一种基于竞争的覆盖算法

张月琴,孙先洋,刘 翔
(太原理工大学 计算机科学与技术学院,山西 太原 030024)

摘 要:与传统人工神经网络的算法相比,覆盖算法有运行速度快、精度高和易于理解的优点,但是覆盖算法的学习顺序是随机选择的,大量实验表明样本的学习顺序对神经网络的性能有着显著的影响。基于竞争的覆盖算法是在覆盖算法的基础上提出的,以消除算法中学习顺序所产生的影响。在该算法中,通过加入竞争机制,神经网络在学习样本的同时会逐步调整覆盖中心以形成更优的覆盖域。实验表明改进后的覆盖算法可以有效减少覆盖数量,减少拒识样本数,提高识别精度。

关键词:神经网络;覆盖算法;学习顺序
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2012)09-0029-03

A Covering Algorithm Based on Competition

ZHANG Yue-qin, SUN Xian-yang, LIU Xiang
(College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan 030024, China)

Abstract: Compared with traditional neural networks, covering algorithm possesses some advantages, such as running fast, high accuracy and easy to understand, but the learning order of covering algorithm is randomly selected. Experiments show that the learning sequence has a significant impact on the network performance. It proposes a new kind of algorithm named covering algorithm based on competition. In this algorithm, sphere domains can be adjusted gradually. Experiments show that this algorithm can effectively reduce the number of sphere domains, decrease the number of rejected samples and improve the recognition accuracy.

Key words: neural network; coverage algorithm; learning sequence

0 引 言

人工神经网络是由大量并行分布、有机相连的神经元组成的计算结构,是在对人脑组织结构和运行机制的认识理解基础之上模拟其结构和智能行为的一种工程系统,并不是人脑功能的真实描述[1]。神经网络的信息处理是由神经元之间的相互作用来实现的;知识与信息的存贮表示成为网络元件之间互联分布式的物理联系;网络的学习和识别决定于各神经元连接权系数的动态演化过程[2]。

传统的神经网络是非构造性的,相邻层之间是全连接的,即每个神经元都接收前一层中全部神经元的输出,因此网络各部分间是紧密联系的,任何部分的微小改动都会对网络其他部分造成影响,这使得网络很难到达理想的状态,而且存在计算量大、训练时间长和泛化能力不足等缺点[3]。张玲教授等在文献[4,5]中根据神经元的几何意义提出多层前向神经网络的覆盖算法。在超平面表示神经元的几何意义的基础上,总结并提出了球形领域表示神经元的几何意义获得成功,并给出了构建神经网络的覆盖算法,在一定意义上解决了作为分类器的多层前向网络的设计问题。与传统的前向神经网络相比,覆盖算法具有运行速度快、精度高、易于理解和维护的特点。但是覆盖算法在构建神经网络的学习训练中,是随机选取覆盖中心的。实验均表明学习顺序会对神经网络的质量产生显著影响。文中在前向神经网络覆盖算法的基础上提出了一个新的覆盖算法——基于竞争的覆盖算法,该方法引入竞争机制在一定程度上消除了学习顺序对最终神经网络的影响。实验表明这种改进的算法,通过对神经元进行优胜劣汰的筛选,可有效提高神经网络的整体质量。

1 覆盖算法

在构造性神经网络中,使用的训练学习方法即覆盖算法[4-6],它是根据样本数据本身的结构构建神经

收稿日期:2012-01-13;**修回日期**:2012-04-20
基金项目:山西省自然科学基金项目(2008011028-1);山西省科技攻关项目(20100322003)
作者简介:张月琴(1963-),女,教授,硕士生导师,研究方向为智能信息系统、数据挖掘等;孙先洋(1986-),男,硕士研究生,研究方向为数据挖掘、人工神经网络。

