

基于群体智能的选择性决策树分类器集成

王丽丽, 苏德富

(广西大学 计算机与电子信息学院, 广西 南宁 530004)

摘要: 尽管选择性集成方法的研究和应用已取得了不少重要成果, 然而其实现方法计算复杂度高、效率低仍是应用该方法的一个瓶颈。为此, 提出了一种新的高速收敛的选择性集成方法。该方法使用 C4.5 决策树分类器作为基学习器, 利用高速收敛的群体智能算法来寻找最优集成模型, 并在 UCI 数据库的多值分类数据集上进行了实验。实验结果表明, 该方法计算效率高, 其精度和稳定性比 Bagging 方法都要高, 可以成为一种高效的选择性集成的实现方法。

关键词: 选择性集成; 群体智能; 蚁群优化算法; Bagging

中图分类号: TP181

文献标识码: A

文章编号: 1673-629X(2006)12-0055-03

Swarm Intelligence - Based Selective Ensemble with Decision Trees Classifiers

WANG Li-li, SU De-fu

(School of Computer and Information Engineering, Guangxi University, Nanning 530004, China)

Abstract: Although a good many important results have been achieved about the research of selective ensemble approach and its application, it remains a computational bottleneck that the implementation of selective ensemble approach costs too much time to find an optimal ensemble. Therefore, a quickly convergent version of selective ensemble algorithm is presented. This algorithm uses convergent SI (swarm intelligence) to find the optimal ensemble with using the C4.5 decision trees classifiers as based learners. Meanwhile, experiments are carried out on UCI data sets. The computer experiments demonstrate that the proposed algorithm achieves high speed, and its accuracy and stability are both higher than Bagging algorithm. It can become a high efficient selective ensemble algorithm.

Key words: selective ensemble; swarm intelligence; ant colony optimization; Bagging

0 引言

集成学习 (Ensemble Learning) 技术是一种新的机器学习范式, 它使用多个 (通常是同质的) 学习器来解决同一个问题, 它能够显著地提高学习系统的泛化能力^[1]。国际上, 很多研究者都投入到集成学习的研究中, 使得该领域迅速越升为当前机器学习领域的四大研究方向之首^[1]。目前已有集成学习算法, 其中以 Bagging^[2] 和 Boosting^[3] 方法影响最大。

近年来, Zhou 等人^[4] 提出了“选择性集成 (selective ensemble)”的概念, 在国内外引起了很大反响, 理论分析和实验结果均表明该方法优于 Bagging 和 Boosting 方法。但由于其实现方法是建立在遗传算法的基础上, 计算复杂度, 效率低。

为解决这一难题, 文中提出了基于群体智能的选择性集成实现方法, 并将这种方法命名为 ACOSSEN (ACO algorithm based Selective ENsemble)。该方法利用群体智能

的典型算法——蚁群优化算法来选择差异大的个体建立最优的集成模型, 不仅使计算效率大幅度提高, 而且得到的集成模型规模小且精度高。实验结果表明, ACOSSEN 计算效率高, 其预测精度和稳定性均比 Bagging 方法高, 可以成为一种高效的选择性集成的实现方法。

1 选择性集成

选择性集成是一种新的集成学习范式, 由 Zhou 等人^[4] 提出, 在国际上引起了很大反响。他们证明了通过选择部分个体学习器来构建集成要优于使用所有个体学习器构建的集成, 这就意味着利用中小规模的选择性集成就可以获得很好的性能。理论分析和实验结果均表明该方法优于 Bagging 和 Boosting 方法。同时他们的“Many Could Be Better Than All”^[4] 的观点也被广为接受, 选择性集成已经成为目前集成学习领域中效果最好的学习范式。

由于选择性集成所面对的学习器可视为对同一问题的不同解决方案, 因此选择性集成其实是一种具有一定普遍性的思想。选择性集成的基本思想^[4] 就是利用多个个体 (解决方案), 并通过对个体进行适当的选择, 将所选择

