一种基于聚类的交叉覆盖算法

吴 涛[1,2,3],尚 丽[1,2],陈黎伟[1,2]
(1.安徽大学 智能计算与信号处理教育部重点实验室,安徽 合肥 230039;
2.安徽大学 数学与计算科学学院,安徽 合肥 230039;
3.南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

摘 要:与传统的前向神经网络相比,覆盖算法具有运行速度快、精度高的特点,但覆盖算法的初始领域中心是随机选取的。实验表明网络性能与学习顺序有密切的关系。在前向神经网络交叉覆盖算法基础上提出了一种新型改进的交叉覆盖算法——基于聚类的交叉覆盖算法。该方法是一种根据聚类结果确定学习顺序的方法。实例表明这种改进的算法是确定性学习方法,可以有效减少覆盖数量,提高交叉覆盖算法的测试速度,减少拒识样本数,提高识别的精度。

关键词:交叉覆盖算法;聚类;模式;初始中心
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)11-0113-04

A Kind of Alternative Covering Algorithm Based on Clustering

WU Tao[1,2,3],SHANG Li[1,2],CHEN Li-wei[1,2]
(1.Ministry of Edu. Key Lab.of Intelligent Computing and Signal Processing,Anhui Univ.,Hefei 230039,China;
2.School of Mathematics and Computational Science,Anhui University,Hefei 230039,China;
3.State Key Lab.for Novel Software Technology,Nanjing University,Nanjing 210093,China)

Abstract:Compared with traditional algorithm of forward neural network (FNN),covering algorithm (CA) possesses some advantages, such as faster speed and higher precision. But the original centers of sphere domains are selected at random. Experiments show that the performance of network is related with the order of study closely. A new kind of algorithm named CABC, which combines covering algorithm and clustering is put forward. Instances show that this kind of algorithm is deterministic learning algorithm. It can reduce the number of sphere domains availably, low down testing time, reduce the number of rejected samples,and improve the recognition precision.

Key words:alternative covering algorithm;clustering;pattern;original center

0 引 言

文献[1,2]根据神经元的几何意义提出的多层前向神经网络的交叉覆盖算法,针对学习样本的特征构造神经网络,在一定意义上解决了多年来一直未解决的作为分类器的多层前向网络的设计问题。使用这种方法,可以十分简便地解决诸如双螺旋线的识别等难学习问题[3],快速有效地完成大量手写汉字的识别等海量数据的处理[4]。社会发展到现在,学习对人们来说是一件再熟悉不过的事。在不断的学习过程中社会才能不断进步。从哲学的角度来说,万事万物都是有它们的内在规律的,学习当然也不例外。掌握了学习的内在规律之一学习顺序,那么将会得到事半功倍的效果,反之,则会事倍功半。在人工智能研究发展的过程中,电脑并行处理问题的能力不断地加强(甚至比人脑更多更快),这就使得学习顺序更加成为人们关心的焦点问题。不论什么系统,稳定性也是同样重要的,尤其是对计算机系统来说,只有稳定了,才能做进一步的研究和发展,这一点是不言而喻的。

对于覆盖算法大家已经不再陌生。与传统的前向神经网络相比,覆盖算法具有运行速度快、精度高的特点,但覆盖算法的初始领域中心是随机选取的,实验表明网络性能与学习顺序有密切关系。文中在前向神经网络交叉覆盖算法基础上提出了一种新型改进的交叉覆盖算法——基于聚类的覆盖算法,该方法根据聚类结果确定学习顺序。实例表明这种改进的算法是确定性学习方法,可以有效减少覆盖数量,提高交叉覆盖算

收稿日期:2008-02-15
基金项目:中国博士后基金面上项目(20070411028);973 计划(2004CB318108);国家自然科学基金(60675031);安徽省高等学校省级自然科学研究项目(2006KJ244B);安徽大学学术创新团队和安徽大学人才队伍建设经费资助项目
作者简介:吴 涛(1970-),男,安徽人,博士,副教授,研究方向为机器学习、智能计算及其应用。

法的测试速度,减少拒识样本数,提高识别的精度。

1 覆盖算法

覆盖算法[1~3]是根据样本数据自身的结构,构造性地建立神经网络模型。设样本空间中的向量的长度都相等,即样本空间中的样本点都位于 $n+1$ 维空间中某个中心在原点的球面 S^n 上(若不然,可通过变换:$T:D\to S^n, T(x)=(x,\sqrt{r^2-|x|^2})$ 将样本点映射到球面 S^n 上,其中,$r\geq\max\{|x_i|\}$)。则 $\omega' x-\theta>0$ 表示球面上由超平面 $\omega' x-\theta=0$ 所分割的正半空间的部分,称为球面上的"球形领域"(见图 1)。