网络。在使用覆盖算法构造性神经网络的过程中,首先对给定的样本集进行符合要求的分类,此过程等价于求出一组邻域,然后对给定样本集中不同类别的样本点用邻域覆盖将它们分割开来。这样就把神经网络的设计问题转变成求邻域覆盖的问题,同时也就将原先基于搜索机制的学习方法转变成构造性学习方法也就是覆盖算法[4]。覆盖算法根据学习样本的特征来构造神经网络,在对给定数据进行处理的过程中,给出网络的结构和参数,也就是说所得到的网络结构是在处理数据的过程中逐步构造的,而不是在学习之前就已经给定的。这样整个神经网络中的每个神经元及其功能均在观察之中,因此使得对获得的神经网络进行深入分析和调整提供了可能性,所以构造性神经网络一经提出就获得了很多关注,并成功在一些领域进行了实际应用[7]。

下面给出一个简单的构造性神经网络学习过程:

给定一个输入集(样本集) $Y = \{(X^1, y^1), (X^2, y^2), \cdots, (X^l, y^l)\}$ (X 为 n 维向量,$l = 0,1,2\cdots$) 作为训练样本。设整个样本集可以分为 S 个类别,使用 $I(t)$ 表示输出 y 相同的 i 的集合 $I(t) = \{i \mid y^i = y^t\}$,相对应的样本点可根据 $p(t) = \{x^i \mid i \in I(t)\}$,$t = 0,1,2,\cdots, S-1$ 进行分类。对于所有的 $x^i \in p(k)$ 的样本点均被覆盖域集合 $\{C_m^k, m = 1,2\cdots\}$ 中的某个覆盖域覆盖,且只能被属于该覆盖域集合中的覆盖域覆盖,所有的覆盖域集合则构成了整个覆盖神经网络 $C(k) = \cup C_m^k$。

具体的算法步骤如下:

步骤 1:在样本集 Y 中随机找一个尚未覆盖的点 x^m,按公式

$$d1(j) = \min_{x^m \notin Y_k} \{d(x^j, x^m)\}$$

$$d2(j) = \max_{x^m \in Y_k} \{d(x^j, x^m) < d1(j)\}$$

$$d(j) = [d1(j) + d2(j)]/2$$

$$\theta(j) = [d1(j) - d2(j)]/2$$

$d1(j)$ 为 x^m 与 $x^j \in p(t)$ 之间的最短距离。

$d2(j)$ 为 x^m 与 $x^j \in p(t)$ 之间最大且小于 $d1(j)$ 的距离。

计算得到以 x^m 为中心,半径 $R = d(j)$,分类间隙为 $\theta(j)$ 的覆盖 $C_m^k = \{x \mid (x, x^m) < d(j)\}$。

步骤 2:C_m^k 求出之后,将所有被 C_m^k 覆盖的点从样本空间中删除,再在 Y_k 中选择一个 x^m,重复上述操作直到所有属于 Y_k 的点均被删除为止。这样,一个类的所有覆盖领域就构造出来了。

步骤 3:针对每种类别的样本点分别按类别按上述 2 个步骤进行处理,最终就可以得到一个覆盖集合 $C(k) = \bigcup_{m \in I(k)} C_m^K, k = 0,1,2,\cdots, k-1$,即为整个构造性人工神经网络中所有覆盖点的集合。

根据这些覆盖域可以构造一个完整的神经网络,每个覆盖域相当于一个神经元,并成为神经网络的隐藏节点。识别的方法是,给出一个输入样本,如果它在某类的覆盖领域之中,则该样本属于该类别;如果该样本不属于任何覆盖领域则作为拒识点进行处理,按照就近原则归属于某一类别;如果被多个覆盖领域覆盖则按照就近原则归属于某一类别[8,9]。

