

基于捕食逃逸 PSO 的贝叶斯网络分类器

孔宇彦¹, 姚金涛², 李强¹, 祝胜林², 张明武²

(1. 南海东软信息技术职业学院 计算机系, 广东 佛山 528225; 2. 华南农业大学 信息学院, 广州 510642)

(justin_yjt@163.com)

摘要:构造精确的贝叶斯网络分类器已被证明为 NP 难问题,提出了一种基于捕食逃逸粒子群优化(PSO)算法的通用贝叶斯网络分类器,能有效避免数据预处理时的属性约简对分类效果的直接影响,实现对贝叶斯网络结构的精确学习和搜索。另外,将所提出的分类器应用于高职院校就业预测分析,并在 Weka 平台上实现对该分类器的构建和验证,与其他几种贝叶斯网络分类器的对比实验结果表明,该分类器具有更好的性能。

关键词:捕食逃逸;粒子群优化;贝叶斯网络分类器;Weka;就业预测

中图分类号: TP301; TP18 **文献标志码:** A

Bayesian network classifier based on PSO with predatory escape behavior

KONG Yu-yan¹, YAO Jin-tao², LI Qiang¹, ZHU Sheng-lin², ZHANG Ming-wu²

(1. Department of Computer, Nanhai Neusoft Institute of Information Technology, Foshan Guangdong 528225, China;

2. College of Information, South China Agricultural University, Guangzhou Guangdong 510642, China)

Abstract: Bayesian network classifier with precise structure has been proven to be NP-hard problem. A Bayesian network classifier based on Particle Swarm Optimization-Predatory Escape (PSO_PE) algorithm was proposed in this paper, which could effectively avoid the direct influence of feature reduction on the performance of classification and complete the precise learning Bayesian network. In addition, the proposed classifier was exploited in employment predication of vocational college and was experimentally tested on Weka. The experimental results show that compared with other Bayesian classifiers, the new classifier is more effective and precise to learn Bayesian network.

Key words: predatory escape; Particle Swarm Optimization (PSO); Bayesian Network Classifier (BNC); Weka; employment predication

0 引言

贝叶斯网络分类器(Bayesian Network Classifier, BNC)^[1]是建立在贝叶斯概率论基础上的基于统计方法的分类模型,可以预测类成员关系的可能性,即计算给定样本属于一个特定类的概率。贝叶斯网络分类器的一个主要优点是在用于大型数据库时,表现了高准确率与高速度。应用贝叶斯网络分类器进行分类主要分成两阶段:1)贝叶斯网络分类器的学习,即从样本数据中构造分类器,包括结构学习和条件概率表(Conditional Probability Table, CPT)学习;2)贝叶斯网络分类器的推理,即计算类节点的条件概率,对分类数据进行分类。这两个阶段的时间复杂性均取决于特征值间的依赖程度,甚至可以是 NP 完全问题,因而在实际应用中,往往需要对贝叶斯网络分类器进行简化。根据对特征值间不同关联程度的假设,可以得出各种贝叶斯网络分类器,如朴素贝叶斯(Naive Bayes, NB)、树扩展的朴素贝叶斯(Tree Augmented Naive Bayes, TANB)、增强的贝叶斯网络分类器(Bayesian Network Naive-Bayes, BAN)、一般贝叶斯网络(General Bayesian Network, GBN)是其中较典型且研究较深入的贝叶斯网络分

类器^[2]。

粒子群优化(Particle Swarm Optimization, PSO)算法^[3]是一种模拟简化社会模型的迭代优化方法,其基本思想是将每个粒子看成问题解空间中的候选解,然后通过更新粒子的速度和位置来改变粒子在解空间中的位置,迭代探索最优解。与遗传算法比较,PSO 没有如交叉和变异等遗传操作,而是通过所有粒子在解空间追随最优粒子进行搜索,只有 *gBest* 把信息单向流动给其他粒子,即整个搜索更新是跟随当前最优解的过程,因此在大多数情况下,所有粒子则可能更快地收敛于最优解。因其概念简单、易于实现、无需梯度信息、参数少、能有效解决复杂优化问题等特点而引起学术界的广泛重视,成为国际上智能优化领域研究的热点,目前已被成功应用于多目标优化、模式识别、信号处理和决策支持等领域,具有良好的应用前景。

