第16卷 第3期
2006年3月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 3
Mar. 2006

基于概率的覆盖算法的研究

周 瑛[1,2],张 铃[2]

(1. 安徽大学 管理学院,安徽 合肥 230039;
2. 安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:通过对前向神经网络交叉覆盖算法的分析,针对该算法的特点,文中提出了扩大覆盖半径以减少拒识样本数据的新方法。另一方面,随着覆盖半径的扩大所出现的一个测试样本属于多个覆盖的情况,使用概率的方法对其进行处理,并对处理的结果用投票的方式来决定样本的最后类别。实验证明,该方法对提高识别率和精确度有比较好的效果。

关键词:覆盖算法;概率;精确度

中图分类号:TP18 **文献标识码**:A **文章编号**:1005-3751(2006)03-0029-02

Study of Cover Algorithm Based on Probability

ZHOU Ying[1,2], ZANG Ling[2]

(1. Management School, Anhui University, Hefei 230039, China;
2. Ministry of Education Key Lab. of IC & SP, Anhui University, Hefei 230039, China)

Abstract: The features of the crossing cover algorithm are analysed and a novel algorithm is presented. It reduces the number of the tested samples that can't be classified by the spherical neighborhood gained before by enlarging the radius of spherical neighborhood. On the other hand, the paper uses a probabilistic way to handle the case that a sample belongs to more than one spherical neighborhood, and it votes the result to classify the sample. The algorithm improves the rates of recognition and accuracy through experiments.

Key words: cover algorithm; probability; accuracy

0 引 言

前向神经网络的交叉覆盖算法是张铃教授根据M-P神经元模型的几何意义而提出的一个应用。该算法的核心思想是构造一个网络,使对给定的样本集进行符合要求的分类,等价于求出一组领域,对所给定的样本集中的点,能按分类的要求用所覆盖的领域将它们分隔开来。根据这个思想,针对学习样本的特征,首先将原空间的样本点向高维空间投影[1,2]。在投影后,每个样本点都落在一个超球面上,再根据投影后的位置来构造神经网络。这种方法可迅速地、构造性地得到对于训练数据几乎完全正确分类的神经网络,而不必象传统的BP算法那样反复地进行迭代训练而未必会有好的结果。该算法解决了双螺旋线识别的学习问题[3],还快速有效地完成了大量手写汉字的识别等海量数据的处理问题[4]。覆盖算法的一个较明显的特点是测试样本中存在一部分样本不属于任何学习所得到的球形领域,即被"拒识"。文中通过扩大覆盖半径,即扩大学习所得覆盖的球形领域,减少拒识的样本数,提高识别率;另一方面,随着覆盖半径的扩大,必然会产生覆盖间的相互交叉,即一个测试样本同时属于多个覆盖的情况,对这些样本的分类,用概率的方法对其进行处理。实验证明,这种方法减少了识别错误率,有效地提高了识别率和精确度。

收稿日期:2005-06-14
基金项目:国家自然科学基金资助项目(60475017)

作者简介:周 瑛(1968—),女,安徽无为人,副教授,博士研究生,研究方向为模糊理论及应用、神经网络、信息检索;张 铃,教授,博导,研究方向为人工智能理论、机器学习理论和方法、智能计算技术、神经网络技术。

1 覆盖算法的主要内容

覆盖算法是基于M-P神经元模型的几何意义提出的[1],该算法的主要步骤是:先将样本从原始空间(n维)向高维空间($n+1$维)投影,该变换可用T来表示,即:

$$T:D\to S^{n+1},T(x)=(x,\sqrt{r^2-|x|^2}),\text{其中 } r\geqslant \max\{|x|,x\in D\} \quad (1)$$