2 改进的覆盖算法

2.1 学习顺序对算法的影响分析

在构造性神经网络中,每个覆盖领域均为隐藏层中的一个神经元,然而通过实验发现在覆盖算法构造神经网络的过程中样本的学习顺序直接影响覆盖领域的大小和个数。下面通过简单的数据分类来说明学习顺序对分类结果的影响[10]。图 1 是随机选取中心得到的识别图,图 2 是通过竞争机制选择中心点,消除学习顺序产生的影响得到的识别图。显而易见图 2 的覆盖效果要好于图 1 的覆盖效果。

图 1 随机选取中心的识别别图

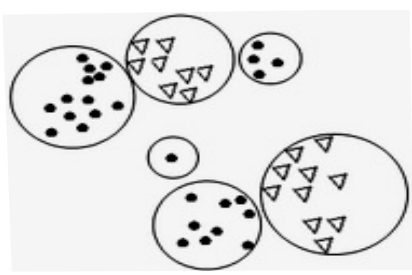

图 2 通过竞争获得的识别图

覆盖领域的大小和个数是衡量整个人工神经网络质量的重要因素。同时由于学习样本的分布是随机的,事先无法知道整个样本集的空间分布,也就无法事先确定适合的学习顺序,这样就使得该问题没有得到较为满意的解决[11]。在覆盖算法中,学习顺序中的初始点样本无法做到最优选择,是否可以在样本学习的过程中进行调优处理? 在覆盖算法的进一步研究中有交叉覆盖算法、多侧面递进覆盖算法等等改进的算法,但在这些算法中覆盖领域计算出来之后即作为“正确的”覆盖中心予以保留并把该领域覆盖的点从样本集中去除[12]。这样就使得接下来构造的覆盖领域无法对已被覆盖的样本点进行再次覆盖,更无法在已被覆

盖的样本点中选取覆盖中心。因此覆盖中心一旦形成之后就固定下来并受到保护,从而无法对由于初始点选择不当导致的错误进行调整。

2.2　改进的覆盖算法

新的覆盖算法与之前的覆盖算法相比改进之处就是:引入竞争机制,最大的不同之处就是覆盖中心形成之后并不将其覆盖的样本点从训练样本集中删除,样本点可以被重复覆盖,每个样本点记录其被覆盖的次数,从当前的覆盖域中选取覆盖次数较多的样本点作为下一个覆盖域的覆盖中心,一个覆盖域一旦被另一个覆盖域完全包含时则被包含的覆盖域将被删除。而为了防止覆盖中心的局部选择现象,将随机选择中心和最多覆盖次数的点作为中心交叉进行,改进的覆盖算法步骤如下:

Step1:按原覆盖算法在学习样本中选择类别为 k 的 Y_k 样本集合进行学习,构造两个覆盖领域 T , T'。

Step2:若在覆盖域 T 中,有被重复覆盖的点则在这些点中选取一点作为中心构造新的覆盖域赋值给 T'。否则直接进入步骤 3。

Step2.1: Y_k 中的样本进行遍历,若被 T' 覆盖则改变该样本的标志数据。

Step2.2:将 T' 与已有覆盖进行比较,删除不合适的覆盖域,进入步骤 3。

Step3: Y_k 中若有未被覆盖的点,则选取一个未被覆盖的点作为中心点,构造一新的覆盖并赋值给 T ,若没有未被覆盖的点进入步骤 4。

Step3.1: Y_k 中的样本进行遍历,若被 T 覆盖则改变该样本的标志数据。

Step3.2:将 T 与已有的覆盖域进行比较,删除不合适的覆盖域, T 通过比较则加入覆盖域集合中,返回步骤 2。

Step4:算法结束。

步骤 2 实现了竞争机制。在算法中,若一直从已覆盖的样本点中选取覆盖次数多的样本点作为下一次构造覆盖域的中心,将使得覆盖中心均是在已覆盖的样本点中选取,极易陷入一个隐性循环(覆盖中心的局部选择现象),因此加入了步骤 3。步骤 3 从未被覆盖的样本点中选择覆盖中心,并且和步骤 2 交叉进行从而防止覆盖中心点陷入局部选择。步骤 2 和步骤 3 中对覆盖领域的有效性判断则可以实现优胜劣汰的目的,即如果覆盖域 A 被另一个覆盖域 B 完全覆盖的话,则可以删除覆盖域 A,减少神经元的个数,使整个神经网络的泛化能力增强。