目前,构造精确的贝叶斯网络分类器已被证明为 NP 问题^[4],因此可以将 PSO 算法引入贝叶斯网络分类器中,通过在训练集上随机属性选取生成若干属性子集,并以这些子集构建相应的贝叶斯网络分类器,然后用 PSO 算法对分类器进行优选,能有效避免数据预处理时的属性约简对分类效果的

收稿日期:2010-08-02; **修回日期:**2010-09-15。 **基金项目:**广东省科技计划项目(2009b020315014);广东省育苗工程项目(wym09066);南海东软信息学院院立科研基金资助项目(NN0811018R)。

作者简介:孔宇彦(1978-),女,湖南长沙人,讲师,硕士,主要研究方向:数据挖掘、数据库;姚金涛(1978-),男,山东临沂人,博士,主要研究方向:进化计算、通信与信息安全;李强(1976-),女,湖南永丰人,讲师,硕士,主要研究方向:数据挖掘、数据库技术;祝胜林(1969-),男,江西进贤人,副教授,博士,主要研究方向:通信与信息安全;张明武(1970-),男,湖北仙桃人,副教授,博士,主要研究方向:通信与信息安全。

直接影响。但是,早熟收敛、收敛效率低、全局收敛性差依然是 PSO 算法面临的一大难题,其主要原因是群体最优 $gBest$ 的唯一支配性信息共享模式无法对称调整社会认知能力,因此,通过在群体中引入捕食粒子来增大逃逸粒子的捕食风险,使各逃逸粒子根据捕食风险和自身能量状态的权衡结果产生相应逃逸行为,提高了粒子群对称调整社会认知能力,进而有效地保持群体多样性,平衡群体的探索和开发能力,使群体避免陷入早熟收敛。利用改进后的捕食逃逸 PSO 算法提出一种基于捕食逃逸粒子群优化(Particle Swarm Optimization with Predatory Escape, PSO_PE)算法的贝叶斯网络分类器,则能更加有效完成对贝叶斯网络分类器模型的精确构建。

为了进一步验证基于 PSO_PE 的贝叶斯网络分类器的分类预测效果,本文将具体应用于高职院校就业预测分析,结果表明采用该分类器所构造的就业预测模型能较准确地预测某毕业生能否就业或某一年整个学校的就业率,对学校 and 毕业生个人来说都是一个很有价值的信息。可见,本文所提出的贝叶斯网络分类器是有效的。

1 贝叶斯网络分类器

贝叶斯网络又称为信念网络(Belief Network, BN),其详细的定义为^[1]:

设 $V = \{X_1, X_2, \dots, X_n\}$ 是值域 U 上的 n 个随机变量,则值域 U 上的贝叶斯网络定义为 $BN(B_s, B_p)$, 其中: B_s 是一个定义在 V 上的有向无环图(Directed Acyclic Graph, DAG) Γ , V 是该有向无环图 Γ 的节点集, E 是 Γ 的边集。如果存在一条节点 X_j 到节点 X_i 的有向边,则称 X_j 是 X_i 的父节点, X_i 是 X_j 的子节点,记的所有父节点为 π_i ; 而 $B_p = P(X_i/\pi_i[0,1] | X_i \in V)$ 。对于 V 中每个节点,定义了一组条件概率分布函数 $P(X_i/\pi_i[0,1])$ 。贝叶斯网络是概率理论和图形理论的结合,主要由以下两部分构成。