收稿日期: 2006-03-17

作者简介: 王丽丽 (1980-), 女, 广西柳州人, 硕士研究生, 研究方向为并行计算与网络安全; 苏德富, 硕士生导师, 博士, 研究方向为并行计算与网络安全。

的结果进行结合从而得到更好的解。

Zhou 等人根据这个思想,利用了遗传算法来选择个体差异大的个体建立集成模型,提出了 GASEN (Genetic Algorithm based Selective ENsemble)^[4] 的选择性集成的实现方法。但是由于该方法是建立在遗传算法的基础上,而遗传算法计算效率低,使得 GASEN 计算量很大,所以该算法计算复杂度高,效率低。鉴于此,文中利用群体智能算法计算效率更高、收敛速度更快的特点,提出了一种新的基于群体智能的选择性集成的实现方法。

2 群体智能与蚁群优化算法

2.1 群体智能

群体智能的概念源于对蜜蜂、蚂蚁、大雁等这类群居生物群体行为的观察和研究,是一种在自然界生物群体所表现出的智能现象启发下提出的人工智能实现模式,是对简单生物群体的智能涌现现象的具体模式研究,即“简单智能的主体通过合作表现出复杂智能行为的特性”^[5-8]。该智能模式需要以相当数目的智能个体来实现对某类问题的求解功能。群体智能的典型实现模式——蚁群优化算法正在受到学术界的广泛关注。由于其概念简明、实现方便,在短期内迅速得到了国际演化计算研究领域的认可,并在组合优化、电力系统、人工智能、冶金自动化等领域得到了广泛应用。

2.2 蚁群优化算法

蚁群优化算法(Ant Colony Optimization, ACO)^[9]是近年来刚刚诞生的群体智能方法,它是一种源于大自然的新的仿生类算法,是由意大利学者 M. Dorigo, V. Maniezzo, A. Colomni 等人^[10-12]根据蚂蚁群体具有智能的特点而首先提出来的,它主要是通过蚂蚁觅食过程中群体之间的信息传递而达到寻优的目的。其原理是一种正反馈机制,通过信息素的不断更新最终达到收敛于最优路径上。具有群体智能的蚁群优化算法还具有如下优点:

①算法中的人工蚁群是一种正反馈机制或称增强型学习系统,它通过信息素的不断更新达到最终收敛于最优路径上;

②它是一种通用型随机优化方法,但人工蚂蚁决不是对实际蚂蚁的一种简单模拟,它融进了人类的智能;

③它是一种分布式的优化方法,不仅适合目前的串行计算机,而且适合未来的并行计算机;

④它是一种全局优化的方法,不仅可用于求解单目标优化问题,而且可用于求解多目标优化问题;

⑤它是一种启发式算法,计算复杂性为 $O(NC \cdot m \cdot n^2)$,其中 NC 是迭代次数, m 是蚂蚁数目, n 是目的节点数目。蚁群优化算法所具有的天然的随机性、自适应性和分布式等特点,使其非常适合并行计算和求精确解。同时有研究表明^[13],该算法在没有任何先验知识的情况下比遗传算法和模拟退火算法的执行效率都要高,系统收敛的速度也要快。

3 基于群体智能的选择性集成(ACOSSEN)

选择性集成方法利用了遗传算法来选择个体差异大的个体建立集成模型,正是由于其实现方法是建立在遗传算法基础上,所以该算法不可避免地具有遗传算法计算复杂度高、效率低的缺点。鉴于此,文中提出基于群体智能的选择性集成实现方法,命名为 ACOSSEN(ACO algorithm based Selective ENsemble)。