经过投影后的样本点都位于$n+1$维空间中某个中心在原点、半径为r的球面S^{n+1}上。而所构造的三层神经网络分类器就等价于求出一组领域,这组领域能将不同类的点分开。先求一个领域C^1,它只覆盖第一类中的点,而不覆盖其它类的点,然后将被C^1覆盖的点删去,对余下的点求第二个覆盖领域C^2,它只覆盖第二类中的点而不覆盖其它类的点,然后将被C^2覆盖的点删去……,如此反复进行覆盖,直到样本中的点被全部删除(覆盖)为止[3]。在设计网络结构时,以每个球形领域作为一个神经元,取

$\sigma(wx-\theta)$ 为其功能函数，$\sigma(x)=\begin{cases}1 & x>0\\ 0 & \text{其它}\end{cases}$，功能函数就是“球形领域”的特征函数。学习过程中构造第 K 类样本 Xk 的球形领域 C^i 的方法是：任一样本中尚未被覆盖的点 $x_i \in Xk$，按(2)式计算：

$$d1(i)=\max_{x\in Xk}|<x_i,x>|,d2(i)=\min_{x\in Xk}|<x_i,x>|<x_i,x>>d1(i)|,d(i)=(d1(i)+d2(i))/2 \tag{2}$$

其中，$<x,y>$ 表示 x、y 的内积，这样就可得到一个以 x_i 为中心，以 $\theta=d(i)$ 为阈值的覆盖 C^i，按此方法求出样本的全部覆盖。

在测试时，给定一个测试样本，若它属于某类覆盖的一个球形领域，即可将该测试样本分为该类；若它不属于任何类别覆盖的任何一个球形领域，则“拒识”。

由于覆盖算法中只覆盖同类点，而不覆盖异类点，使得在投影后的空间中覆盖的范围受到限制，虽然在学习时对训练样本几乎可以完全识别，但在测试时，产生的“拒识”样本数较多。为了提高识别率，减小误差，通过增大覆盖的阈值，扩大领域的半径，提高识别率；通过使用概率的方法，提高精确度，减小误差率。

2 基于概率的覆盖算法

2.1 基于概率的覆盖算法的训练步骤

设学习样本 X 分为 m 类，即 $X=\{X_1,X_2,\cdots,X_m\}$，求出样本 X 的所有覆盖的算法为：

①求出样本 X 中的最大模 r，并将 X 中的点投影到中心在原点、半径为 $2r$ 的球面上(为避免测试集中可能出现模更大的样本，取半径为 $2r$ 比取 r 合适)；

②初始化覆盖个数 $j=1$，类别数 $i=1$；

③取第 i 类的样本，构造第 j 个覆盖 $C(j)$；

④若 X_i 中没有尚未覆盖的点，转⑧；否则，任取 X_i 中尚未被覆盖的一点 x_i；

⑤按公式(2)计算，作以 x_i 为中心、θ 为阈值的覆盖 $C(j)$；

⑥若 $C(j)$ 覆盖的同类点的个数大于 2，则增加若干个异类点，扩大覆盖半径，同时保存该覆盖所包含的同类点及异类点个数；

⑦$j=j+1$，转③；

⑧训练结束。

在步骤⑥中，增加异类点的个数要根据具体问题来确定，根据经验，增加 2～4 个异类点，效果较好。

2.2 基于概率的覆盖算法的测试步骤

对所测试的样本，一般可能出现 3 种情况：①只属于某一类领域的某个覆盖；②属于多类领域的多个覆盖；③不属于任何覆盖。对于第①种情形，直接由覆盖中心确定其类别，对于第③种情形，定为“拒识”。由于在基于概率的覆盖算法的训练算法中，阈值 θ 比原来的覆盖算法要大，所以第②种情形出现的几率比原覆盖算法要大。对于②，又有两种情形，如图 1 所示。