1) 有向无环图 Γ , 通常称为贝叶斯网络结构。由若干个节点和连接这些节点之间的有向边组成,节点代表问题领域的随机变量,每个节点对应一个变量。变量的定义可以是问题中感兴趣的现象、部件、状态或属性等,具有一定的物理和实际意义。连接节点之间的有向边代表节点之间的依赖或因果关系,连接边的箭头代表因果关系影响的方向性(由父节点指向子节点),节点之间缺省连接则表示节点所对应的变量之间条件独立。

2) 局部概率分布集,通常称为条件概率表。概率值表示子节点与其父节点之间的关联强度或置信度,没有父节点的节点概率为其先验概率。贝叶斯网络结构是将数据实例抽象化的结果,是对问题领域的一种宏观描述;而概率参数是对变量(节点)之间关联强度的精确表达,属于定量描述的部分。

GBN 分类器与 NB、TAN、BAN 分类器的较大区别是,后者 3 类分类器中均将类变量所对应的节点作为一个特殊的节点,即是各特征节点的父节点,而 GBN 将分类节点当成和其他节点一样的属性节点来对待。这种结构学习需要获得一个完整的贝叶斯网络,它把分类问题看做一种特殊的推理过程或决策问题,本文主要讨论 GBN 分类器。

2 基于捕食逃逸 PSO 的贝叶斯分类器

2.1 捕食逃逸 PSO

2.1.1 PSO_PE 算法的基本原理

捕食与被捕食在生物界普遍存在,捕食风险的影响涵盖了几乎所有动物性的决策方式,被捕食者必须在自身能量状态和捕食风险间进行权衡,并最终做出行为上的改变。一般来说,被捕食者会回避捕食风险较高的搜寻区域,从而始终与捕食者保持一定的逃逸开始距离^[5](Flight Initiation Distance, FID),即允许捕食者靠近的最大距离。当双方距离等于或接近 FID 时,被捕食者将根据自身的能量状态做出不同反应,能量状态越大则逃逸速度越快;能量状态越小将承受越大的捕食风险缓慢移动或静止不动,被捕食概率加大。本文受捕食逃逸现象启发而提出的捕食逃逸粒子群优化算法把原单一粒子群分成捕食粒子群(Predator Swarm, PS)和逃逸粒子群(Escaping Swarm, ES)两个子群体。PS 粒子和 ES 粒子的行为将依据各自定义的简单规则加以约束,其中 PS 粒子追捕 ES 的 $gBest$ 粒子,因而对 ES 粒子造成不等的捕食风险,即 $gBest$ 粒子也能从 PS 粒子获取信息,实现了群体的对称社会认知。当 ES 粒子与 PS 粒子的距离接近 FID 时产生逃逸,逃逸速度依赖于自身的能量状态(适应度),能量越大逃逸能力越强;若 ES 粒子与 PS 粒子的距离小于 FID,则对 ES 粒子进行确定性变异,变异前后的 ES 粒子优胜劣汰。因而,在进化前期,算法具有较好的全局搜索能力。随着迭代增加,将逐步降低 PS 粒子对 ES 粒子的影响,以增加群体的局部搜索能力,进而平衡算法的局部和全局搜索能力,最终实现算法的全局收敛。

2.1.2 相关定义

定义 1 捕食风险也称为捕食压力是指在一定时间内 ES 粒子被捕食的概率,即:

$$P_i^{ES}(t) = \exp(-\alpha_i kt') \quad (1)$$

其中: α_i 表示 ES 粒子 i 与 PS 粒子相遇的概率,取决于它们之间的距离和当前 PS 粒子的密度,即 $\alpha_i = \exp\left(-\frac{distance}{n_1} \times \beta\right)$; β 为控制参数; n_1 为 PS 的规模; $distance$ 为 ES 粒子 i 与最近 PS 粒子之间的距离; k 表示 PS 粒子攻击 ES 粒子的概率(固定为 1); $t' = (t + T)/T$, t 为当前代数, T 为最大代数,迭代会逐步降低捕食粒子对被捕食粒子的影响。