ACOSSEN 是利用群体智能中的典型算法“蚁群优化算法 ACO”来选择个体差异大的个体建立集成模型的一种方法。ACOSSEN 首先在训练集上利用 bootstrap 方法^[14]训练多个学习器,然后用 ACO 对这些学习器进行选择,选出其中差异最大的个体组成选择后的个体集合。选择好个体后,通过多数投票法,建立集成模型。在用 ACO 选择学习器时,将每个学习器当作一个二进制,若该学习器被选择的话,则用 1 代表,否则用 0 代替,所以优化后的学习器组合可用一串二进制 0101...来代替,在找到最优的那组学习器后,在该组以 1 为代表的学习器作为集成模型输出。

ACOSSEN 算法描述如下:

(1) 从训练集 S 中通过 bootstrap 方法产生子集 S_i ,用 S_i 训练后得到学习器 L_i ,训练 T 轮后得到一组基学习器 L 。

(2) 采用蚁群优化算法 ACO 对学习器 L 进行选择:

①初始化相关参数:蚂蚁的数目,迭代的次数,初始信息素浓度,先验概率等;

②每只蚂蚁都同时寻找和选择学习器个体;

③每只蚂蚁 k 都通过遍历各个学习器而形成一个解,并在遍历的过程中,将遍历到的学习器的值(0/1)保留在 $Solution_k$ 中;

④蚂蚁 k 对学习器 i 是否进行选择,要根据情况而定。首先系统产生一个随机数,如果它大于或等于先验概率 Q_0 ,则根据下列的概率公式(1)判断是否选择当前学习器 i :如果 P_1 大于 P_0 ,则选择当前学习器 i ;否则不选择。

$$P_0 = [\tau_{i,0}]^\alpha \cdot [d_i + 1]^\beta \quad P_1 = [\tau_{i,1}]^\alpha \cdot [d_i]^\beta \quad (1)$$

其中, $\tau_{i,0}$ 和 $\tau_{i,1}$ 分别表示没有被选择或被选择的学习器 i 的信息素浓度, d_i 为到目前为止蚂蚁 k 选择所有学习器的个数即 i 的个数: $d_i = \sum_{j=1}^{i=1} Solution_{k,j}$, α, β 是系统参数,分别表示信息素浓度、学习器的个数对蚂蚁选择这个学习器的影响程度。如果系统产生随机数小于先验概率 Q_0 ,则由系统产生的一个随机数 0/1 来确定是否选择学习器 i 。

⑤当所有蚂蚁选择出一组学习器后,就用一个验证集 validation set 对每个蚂蚁所选择的一组学习器进行验证,每组学习器看作一个集成,而验证后的正确识别率就是蚂蚁所求的解。如果系统采取保留最优解的策略,则判断当代蚂蚁中是否存在比到目前为止最优解还要好的解,

如果存在,则保留这个最优解,并对这只蚂蚁所选择的每个学习器上的信息素强度进行增强。

⑥ 当所有蚂蚁都求解后,则在每个学习器 i 上,更新信息素浓度。文中采用了两种更新方法:局部更新和全局更新。在蚂蚁 k 完成一次求解后,蚂蚁应用式(2)的局部更新规则对它选择的学习器 i 进行信息素更新。公式如下:

$$\tau_{i,0}^k = (1 - \rho)\tau_{i,0} + \rho\Delta\tau_{i,0}^k \quad \tau_{i,1}^k = (1 - \rho)\tau_{i,1} + \rho\Delta\tau_{i,1}^k \quad (2)$$

其中, $\Delta\tau_{i,1}^k$ 和 $\Delta\tau_{i,0}^k$ 分别表示蚂蚁 k 在本次求解过程中选择学习器 i 与不选择学习器 i 上的信息素浓度, ρ 为信息素浓度的挥发系数, $0 < \rho < 1$ 。如果学习器 i 被蚂蚁 k 选择了,则蚂蚁 k 留在学习器上的信息素强度 $\Delta\tau_{i,1}^k = R_k/Q$, $\Delta\tau_{i,0}^k$ 为 0。反之,学习器 i 没有被蚂蚁 k 选择,则蚂蚁 k 留在学习器 i 上的信息素强度 $\Delta\tau_{i,0}^k = R_k/Q$, $\Delta\tau_{i,1}^k$ 为 0。这里的 Q 为常量,表示单位信息素浓度; R_k 为蚂蚁 k 所求的正确识别率。