3　实验及结果分析

在本次试验中将改进的算法和覆盖算法进行比较以获得更加直观的信息。选取 UCI 数据库中的 Iris 数据集和 Wine 数据集进行实验分析。在 Iris 数据集中,数据被分为三种类型,每种类型有 50 条样本数据,每一条数据具有四个属性值,使用一半的数据作为训练样本,其余的数据作为测试样本。Wine 数据集中,数据同样分为 3 种类型,每条数据有 13 个属性,使用 10-fold 交叉算法(将样本平均分为 10 份,其中 9 份作为训练样本,1 份作为测试样本,取 10 次测试的平均值作为结果)。使用获得的实验结果如表 1 所示:

表 1　实验结果

数据集	方法	覆盖域	拒识数	最大识别率(%)	平均识别率(%)
Iris	一般算法	15.3	7.5	96.67	90.6
	改进算法	11.5	3.3	98.7	96
Wine	一般算法	23.8	2.1	96.1	93.8
	改进算法	17.2	1.8	97.9	97.3

从实验结果可以看出,改良后的覆盖算法在覆盖数、最大识别率和平均识别率上均有提高,在覆盖半径和覆盖点样本数方面进行比较,多次覆盖样本点作为中心明显比随机选取覆盖中心的数据要好,说明通过覆盖数来选取的覆盖中心有着较好的泛化能力,促使整个神经网络的质量得到提升。由此可见改进后的覆盖算法确实取得了较好的效果。

4　结束语

改进的覆盖算法最大的优点就是引入竞争机制。在新的覆盖中心产生之后会与已有的覆盖中心进行比较,以达到优胜劣汰的目的。这样产生的覆盖中心更加合理同时整个神经网络的泛化能力得到提高,也符合物竞天择的进化观点。在社会应用中,覆盖算法已经在经济预测、模式识别、信号处理等领域得到了大量应用并取得一定成果,对该算法的研究和改进将产生良好的社会和经济效益。

参考文献:

[1]　杨明辉. 智能计算几种经典算法解析[J]. 电脑知识与技术(学术交流),2007(15):816-816.

[2]　杨雪洁,赵　姝,张燕平. 基于构造神经网络的时间序列混合预测模型[J]. 计算机应用与研究,2008,25(10):2921-2931.

[3]　张燕萍,张　铃,吴　涛. 机器学习中的多侧面递进算法 MIDA[J]. 电子学报,2005(2):327-331.

[4]　张　玲,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.

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

[6]　Zhang Ling , Zhang Bo . A Geometrical Representation of

(下转第 36 页)

相结合后,在两个语料集上都使得召回率达到了实验中的最大值。然而,精确率指标在 $Corpus_{wiki}$ 上在此过程中不升反降,这说明基于阈值 θ_1 的补充召回带入了较大的噪声,降低了精确度。

5)由方案 5 和 6 的结果可以看出,通过设定软匹配阈值来进行候选定义句的召回和过滤,存在着对精确率和召回率方面的权衡,一个指标的上升总是伴随着另一个指标的下降。这也从一个侧面说明,试图在单特征上使用线性的分类指标去判别定义句和非定义句是不合理的,要达到更好的定义句抽取效果,势必需要加入更多的特征。这也是未来研究工作的一个方向。

实验结果表明,文中提出的术语定义句抽取方法在两个语料集上都达到了较好的效果,是明显优于单独的硬模板匹配以及软模板匹配方式的。

5 结束语

文中在分析现有的基于规则的术语定义句抽取方法不足的基础上,提出了一种将硬模板匹配与软模板匹配技术相结合的综合术语定义句抽取方法。在两个语料集上的实验结果验证了文中方法的有效性,以及相对于单独的硬模板匹配以及软模板匹配方式的优越性。

在未来的研究工作中,会尝试加入更多的语料和其他新的特征,来综合进行定义句和非定义句的分类,以期达到更好的效果。

参考文献:

[1] 荀恩东,李 晟. 采用术语定义模式和多特征的新术语及定义识别方法[J]. 计算机研究与发展,2009,46(1):62-69.