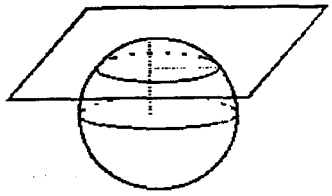

图 1 球形领域示意图

若 ω 与 x 等长,则 ω 就是这个球形领域的中心。以每个球形领域作为一个神经元,取 $\sigma(\omega' x-\theta)$ 为其功能函数,其中,$\sigma(x)=\begin{cases}1 & x>0\\ 0 & \text{其它}\end{cases}$,则功能函数就是"球形领域"的特征函数。学习过程中构造第 k 类学习样本 X_k 的"球形领域"的方法是:任取 X_k 中尚未被覆盖的点 a_i,按公式

$$d^1(\omega)=\max_{x\in X_k}\{<a_i,x>\}$$

$$d^2(\omega)=\max_{x\in X_k}\{<a_i,x>|<a_i,x>>d^1(\omega)\} \quad (1)$$

$$d(\omega)=\frac{1}{2}(d^1(\omega)+d^2(\omega))$$

计算,作以 a_i 为中心、阈值 $\theta=d(\omega)$ 的覆盖 $C(a_i)$:$\omega' x-\theta>0$,并通过求球形领域的重心和平移领域中心位置,使之可以覆盖更多的样本点。并按此方法求出样本的全部覆盖。识别的方法是:给定一个样本,若它属于某类覆盖的一个"球形领域",即可确定其类别,否则,若它不属于任何类别覆盖的任何一个"球形领域",则按距离就近原则确定其类别归属。

2 改进的覆盖算法

2.1 学习顺序对覆盖算法的影响分析

学习顺序是人们关心的焦点问题。但是它到底有什么作用呢?下面给出简单的数据分类来形象说明学习顺序对分类结果的重大影响。图 2 是随机选取中心时得到的识别图,图 3 是经过一定的学习顺序选取中心之后得到的识别图。显而易见,图 3 比图 2 的覆盖数更少,拒识数也更少,也就是说效果更好。

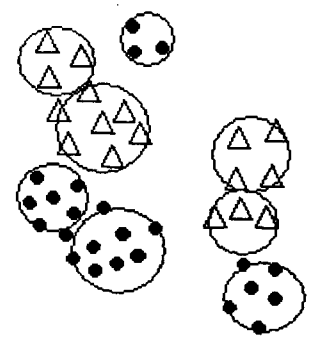

图 2 随机选取中心时得到的识别图

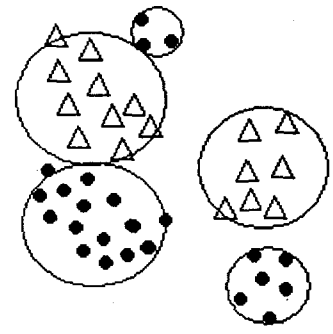

图 3 按照一定学习顺序选取中心后的识别图

新型改进的覆盖算法与以前的覆盖算法相比改进之处就在于:考虑到学习顺序对识别结果的重大影响,在原来的覆盖算法中添加了学习顺序的选择。具体来说就是:以前的交叉覆盖算法初始中心点的选择是随机的,这就使得算法的结果具有一定的随机性和不稳定性(即构造覆盖领域时,领域的中心是随机选取的,这就使得每次构造的领域大小、个数均不相同)。为解决这个问题,在算法的初始中心的选择上做了重大改进,使得整个算法更为优化。下面就改进的地方给出详细的阐述。

2.2 初始点的选择

初始点的选择上采用的新方法是[4]:先对数据进行聚类,然后采用聚类的中心作为初始点。这样一来就克服了旧的覆盖算法初始点的随机选择对结果带来的不良影响,使得结果非常稳定。

聚类(Clustering)分析是由若干模式(Pattern)组成的,通常,模式是一个度量(Measurement)的向量,或者是多维空间中的一个点。聚类分析以相似性为基础,在一个聚类中的模式之间比不在同一聚类中的模式之间具有更多的相似性。聚类分析的算法一般可以分为以下几种:

1)划分方法(Partitioning Method)。给定一个 n 个对象或元组的数据库,一个划分方法构建数据的 k 个划分,每个划分表示一个聚簇,并且 k 小于等于 n。也就是说,它将数据划分为 k 个组,同时满足如下要求:每个组至少包含一个对象,每个对象必须属于且只属

于一个组。为了达到全局最优,基于划分的聚类会要求穷举所有可能的划分,比较流行的启发式算法为 k - 平均算法和 k - 中心点算法。k - 平均算法中每个簇用该簇中对象的平均值来表示,k - 中心点算法中用接近聚类中心的一个对象来表示。