同时还采用全局更新规则对找到最优解的蚂蚁所选择的学习器进行信息素更新。更新公式如下:

$$\tau_{i,0} = (1 - \rho)\tau_{i,0} + \rho\Delta\tau_{i,0} \quad \tau_{i,1} = (1 - \rho)\tau_{i,1} + \rho\Delta\tau_{i,1} \quad i \in \text{全局最优解中的学习器} \quad (3)$$

⑦ 判断是否满足停止。停止条件是连续迭代 10 次取得的最优正确识别率要有所改善;否则,返回第 ② 步。

(3) 输出最优的那组学习器作为集成模型。

4 实验结果及分析

文中在通用的机器学习实验平台 WEKA 上实现了 ACOSSEN 算法,并在 UCI 的标准数据集上进行了测试。

4.1 UCI 标准数据集

根据问题的类型、样本大小等因素,选择了 UCI 数据库中的 10 个数据集来测试 ACOSSEN 算法,并在 WEKA 平台上对 UCI 数据库中的 10 个数据集进行了处理,使之满足 WEKA 的格式要求,如表 1 所示。按照每类个数,将每个数据集随机分为两部分:一部分用来作训练集,另一部分用来测试。该操作在每个数据集上重复做了 50 次。

4.2 实验结果及分析

采用目前效果较好的 C4.5 决策树算法,在 WEKA 平台上利用 bootstrap 方法训练了 20 个 C4.5 分类器作为基学习器。ACOSSEN 在每个数据集上利用训练集训练,验证集是训练集上随机抽取的,用于选择最优的集成。

实验的参数设置如下:蚁群优化算法 ACO 中的蚂蚁总数 $M = 20$, $\alpha = 1$, $\beta = 2$, $\rho = 0.1$, $Q_0 = 0.5$,单位信息素浓度 Q 为 10,初始信息素浓度为 10,迭代次数为 50 次。

实验是在一台 P4 2.4G 的计算机上进行的,数值仅供参考。实验结果如表 1 所示。

由表 1 可以看到:从正确率来看,ACOSSEN 获得了较好的精度和稳定性,其平均正确识别率都优于标准 Bag-

ging;从时间上来看,系统收敛速度较快,时间效率得到很大提高。

表 1 ACOSSEN 在 UCI 数据集上的实验结果

数据集名称	数据集			实验结果				
	示例数	属性数	分类数	单个分类器的正确识别率 (C4.5 决策树分类器)	标准 Bagging 的正确识别率	ACOSSEN 的最优正确识别率	ACOSSEN 的平均正确识别率	时间 /s
labor	57	17	2	89.4737	91.2281	96.4912	96.0234	0.291
lymph	148	19	4	91.2162	93.9189	95.9459	94.3749	0.385
waveform-5000	5000	41	3	85.72	93.02	94.9617	93.3207	12.638
splice	3190	62	3	96.3323	97.0219	98.8769	97.5862	11.852
glass	210	10	7	71.4953	81.7757	85.9813	84.6962	1.213
audiology	226	70	24	91.1054	92.0354	95.1327	94.4248	0.437
iris	150	5	3	96	96.6667	98.767	97.3333	0.28
lung-cancer	32	57	2	93.75	96.875	100	98.4375	0.325
segment	2310	20	7	95.974	97.5325	99.2614	98.9221	5.137
ionosphere	351	35	2	92.3077	95.1567	98.151	96.8661	1.013

同时从集成的规模上来说,ACOSSEN 使用了较少的个体学习器就达到了较高的预测精度,如图 1 所示。一般 ACOSSEN 找出的个体学习器数目仅是 Bagging 使用数目的 20% 左右。