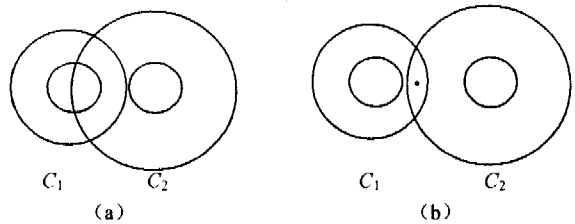

图 1 测试样本属于多个覆盖示意图

在图 1 中，阴影部分表示该覆盖中的同类点。图 1(a)中，测试样本位于 C_1 的同类点与 C_2 的异类点的交叉处，这时确定该点属于 C_1 所在的类；图 1(b)中，测试样本位于 C_1 的异类点与 C_2 的异类点的交叉处，这时考虑该点属于 C_1、C_2 类的概率 p。令 p_i 为 $C_i(i=1,2)$ 中同类点个数占该覆盖中总样本个数的比例，即 $p_i=$ 同类点个数 / 该覆盖的总样本数，则 $p=p_1/(p_1+p_2)$，即为该点属于 C_1 的概率，$0<p<1$。在程序的实际运行时，产生 100 个(0，1)上的随机数，若该数小于等于 p，即认为该点属于 C_1，否则属于 C_2。然后对此结果进行投票，以决定该样本点的分类，具体步骤如下：

①将待测试的样本点投影到中心在原点、半径为 $2r$ 的球面上；

②对每个样本 x，计算 $d(x,C_i)=<x,x_i>,i=1,2,\cdots,j$，其中 x_i 为覆盖 C_i 的中心，j 为覆盖总数；

③若 x 只属于一个覆盖，判 x 属于 x_i 所在的类，转⑥；

④若 x 不属于任一覆盖，判 x 为“拒识”，转⑥；

⑤若 x 属于多个覆盖，且 x 位于某一覆盖的同类点所在的半径内，则判为该类；否则，x 位于两个覆盖的异类点所处位置的交叉处，则取密度为 $p=p_1/(p_1+p_2)$，产生 100 个(0,1)间的随机数，统计大于 p 和小于 p 的个数，再进行投票，最后决定 x 的类别；

⑥统计识别的误差率。

3 实验结果

根据基于概率的交叉覆盖算法的思想，使用 UCI 数据库中的 pima－indians－diabetes 数据[5]，该数据库中的数据包括 8 个属性，结果分两类，共有 768 个样本。用前 576 个样本作为训练样本，后 192 个样本作为测试样本。作为对比，同时也用原覆盖算法对该数据进行了学习与测试，实验结果如表 1 所示。

表 1 pima 数据库实验结果表

项目	CA	PCA
训练样本数	576	576
测试样本数	192	192
覆盖个数	255	221
训练时间(s)	0.313	0.578
拒识样本数	65	36
正确样本数	98	121
拒识率	0.338	0.187

(下转第 33 页)

$A = 1, B = 10, C = 15$,从这个网络参数的实际出发,这样的参数设置突出了对延时的敏感性,即在满足 QoS 要求的前提下,延时小的路径将得到优先选择。如何选取网络各边的初始信息素值,目前对此还没有理论指导,只能通过实验确定较好的初始值。在文中认为各边具有相同的信息素初始值,通过大量实验后认为取信息素初始值=50.0 可取得较好的结果。同时,将每次进行全局更新时的信息素的增量设为 10。

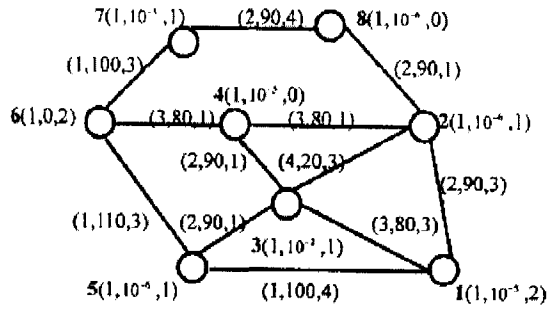

图 1　网络拓扑与参数

首先进行了(1,6)两点之间的最优路由选择,通过对结果的分析,发现该算法可找出(1,6)之间的几乎所有可能路由,充分说明了该算法对路径搜索的充分性,可以防止陷入局部最优。最后找到的最优路径如表 1 所示。