[2] 张 榕,宋 柔. 一种被定义项的识别策略[J]. 当代语言,2007,9(1):33-38.

[3] Gangemi A, Navigli R, Velardi P. The OntoWordNet project: extension and axiomatization of conceptual relations in WordNet[C]//Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics(ODBASE 2003). Catania, Italy:[s. n.],2003:820-838.

[4] Snow R, Dan Jurafsky D, Ng A Y. Learning syntactic patterns for automatic hypernym discovery[C]//Proceedings of Advances in Neural Information Processing Systems. [s. l.]:MIT Press,2005:1297-1304.

[5] Cui Hang, Kan Min-Yen, Chua Tat-Seng. Soft pattern matching models for definitional question answering[J]. ACM Transactions on Information Systems(TOIS),2007,25(2):1-30.

[6] 陈 议. 开放域的自动问答系统的研究[D]. 重庆:重庆大学,2006.

[7] Liu Bing, Chin C W, Ng H T. Mining topic-specific concepts and definitions on the web[C]//Proceedings of the 12th international conference on world wide web. Budapest, Hungary:[s. n.],2003.

[8] 贾爱平. 科技文献中术语定义语言模式研究[D]. 北京:北京语言文化大学,2002.

[9] 张 艳,宗成庆,徐 波. 汉语术语定义的结构分析和提取[J]. 中文信息学报,2003,17(6):9-16.

[10] Navigli R, Velardi P. Learning word-class lattices for definition and hypernym extraction[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden:[s. n.],2010:1318-1327.

[11] Cui Hang, Kan Min-Yen, Chua Tat-Seng. Unsupervised learning of soft patterns for generating definitions from online news[C]//Proceedings of the 13th international conference on world wide web. New York, NY, USA:[s. n.],2004.

[12] 张 榕,宋 柔. 基于互联网的汉语术语定义提取研究[C]//全国第八届计算语言学联合学术会议. 北京:清华大学出版社,2005:428-434.

(上接第 31 页)

McCulloch-Pitts Neural Model and Its Applications[J]. IEEE Trans. on Neural Networks,1999,10(4):925-929.

[7] Wu Tao, Mao Junjun, Gao Liang, et al. Covering Algorithm Based on Neighborhood Search and Its Applications[C]// Third International Conference on Natural Computation (ICNC 2007). [s. l.]:[s. n.], 2007:115-119.

[8] 李丽芳,周鸣争. 一种基于构造性核覆盖的聚类算法[J]. 计算机技术与发展,2009,19(1):88-91.

[9] Wang Di. Fast constructive-covering algorithm for neural networks and its implement in classification[J]. Applied Soft Computing,2008(8):166-173.

[10] 贾瑞玉,冯伦阔. 基于集成学习的覆盖算法[J]. 计算机技术与发展,2009,19(7):76-79.

[11] 李文娟,胡春生. 基于聚类优化覆盖的集成学习方法[J]. 计算机技术与发展,2010,20(11):51-54.

[12] 赵 姝,张燕平. 覆盖聚类算法[J]. 安徽大学学报(自然科学版),2005,29(2):28-32.

一种基于竞争的覆盖算法

万方数据 WANFANG DATA 文献链接

作者： 张月琴， 孙先洋， 刘翔
作者单位： 太原理工大学 计算机科学与技术学院,山西 太原 030024
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2012(9)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_wjfz201209010.aspx