2)层次法(Hierarchical Method)。层次的方法对给定的数据对象集合进行层次的分解,可分为凝聚的和分裂的两种方法。凝聚的的方法也称自底向上的方法,一开始将每个对象作为单独的一个组,然后相继合并相近的对象或组,直到所有的组合并为一个(层次的最上层),或者达到一个终止条件。分裂的方法也称自顶向下的方法,一开始将所有的对象置于一个簇中。在迭代的每一步中,一个簇被分裂为更小的簇,直到最终每个对象在单独的一个簇中,或者达到一个终止条件。改进的层次方法有:CURE,CHAMELEON 中的做法以及 BIRCH 中的方法。

3)基于密度的方法(Density - based Method)。绝大多数划分方法基于对象之间的距离进行聚类。这样的方法只能发现球状的簇,而在发现任意形状的簇上遇到了困难。随之提出了基于密度的另一类聚类方法,其主要思想是:只要临近区域的密度(对象或数据点的数目)超过某个阈值,就继续聚类。也就是说,对给定类中的每个数据点,在一个给定范围的区域中必须至少包含某个数目的点。这样的方法可以用来过滤"噪音"孤立点数据,发现任意形状的簇。如 DBSCAN,OPTICS 等。

4)基于网格的方法(Grid - based Method)。基于网格的方法把对象空间量化为有限数目的单元,形成了一个网格结构。所有的聚类操作都在这个网格结构(即量化的空间)上进行。这种方法的主要优点是它的处理速度很快,其处理时间独立于数据对象的数目,只与量化空间中每一维的单元数目有关。基于网格的方法有:STING,CLIQUE 和 WAVECLUSTER 等。

5)基于模型的方法(Model - based Method)。基于模型的方法为每个簇假定了一个模型,寻找数据对给定模型的最佳拟合。一个基于模型的算法可能通过构建反映数据点空间分布的密度函数来定位聚类。它也基于标准的统计数字自动决定聚类的数目,考虑"噪声"数据或孤立点,从而产生健壮的聚类方法。

一些聚类算法集成了多种聚类方法的思想,所以有时将某个给定的算法划分为属于某类聚类方法是很困难的。此外,某些应用可能有特定的聚类标准,要求综合多个聚类技术。

2.3　改进算法的训练

设学习样本 X 分为 k 类,即:$X = \{X_1, X_2, \cdots, X_k\}$,构造样本 X 的覆盖:

1) 求学习样本 X 中样本的最大模 r,并将 X 中的点投影到中心在原点、半径为 r 的球面上;

2) 对样本点进行聚类,并选取聚类的中心点作为初始点 a_0;

3) 取类别号 $i = 1$,构造覆盖 $C(i)$;

4) 若 X_i 中没有尚未覆盖的点,转 9),否则,任取 X_i 中尚未被覆盖的一点 a_i;

5) 按(1) 式计算,作以 a_i 为中心、θ 为阈值的覆盖 $C(a_i)$;

6) 求 $C(a_i)$ 所覆盖的点的重心,并将其映射到球面上,设投影点为 a'_i,按(1) 式计算其阈值 θ',得球形领域 $C(a'_i)$;

7) 若 $C(a'_i)$ 覆盖的点数大于 $C(a_i)$ 所覆盖的点数,则令 $a'_i \to a_i, \theta' \to \theta$,返回 6),否则,转 8);

8) 求 ω 的平移点 ω'',并求对应的球形领域 $C(a''_i)$,若 $C(a''_i)$ 覆盖的点数大于 $C(a_i)$ 所覆盖的点数,转 6),否则,得 $C(i)$ 的一个覆盖;若 $i < k$,则 $i+1 \to i$,转 4),否则转 9);

9) 训练结束。

2.4　识别与检验

一个样本,可能有三种情况:只属于某一类领域的覆盖、不属于任何类型的覆盖、属于多种类型领域的交集。第一种情况可直接确定类别,第二种情况按就近原则确定类别,第三种情形,根据 $\max\{\sigma(\frac{\omega' x - R}{R})\}$ 确定其类别的归属。由此采取以下步骤进行识别与检验:

1) 将待识别的样本投影到中心在原点、半径为 r 的球面上;

2) 对每一个样本 x,计算:

$d(x, C(i)) = \max\limits_{C \in C(i)} \{f(x, C)\}$,其中:

$$f(x, C) = \frac{\omega' x - R}{R} \tag{2}$$

3) 求 $\max\limits_{1 \le i \le k} d(x, C(i))$ 所对应的 i,确定样本 x 的类别;