表 1　实验结果 1

最佳路由	费用	延时	丢包率
1 2 4 6	8	8	0.000011

可以看到,对于路由请求(1,6),还存在着路由 1,5,6,其总费用仅为 2,但延时达到了 9,而文中的算法是在满足 QoS 的条件下延时敏感的,虽然 1,2,4,6 这条路径费用达到了 8,但延时只有 8,所以被选择。同时看到存在路由 1,3,4,6,其费用和延时和 1,2,4,6 完全一样,但丢包率比较大,达到 0.010010,不满足预设的 QoS 要求,所以不被选择。由此可以看出,本算法具有良好的性能,同时也具有高度的灵活性,只需对 A, B, C 等参数适当取值,就可以做到在满足 QoS 的前提下对某一参数特别考虑。在算法进行的过程中,可以看到免疫算法对结果的影响,初始生成的染色体具有很大的随机性,最优路径命中率分布不均匀,高的可达 90% 以上,低的只有 3%,平均不超过 50%,性能一般。经过几代进化选择后,染色体趋于几个最佳值,命中率也大大提高,达到了 95% 以上,说明了免疫算法对蚂蚁算法参数进行选择的有效性。同样,对于路由请求(3,8),最后也能找到最佳路由(如表 2 所示)。

表 2　实验结果 2

最佳路由	费用	延时	丢包率
3 4 2 8	7	6	0.000012

由结果可知,本算法能很好地找到网络最优解,同时由于免疫算法的融合,本算法有较好的收敛性,有效性大大提高。

5　结束语

文中从用免疫算法对蚁群算法的参数进行控制的角度,探讨了对蚁群算法的改进方法,并通过对 QoS 单播路由问题的求解验证其有效性。

参考文献:

[1] 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.

[2] 张素兵,吕国英,刘泽民,等.基于蚂蚁算法的 QoS 路由调度方法[J].电路与系统学报,2000,3(1):1－5.

[3] Dorigo M, Maniezzo V, Colony A. The Ant System: Optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybermetrics, 1996, 26(1): 1－13.

[4] 葛　红,毛宗源.免疫算法的改进[J].计算机工程与应用,2002,14(7):47－50.

[5] 邓志成,周　棋,张凌云,等.QoS 单播路由算法的研究[J].通信学报,2001,22(8):122－128.

[6] 周　正,刘泽民.智能蚂蚁算法及其在电信网动态路由优化中的应用[J].电信科学,1998,11(11):10－13.

[7] 王征应,石冰心.基于启发式遗传算法的 QoS 组播路由问题求解[J].计算机学报,2001,24(1):55－61.

(上接第 30 页)

表中 CA 表示覆盖算法,PCA 表示基于概率的交叉覆盖算法,对 PCA 的实验共做 10 次,最后取平均值。

实验的结果表明,基于概率的覆盖算法减少了覆盖个数和拒识样本数,提高了识别率和精确度,达到了预期的效果。但如何选择合适的异类点个数以及如何选择合适的概率 p 来提高 PCA 的识别精度,仍是今后工作中进一步的研究目标之一。

参考文献:

[1] 张　铃. A Geometrical Representation of McCulloch－Pitts Neural Model and Its Applications[J]. IEEE Trans on Neural Networks, 1999, 10(4): 925－929.

[2] 张　铃,张　钹. M－P 神经元模型的几何意义及其应用[J].软件学报,1998,9(5):334－338.

[3] 张　铃,张　钹,殷海风.多层前向网络的交叉覆盖算法[J].软件学报,1999,10(7):737－742.

[4] 吴鸣锐.大规模模式识别问题的分类器设计研究[D].北京:清华大学计算机系,2000.

[5] Blake C, Merz C. UCI repository of machine learning databases, URL[EB/OL]. http://www.ics.uci.edu/(mlearn/ML-Repository.html. 1998.