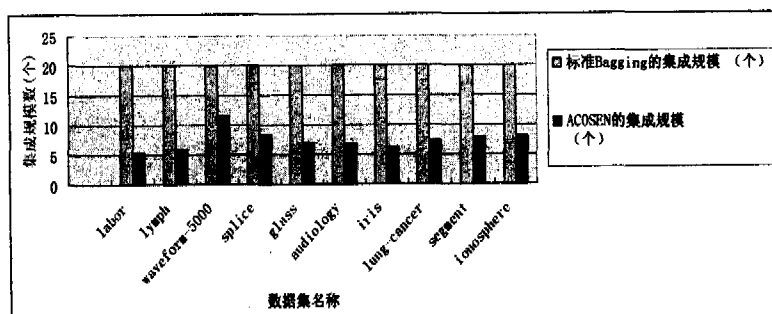


图 1 ACOSSEN 与 Bagging 的集成规模比较

5 结束语

文中采用群体智能方法进行模型选择,选择出差异大的个体建立最优的集成模型。从正确率、时间效率和集成规模三方面的实验结果均表明,该方法不仅使计算效率大幅度提高,而且得到的集成模型规模小且精度高,可以成为一种高效的选择性集成的实现方法。

然而,选择性集成还是一个比较新的研究课题,其实现方法上还存在很大的改进空间。除了聚类算法,文中采用蚁群优化算法,相信结合其他更新的更高效的算法进行模型选择将会得到更强有力的选择性集成算法,从而促进选择性集成在更多领域中的应用。

参考文献:

- [1] Dienerich T G. Machine learning research: Four current directions[J]. AI Magazine, 1997, 18(4): 97-136.
- [2] Breiman L. Bagging predictors[J]. Machine Learning, 1996 (2): 123-140.

(下转第 60 页)

M_{camera}), d 为物体的顶点到光源的距离。

2.3.3 聚光的光照方程

聚光是三种光源中计算量最大的一种光源。一方面它同样会随着物体的远近而发生衰减;另一方面,还应判断聚光源到顶点的向量在其发光锥体的哪一部分。即计算在内外锥体上的衰减因子 SpotFactor, 设变量 $Rho = -P \cdot S_{\text{Direction}}$, 则:

$$\text{SpotFactor} = \left(\frac{\max(Rho - \cos(\Phi/2), 0)}{\cos(\Theta/2) - \cos(\Phi/2)} \right)^{\text{Falloff}} \quad (6)$$

由方程(5),(6)得到的聚光的光照方程为:

$$I = M_{\text{Emissive}} + M_{\text{Ambient}} \odot A_{\text{glob}} + (S_{\text{Ambient}} + \max(N \cdot P, 0) * S_{\text{Diffuse}} \odot M_{\text{Diffuse}}) * (\text{Attenuation0} + \text{Attenuation1} * d + \text{Attenuation2} * d^2)^{-1} * \left(\frac{\max(Rho - \cos(\Phi/2), 0)}{\cos(\Theta/2) - \cos(\Phi/2)} \right)^{\text{Falloff}} \quad (7)$$