4) 统计识别的正确率。

3　应用实例

下面使用 UCI 数据库中的数据给出 5 个例子,通过与一般覆盖算法的结果以及支持向量机(SVM)算法的结果进行比较,说明改进算法的效果。

支持向量机方法[5]是建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精

度,Accuracy)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折中,以期获得最好的推广能力(Generalization Ability)。支持向量机方法的主要优点有:可以解决小样本情况下的机器学习问题,提高泛化性能,解决高维问题,解决非线性问题,避免神经网络结构选择和局部极小点问题。

3.1 实验数据与结果

采用数据库中的 Ionosphere, Wine, Iris, Mushroom 这五组数据来做试验。其中 Ionosphere 数据库以其前 200 个样本作为训练样本,后 150 个样本作为测试样本。其他三个数据都是采用 10 - fold 交叉算法(即将样本平均分为 10 份,其中 9 份作为训练样本,1 份作为测试样本,重复 10 次,测试结果取 10 次的平均值)。所用数据信息见表 1。

表 1 实验数据表

数据库	提供者	维数	类数
Mushroom	Jeff Schlimmer	21	2
Ionosphere	Vince Sigillito	34	2
Wine	Stefam Aeberhard	13	3
Iris	R. A. Fisher, Michael Marshall	4	3

3.2 实验结果及分析

实验结果见表 2。

实验结果表明改进算法对 Ionosphere 和 Mushroom 的效果非常好,提高了识别率:平均识别率分别从原来的 93.264% 和 96.347% 提高到了 99.338% 和 99.176%;而 SVM 的精度只有 87.778% 和 96.273%,缩减了测试时间:分别从原来的 0.0047s 和 0.043586s 减少到 0.003625s 和 0.0046875s,而且减少了覆盖数:从原来的 82.2 个和 69.406 个减少到 47 个和 26 个。对其他的几个数据也减少了覆盖个数,识别率也有一定的提高,并且结果非常稳定。由此可见,学习顺序确实对实验结果有重大的影响。对原覆盖算法采用先聚类,再求覆盖的修改方法取得非常好的效果。

4 结束语

覆盖算法已被广泛应用于模式识别、信号处理、模糊控制、金融预测等方面,并取得一定的成就。改进之后的基于聚类的覆盖算法以其独特的稳定性将在这些领域中得到更大更广泛的应用。

表 2 实验结果

		改进算法	一般算法	SVM	
Ionosphere	覆盖数	47	82.2	精度(%)	87.778
	拒识数	5.1	5.8		
	测试时间(s)	0.003625	0.0047		
	最大识别率(%)	99.338	95	测试时间	0.0010014
	平均识别率(%)	99.338	93.264		
Wine	覆盖数	18.4	23.80	精度(%)	97.778
	拒识数	1.8	1.8		
	测试时间(s)	0.003125	0.0016		
	最大识别率(%)	97.222	96.1111	测试时间	0.0010014
	平均识别率(%)	97.222	94.0444		
Iris	覆盖数	11.8	16.8	精度(%)	96.667
	拒识数	0.9	0.8		
	测试时间(s)	0.001563	0.0031		
	最大识别率(%)	98	96.6667	测试时间	0
	平均识别率(%)	96.02	94.56		
	识别率(%)	92.4	91.9		
Mushroom	覆盖数	26	69.406	精度(%)	96.273
	拒识数	24.7	41.831		
	测试时间(s)	0.0046875	0.043586	测试时间	0.2734
	识别率(%)	99.176	96.347		

参考文献:

[1] 张燕平, 张 铃, 吴 涛. 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] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明 等译. 北京: 机械工业出版社, 2001: 231 - 235.

[5] Vapnik V N. 统计学习理论的本质[M]. 张学工译. 北京: 清华大学出版社, 2000.

(上接第 112 页)

参考文献:

[1] 胡建强, 郭长国, 王 怀, 等. 一种基于 P2P 网络的服务发现方法[J]. 电子学报, 2005, 33(12A): 2503 - 2507.

[2] Schmidt C, Parashar M. A peer - to - peer approach to Web service discovery[J]. World Wide Web, 2004, 7(2): 211 - 229.

[3] Stoica I, Morris R, Liben - Nowell D, et al. Chord: A scalable peer - to - peer lookup service for internet applications [C]// Proc. of ACM SIGCOMM. New York: ACM Press, 2001: 149 - 160.

[4] EI - Ansary S, Alima L O, Brand P, et al. Efficient Broadcast in structured P2P Networks[C]//Proc. of 2nd International Workshop on Peer - to - Peer Systems, vol. 2735. Berlin: Springer - Verlag, 2003: 304 - 314.

[5] 刘志忠, 王怀民, 周 斌. 一种双层 P2P 结构的语义服务发现模型[J]. 软件学报, 2007, 18 (8): 1922 - 1931.