定义 2 能量状态指 ES 粒子当前饥饿状态,表现为该粒子的适应度(考虑最小化问题)与 ES 平均适应度的比值,即:

$$E_i^{ES}(t) = \frac{f_i^{ES}(t)}{f_{avg}^{ES}(t)} \quad (2)$$

定义 3 警戒距离(Alert Distance, AD)反映了 ES 粒子对 PS 粒子的警惕能力,是一种普遍的社群现象,其大小随群体密度和群体规模增加而减小,即:

$$D^{ES} = FID \times \left(1 + \frac{n_1}{\rho \times n_2}\right) \quad (3)$$

其中: ρ 表示当前群体局部密度; n_1 、 n_2 分别表示 PS 粒子和 ES 粒子的规模。

2.1.3 算法描述

1) 随机产生并初始化 n_1 个 PS 粒子和 n_2 个 ES 粒子, $m = n_1 + n_2$, 设置 $t = 0$, β , FID 控制参数。

2) 计算每个粒子的适应度。

3) 对每个粒子 i , 将其适应度值与其历史最好位置 $pBest_i$ 进行比较, 更新 $pBest_i$ 和 $gBest$ 。

4) 对每个 PS 粒子 i , 按照式(4)更新其速度和位置:

$$\begin{aligned} V_{id}^{PS}(t+1) &= \omega V_{id}^{PS}(t) + c_1 r_1 (pBest_{id}^{PS}(t) - X_{id}^{PS}(t)) + \\ &\quad c_2 r_2 (gBest_{id}^{PS}(t) - X_{id}^{PS}(t)) \\ X_{id}^{PS}(t+1) &= X_{id}^{PS}(t) + V_{id}^{PS}(t+1) \end{aligned} \quad (4)$$

5) 对每个 ES 粒子 i :

①若 $distance_d \geq FID$, 按照式(5)更新其速度和位置:

$$\begin{aligned} V_{id}^{ES}(t+1) &= \omega V_{id}^{ES}(t) + c_1 r_1 (pBest_{id}^{ES}(t) - X_{id}^{ES}(t)) + \\ &\quad c_2 r_2 (gBest_{id}^{ES}(t) - X_{id}^{ES}(t)) + \\ &\quad c_3 r_3 sign(D^{ES} - distance_d) \times \\ &\quad E_i^{ES}(t) X_{max} (1 - P_i^{ES}(t)) \\ X_{id}^{ES}(t+1) &= X_{id}^{ES}(t) + V_{id}^{ES}(t+1) \end{aligned} \quad (5)$$

其中: $distance_d$ 表示 ES 粒子 i 与第 d 维最近 PS 粒子之间的距离; $sign()$ 为 0-1 阈值函数; X_{max} 表示位置的最大取值; c_3 为捕食影响因子; r_3 为 $[0, 1]$ 范围内均匀分布的随机数。

②若 $distance_d < FID$, 则粒子 i 被捕食, 即对其位置进行变异, 变异前后的粒子优胜劣汰, 但维持变异前粒子的速度 $V_i^{ES}(t)$ 和 $pBest_i$ 。

6) 判断是否满足终止条件, 若未满足, 则 $t = t + 1$, 转 2)。

2.2 基于捕食逃逸 PSO 的贝叶斯分类器

利用 PSO_PE 算法对 GBN 分类器进行学习, 首先要解决以下几个问题: 1) 确定位置与编码方式, 即如何将 GBN 网络结构编码到粒子的位置; 2) 设定初始群体; 3) 构建适应度函数来度量粒子的位置所代表的贝叶斯网对训练数据的“匹配”程度; 4) 确定粒子的运动模式; 5) 选取各种控制参数。

