad (6)$$

由于分类节点是所有属性节点的父节点, 而且在 GBAN 算法的进化过程中这种特性不会发生改变, 即不参与遗传操作, 所以邻接矩阵 C 中不对分类节点进行编码。采用这种编码, 遗传算法的种群中一个个体可以用 0/1 字符串来表示:

$$C_{11} C_{21} \dots C_{n1} C_{12} C_{22} \dots C_{n2} \dots C_{1n} C_{2n} \dots C_{nn}$$

邻接矩阵(被表示为 $C_{11} C_{21} \dots C_{n1} C_{12} C_{22} \dots C_{n2} \dots C_{1n} C_{2n} \dots C_{nn}$ 的字符串)可以看作一个染色体, 其中每一行是一个基因, $C_{1i} C_{2i}, \dots, C_{ni}$ 相当于等位基因。

2.3 遗传操作

(1) 选择

采用等级比例法^[8]对个体进行选择。该方法按适应度大小将个体分成不同的等级, 每个等级的选择概率不同。这样选择概率与个体适应度的等级有关, 与适应度的绝对大小无关, 避免了超常个体选择概率过大, 出现早熟现象。

如果 S_j^t 表示第 t 代第 j 个个体, 那么 $\text{rank}(F(S_j^t))$ 表示 S_j^t 适应度的等级, 则 S_j^t 被选择的概率 $p_{j,t}$ 为(其中 λ 表示初始群体规模)

$$p_{j,t} = \frac{\text{rank}(F(S_j^t))}{\lambda(\lambda + 1)/2} \quad (7)$$

(2) 交叉

选择操作选出的个体随机配对, 并按概率 p_c 决定是否进行交叉操作。本文采用普通的单点交叉的方法。

(3) 变异

变异操作是随机改变个体的性状, 增加群体的多样性以避免局部极值。具体做法是按接近于 0 的变异概率 p_m 随机改变某些个体串的某些基因座的基因值。本文采用均匀变异。

个体在经过交叉和变异的遗传操作后可能产生出非法个体, 这种非法个体有两种情况, 一种是个体不再是有向无环图 DAG, 另一种是个体中某些节点(属性节点)包含多于 m 个父节点(不计算分类节点)。为此引入修复操作。

(4) 修复

对于个体不再是有向无环图 DAG, 在每次遗传操作后对新产生的个体进行 DAG 的检验, 如果该个体不是 DAG, 则简单地取消刚刚执行的遗传操作。

由于在表示个体的邻接矩阵中寻找最佳的或者近似最佳的边的子集使得该个体的结构为 DAG 需要很大的开销, 如果在每步遗传中均重复这种基本操作, 会使遗传算法的效率很低; 同时, 没有从群体中删除这样的个体可以保留该个体局部优秀的结构, 该结构可以遗传给下一代个体。

对于某些节点(属性节点)包含多于 m 个父节点(不计算分类节点)的个体, 选择这种节点的父

节点集合中最优的 m 个父节点,删除其他的父节点指向该节点的边。

2.4 初始群体的设定

采用随机生成初始群体的方法,并使用 2.3 节所介绍的方法对初始群体进行有向无环性和父节点个数的限制。

因为适应度的计算复杂度高,所以群体规模 λ 不能太大。另一方面,为避免未成熟收敛,群体规模不能太小,一般选在 10-100 之间。算法终止条件可以选做已经进化产生了 10 000 个后代或连续的 g 代最佳的网络结构没有变化(g 是常数,一般取为 20)。控制网络属性节点的父节点个数 m 的选择也根据所要学习的数据库的属性个数,数据量等信息做调整,但始终保持 $m \leq 4$ 。

2.5 GBAN 算法描述

PROCEDURE GBAN /* λ 表示群体规模, t 表示当前是第几代 */

BEGIN

(1) 随机初始化初始群体 $pop(0)$ 。

(2) 对 $pop(0)$ 中父节点个数超过 m 和非 DAG 的个体 I_0 ,按照 2.3 节所述修复操作的方法转化为合法的结构。

(3) 对于所有 $pop(0)$ 中的个体,计算其适应度, $t=0$ 。

(4) WHILE NOT stop_condition DO

BEGIN。

(5) 执行选择操作,选择出用于生成下一代的父亲 $parent(t)$ 。

(6) 对于父亲 $parent(t)$ 中的个体执行交叉操作,然后执行修复操作。

(7) 对于第(6)步产生的新个体,先后执行变异操作和修复操作。

(8) 将经过第(6),(7)步产生的新个体加入到 $pop(t)$ 中生成 $pop(t+1)$,并采用精英约减标准使 $pop(t+1)$ 的大小减少到 λ 。

(9) $t=t+1$,保存第 t 代最好的个体。

END

(10) 输出最好的个体。

END

3 实验结果

作者选用 UCI 数据库^[9]中提供的 11 个数据集作为实验数据集,并使用 Jie Cheng 提供的数据预处理工具 PreProcessor^[10]进行了离散化和丢失数据的预处理。采用分类错误率作为评价分类器性能的指标,针对大数据集采用 hold out(1/3)方法,小数据集采用 5-fold 方法进行测试,分别进行模型评估。采用上述方法比较了 GBAN 算法和朴素贝叶斯分类器、TAN 分类器的性能,实验结果如图 1,2 所示。