通过使用方程(7)计算的光照效果使光线在发光锥体轴线上的亮度明显增强,更加接近现实世界。

如果一个顶点受多光源的照射,那么最终的颜色值 I_{tot} 应该是多个光源光照效果的叠加。

$$I_{\text{tot}} = \sum_{i=1}^n I_i$$

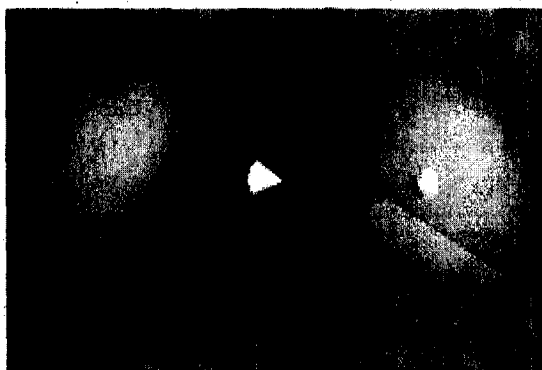


图2 shader 程序实现的聚光光照效果

3 总结与展望

通过顶点 shader 程序可以比较灵活地实现更为真实的光照效果。图2、图3对比了运用标准图形管线和 shader 程序实现的聚光的光照效果的不同。在图2中,靠近发光锥体轴线的部分,光线强度是逐渐增强的,这种效果更加贴近现实世界。但是还存在需要改进的地方,比如,墙壁较光滑时,还需要考虑它的镜面反射光。笔者将在以后的研究中做进一步的改进。

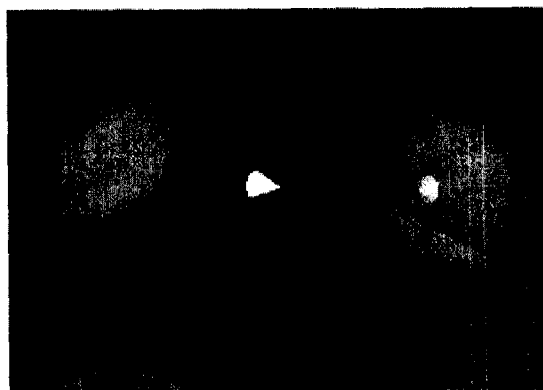


图3 标准图形管线实现的聚光光照效果

参考文献:

- [1] Fernando R. GPU 精粹——实时图形编程的技术、技巧和技艺[M]. 北京:人民邮电出版社,2006:100-106.
- [2] 陈卡. DirectX9 3D 图形程序设计[M]. 上海:上海科学技术出版社,2003:92-100.
- [3] 朱腾辉,刘学慧,吴恩华. 基于像素的光照计算技术[J]. 计算机辅助设计与图形学学报,2002,14(9):861-865.
- [4] Akenine-Moller T, Haines E. 实时计算机图形学[M]. 北京:北京大学出版社,2004:20-21.
- [5] Hanrahan P. An Introduction to Ray Tracing[M]. London: Academic Press Inc, 1989:79-120.

(上接第 57 页)

- [3] Schapire R E. The strength of weak learnability[J]. Machine Learning, 1990(2): 197-227.
- [4] Zhou ZH, Wu J, Tang W. Ensembling neural networks: Many could be better than all. Artificial Intelligence, 2002, 137(1-2): 239-263.
- [5] 吴启迪,汪 镭. 智能微粒群算法的研究及应用[M]. 南京:江苏教育出版社,2005.
- [6] Kennedy J, Eberhart R C. Swarm intelligence[M]. San Francisco: Morgan Kaufmann, 2001.
- [7] 康 琦. 微粒群优化算法的研究与应用[D]. 上海:同济大学,2005.
- [8] 吴启迪,汪 镭. 智能蚁群算法及应用[M]. 上海:上海科技教育出版社,2004.
- [9] Dorigo M, Caro G D. Ant colony optimization: a new meta-heuristic[C]//In: Proc. of the 1999 Congress on Evolutionary Computation, Vol 2. Washington: IEEE Press, 1999: 1470-1477.
- [10] Dorigo M, Caro G D, Gambardella L M. Ant algorithms for discrete optimization[J]. Artificial Life, 1999, 5(2): 137-172.
- [11] Dorigo M, Maniezzo V, Colomi A. The ant system: optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 1996, 26(1): 29-41.
- [12] Dorigo M, Gambardella L M. Ant colony system: a cooperative learning approach to the traveling salesman problem[J]. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 53-66.
- [13] 燕 忠,袁春伟. 增强型的蚁群优化算法[J]. 计算机工程与应用, 2003, 39(23): 62-64.
- [14] Efron B, Tibshirani R. An Introduction to the Bootstrap[M]. New York: Chapman & Hall, 1993.