2.2.1 编码方法以及网络结构的限制

GBN 将分类节点当成和其他节点一样的属性节点来对待, 这种结构学习需要获得一个完整的贝叶斯网络。因此可认为每一种粒子位置状态对应一个网络结构, 状态空间是所有可能结构的集合。对于网络结构的编码, 由贝叶斯网的定义知道网络结构是有向无环图, 所以可将网络中每一个变量的编码用其父节点集合来表示。整个网络的编码就是按一定顺序排列的节点的父节点集合。图 1 为一个简单的贝叶斯网络, 其中 A_0 是类节点 (在图中未将它作为特殊阶段), A_1 至 A_4 是属性节点。按照上面的编码规则, 图 1 的编码为: $[A_0: A_{\pi(0)}; A_1: A_{\pi(1)}; A_2: A_{\pi(2)}; A_3: A_{\pi(3)}; A_4: A_{\pi(4)}]$, A_i 的父节点用 $A_{\pi(i)}$ 表示, 其中 $A_{\pi(0)} = [A_4]; A_{\pi(1)} = [A_0, A_2]; A_{\pi(2)} = [A_0]; A_{\pi(3)} = [A_2, A_4]; A_{\pi(4)} = [\emptyset]$; 可以将每个节点的局部结构 (该节点和父节点集) 表示成位置的一个分量的形式。

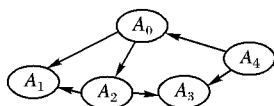


图1 一个简单的 GBN 分类器结构

n 个节点能构成有向无环图的贝叶斯网络数目可由式(6)得到:

$$f(n) = \sum_{i=1}^n (-1)^{i+1} C_i^n 2^{i(n-i)} f(n-i) \quad (6)$$

由式(6)可知, 当节点个数增长时, 需要搜索的贝叶斯网络模型的个数将呈指数级增长。为了尽量地减少搜索空间,

最有效地找到最优解, 可采取以下解决方案:

1) 限制每个节点的父节点数目, 为所有的节点设置一个最大父节点个数 m ($m < N-1$) 来减少网络结构的复杂性, 在本文中设置每个节点的父节点数不能超过 3;

2) 通过计算节点间的相关性, 根据阈值确定没有依赖关系的节点不能成为父子节点:

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \lg \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \quad (7)$$

根据式(7)计算出节点间相关性, 当相关性小于设定的阈值 δ 时, 就认定节点 X_i 和 X_j 之间没有依赖关系, 从而不能成为父子节点关系。通过相关性的计算, 又可以进一步减少搜索空间。

2.2.2 初始群体的设定

PSO_PE 算法中有两种类型的粒子群, 被捕食粒子群和捕食粒子群, 其中被捕食粒子群用来获得 GBN 分类器结构的最优解。被捕食粒子群可以随机生成, 或专家先验知识给出, 或由领域专家给出想象数据集; 然后通过一个程序帮助用户创建一个假想的完备数据集, 在完备数据集条件下选择出一定数量的较好的网络结构作为初始群体。本文随机生成被捕食粒子群和捕食粒子群, 捕食粒子群是为了增加被捕食粒子群的局部搜索能力, 所以捕食粒子群个数不能太多, 取值为 1~3 较为恰当。对于所有的初始粒子群, 都要进行有向无环形、父节点个数限制和计算出节点间相关性等处理。

2.2.3 适应度函数

适应度通常用来度量群体中各个个体在优化计算中有可能达到或接近于找到最优解的优良程度。适应度函数是用来评估个体的适应度, 即区分群体中个体好坏的标准。本文是将 PSO_PE 算法封装在 Weka 平台上, 用 Weka 自带的适应度函数 (封装在 `weka.classifiers.bayes.net.search.global.GlobalScoreSearchAlgorithm` 类 `calcScore` 方法里), 更易实现 PSO_PE 算法。

2.2.4 速度算子的确定

1) 速度^[6]。表示一个集合的形式, 该集合中每一个元素为一个节点及其父节点集。 v 表示速度, $\|v\|$ 表示该速度所包含的元素个数, N 为网络中节点的个数, 则该速度可表示为:

$$\begin{aligned} v &= \{(A_{i1}: \pi_{A_{i1}}), (A_{i2}: \pi_{A_{i2}}) \cdots (A_{i\|v\|}: \pi_{A_{i\|v\|}})\} \\ ij &\in \{1, 2, \dots, N\}, j = 1, 2, \dots, \|v\| \end{aligned} \quad (8)$$