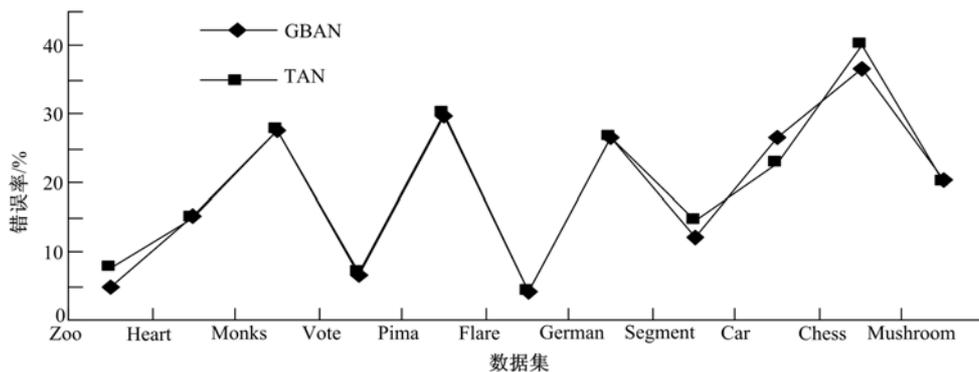


图 1 GBAN 和 TAN 分类错误率曲线图

Fig. 1 Classification error rate of GBAN and TAN

从图 1 的错误率曲线图中可看出,在属性个数较多的数据集,如 Zoo, Vote, Segment, Chess 等数据集,GBAN 分类性能比 TAN 的分类性能要好。而在其他的数据集上,GBAN 的分类性能与 TAN 分类性能相当。同时注意到每个点都偏离不是很远,说明 TAN 和 GBAN 对相同的数据集分类性能差距很小。

从图 2 的错误曲线图中可以看出,对于大多数数据集,GBAN 的分类性能比 NB 的分类性能好,并且图中表示两者的错误率的某些点偏离较远,这说明对于某些数据集,特别是属性变量之间存在较强依赖关系的数据集(偏离较远的点表示的数据集),GBAN 比 NB 的分类性能要好得多。

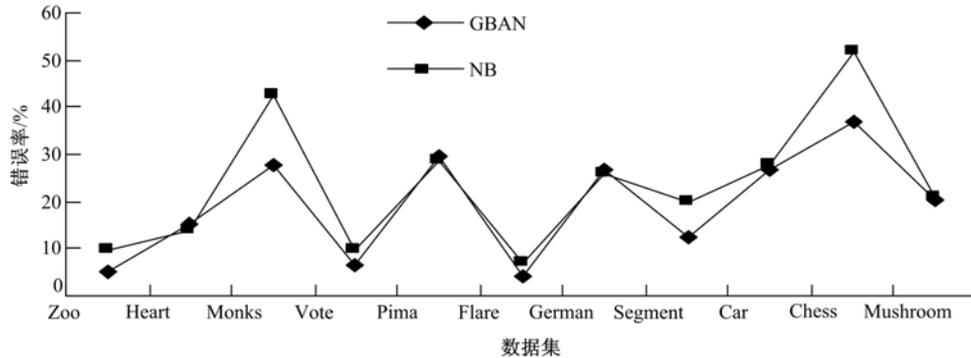


图2 GBAN 和 NB 分类错误率曲线图

Fig. 2 Classification error rate of GBAN and NB

4 结束语

提出了一种基于遗传算法的受限制的 BAN 分类器学习算法 GBAN, 实验结果表明, 在属性间关系复杂的数据集上, GBAN 分类性能比 TAN 分类性能要好。进一步的研究目标是优化适应度函数的计算以提高效率。

参考文献:

- [1] Bouckaert R. Naive bayes classifiers that perform well with continuous variables[C] // Proc Seventeenth Australian Joint Conference on Artificial Intelligence (AI 2004), Advances in Artificial Intelligence. Cairns, Australia: Springer, 2004: 1089-1094.
- [2] Remco R Bouckaert. Bayesian network classifiers in weka[DB/OL]. <http://citeseer.ist.psu.edu/705669.html>, 2005-04-21.
- [3] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers[J]. Machine Learning, 1997, 29(2/3): 131-161.
- [4] Cheng J, Greiner R. Comparing Bayesian network classifiers[C] // Proc of the 15th Conf on Uncertain-

ty in Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1999: 101-108.

- [5] Holland J H. Adaptation in Natural and Artificial Systems[M]. Cambridge: MIT Press, 1992.
- [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.
Liu Da-you, Wang Fei, Lu Yi-nan, et al. Research on learning bayesian network structure based on genetic algorithms[J]. Journal of Computer Research and Development, 2001, 38(8): 916-922.
- [7] Rissanen J. Modeling by shortest data description[J]. Automatica, 1978, 14(5): 465-471.
- [8] Whitley D. The genitor algorithm and selection pressure: why rank-based allocation of reproductive trials is best[C] // Proceedings of the Third International Conference on Genetic Algorithms. San Francisco: Morgan Kaufm Publish, 1989: 116-121.
- [9] Murphy P M, Aha D W. UCI repository of machine learning databases[DB/OL]. <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [10] Cheng J. Power constructor system[DB/OL]. <http://www.cs.ualberta.ca/~jcheng/bnpc.htm>.