其中: 对于任意 j_m, j_n , 如果 $j_m \neq j_n$, 那么 $ij_m \neq ij_n$; 空速度为一个空表, 表示不做任何替换操作。

2) 速度的加法。设 v_1 和 v_2 为两速度, 速度的加法 $v_1 + v_2$ 为两个速度的并集。

3) 速度的乘法。当 $c \geq 1$ 时, $c * v = v$; 当 $c < 1$ 时, 在 v 中随机取 $\lceil c * \|v\| \rceil$ 个节点及其父节点集构成 $c * v$, 其中 v 表示速度, c 为大于 0 的实数, $c * v$ 为速度的乘法。

4) 速度与位置的加法。 $P' = P \oplus v$, 即用速度 v 中各节点的父节点集替换原位置 P 所表示的网络中相应节点的父节点集。得到新位置 P' 后, 它所表示的网络结构中很可能出现环路的情况, 这时需要打破这些环路, 使网络结构合法化。

5) 位置与位置的减法。设 P_1, P_2 分别代表 2 个粒子的位

置,则 $P_1 \ominus P_2 = v$, 速度 v 是由 P_1 中那些节点的局部适应度高于 P_2 中相应节点局部适应度的节点及其父节点集构成的集合。

2.2.5 其他参数的设定

1) 全局极值。只有一个全局极值即被捕食粒子全局极值 $gBest_{id}^{ES}$, 被捕食粒子全局极值记录了目前为止最佳的合法网络结构, 当算法结束时, 该值就是所得到的最优网络结构。

2) 个体极值。捕食粒子个体极值为 $pBest_{id}^{ES}$, 被捕食粒子个体极值为 $pBest_{id}^{ES}$, 在粒子运动过程中, 通过个体极值记载每个粒子的历史最佳位置, 即粒子形成的最佳贝叶斯网络结构。

3) 被捕食粒子位置突变。当捕食粒子和被捕食粒子 i 间的距离 $distance_d < FID$ 时, 则粒子 i 被捕食, 即对其位置进行突变。被捕食粒子这种位置的突变, 本文采取随机产生一个新的被捕食粒子, 即随机产生一个符合网络结构限定的贝叶斯网络。

4) FID 捕食距离。 FID 捕食距离决定被捕食粒子在什么范围内将被捕食, 本文中 FID 设定为捕食粒子和被捕食粒子贝叶斯网络结构的相似度。本文设定当相似度大于 90% 时被捕食。

5) 粒子群的个数和迭代的次数。被捕食粒子用来获得最优贝叶斯网络结构。根据本文整理出来的高职院校毕业生数据, 学习 GBN 分类器网络结构时, 被捕食粒子的个数设定为 40, 捕食粒子的个数为 1。迭代的次数为 80, 停止条件为计算到最大的迭代次数为止。

3 实验结果

3.1 PSO_PE 算法在 Weka 平台上的实现

Weka^[7] 作为一个公开的数据挖掘工作平台, 集合了大量能承担数据挖掘任务的机器学习算法, 包括对数据进行预处理、分类、回归、聚类、关联规则以及在新的交互式界面上的可视化。Weka 提供了贝叶斯网络分类器搜索算法的抽象父类供实验者继承, 在这个抽象类的基础上, 很容易能实现自己的算法。该类在包 `weka.classifiers.bayes.net.search` 下, 类名为 `SearchAlgorithm`, 该类是一个抽象的类。该类有一个抽象子类在包 `weka.classifiers.bayes.net.search.global` 下, 类名为 `GlobalScoreSearchAlgorithm`, PSO_PE 搜索算法就是继承了这个 `GlobalScoreSearchAlgorithm` 抽象子类。 PSO_PE 算法能够设置相应的参数, 设置参数的方法是固定的, Weka 平台已经做好了相应的接口, 继承并实现该接口就行, 接口的名称为 `OptionHandler`, 该接口有 `setoptions`、`getoptions`、`listoptions` 3 个方法, 其作用分别是设置分类器参数、获取当前分类器设置的参数以及给出当前分类器所有能设置的参数列表。下面列出关键的成员变量和成员方法。 PSO_PE 算法与遗传算法类似, 是一种迭代搜索算法, PSO_PE 算法的终止条件是满足最大迭代次数。

3.2 贝叶斯网络分类器在高职院校就业预测上的应用

为了进一步验证基于捕食逃逸 PSO 的贝叶斯网络分类器的分类预测效果, 本文将其具体应用于南海东软信息技术学院就业预测分析, 选用了预处理过的该院毕业生数据作为实验数据集, 特征变量和类变量的值都是离散的且没有缺失数据项。在本文中类别变量为 $C = \{c_1, c_2\}$, 特征变量为 $Y =$

{专业名称, 实训成绩, 专业必修成绩, 公共必修成绩, 专业选修成绩, 公共选修成绩, 是否参加 SOVE}。采用的评价方法是 10-fold 交叉验证法^[9], 采用分类正确率^[10] (Classification Accuracy, CA) 为分类预测准确度指标, 各种分类器分类正确率的实验结果可以参照表 1 的数据进行比较。

从表 1 中可以看出, 基于 PSO_PE 的 GBN 分类器相对于决策树、NB、TAN、BAN 均有较高的分类正确率, 体现出较好的分类能力。其主要原因在于 PSO_PE 通过在群体中引入捕食粒子来增大逃逸粒子的捕食风险, 使得各逃逸粒子根据捕食风险和自身能量状态的权衡结果产生相应逃逸行为, 提高了粒子群对称调整社会认知能力, 进而有效地保持群体多样性, 平衡群体的探索和开发能力, 使群体避免陷入早熟收敛, 具有较好的全局收敛效果, 因而基于 PSO_PE 的 GBN 分类器能够去除不相关和多余的属性变量, 进而实现对贝叶斯网络结构的精确构造和搜索, 既提高了抗噪声能力, 又避免了对数据的过度拟合, 因此具有良好的分类效果。

表 1 分类器实验结果对照表

分类器	搜索算法	评估标准	CA/%
决策树	ID3		89.7456
	C3.4		92.6868
NB	NaiveBayes		90.8585
TAN	K2	Bayes	92.2893
	K2	MDL	92.6073
	TAN	Bayes	93.0843
	TAN	MDL	93.0048
BAN	K2	Bayes	92.0509
GBN	GeneticSearch	Bayes	90.8585
	PSO_PE	Bayes	94.1637

表 1 中“MDL”为最小描述长度 (Minimum Description Length)。

4 结语

所提出的捕食逃逸 PSO 算法具有较好的全局搜索能力, 将其用于对贝叶斯网络结构的学习, 能构造出与数据集精确匹配的网络结构。通过在高职业院校就业预测分析上应用的实验结果表明, 基于捕食逃逸 PSO 的贝叶斯分类器能够获得比决策树、NB、TAN、BAN 更好的分类结果, 方法有效可行。

参考文献:

- [1] HAN JIAWEI, KAMBER M. Data mining: concepts and techniques [EB/OL]. [2010-05-10]. <http://www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf>.
- [2] JIE CHENG, GREINER R. Comparing Bayesian network classifiers [C]// Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1999: 101-108.
- [3] EBERHART R, KENNEDY J. A new optimizer using particle swarm theory [C]// Proceedings of the Sixth International Symposium on Micro Machine and Human Science. Washington, DC: IEEE Computer Society, 1995: 39-43.
- [4] HAI ZHUGE. The future interconnection environment [J]. IEEE Computer, 2005, 38(4): 27-33.
- [5] JOHN A. Animal behavior [M]. Sunderland: Sinauer Associates, 1993: 321-395.
- [6] 刘欣, 贾海洋, 刘大有. 基于粒子群优化算法的 Bayesian 网络结构学习 [J]. 小型微型计算机系统, 2008, 29(8): 1516-1519.