基于聚类优化覆盖的集成学习方法

李文娟,胡春生
(解放军炮兵学院,安徽 合肥 230031)

摘 要:传统的覆盖方法形成的覆盖都是"优簇",但是无法形成非球状的覆盖;而聚类求覆盖的方法就可以得到非球状覆盖,但是由于很难事先找到合适的相似度,所以无法求得全部"优簇"。文中把两者的优点结合起来并加以推广,与SVM,NaiveBayes,交叉覆盖等学习方法相结合,形成基于聚类优化覆盖的集成学习方法,这样求得的覆盖既可以是非球状覆盖,又是全"优簇",优化了覆盖领域。实验证明该方法产生的覆盖不仅数量上较少,并且覆盖的准确率较高,具有一定的抗噪声能力。

关键词:聚类;覆盖;相似度
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2010)11-0051-04

A Combined Learning Algorithm of Optimum Covering Based on Clustering

LI Wen-juan, HU Chun-sheng
(People's Liberation Army Artillery Institute, Hefei 230031, China)

Abstract: The pure clusters can be gotten by the traditional covering algorithm, though the shape of the coves is just ball alike. And the different shapes of covers can be attained if the clustering method is proposed to get covers. Combine the two to get optimum covers, and then extend to other learning algorithms such as NaiveBayes, SVM, cross-coverage, etc. It is called a combined learning algorithm of optimum covering based on clustering. Not only it can get covers of different shapes, but also get clusters of all pure. The experiment proves the algorithm generates not only fewer covers, but also of greater precision. It has some ability of anti-noise.

Key words: clustering; covering; similarity

0 引 言

张铃、张钹教授提出的设计神经网络的覆盖算法[1],是把神经元与几何上样本的球形邻域对应起来,这样就可以通过球面投影变换将神经网络的最优设计问题转化为某种最优覆盖问题,并且证明出任何一种求(次优)最小覆盖的方法与球形邻域法相结合,都能给出一个神经网络的学习算法。国内很多学者对覆盖算法提出改进,比较经典的有核覆盖算法[2,3]、覆盖融合算法[4]、交叉覆盖算法[5,6]。核覆盖方法是把核函数引入到覆盖算法中,使得原样本空间不可分的样本在核空间变得可分,并且核函数法求得的是最大分割解;覆盖融合方法相当于在覆盖的基础上的一次求优过程,减少了覆盖数量,平滑了覆盖边界;交叉覆盖算法是通过不断构造正反覆盖的方法使得覆盖范围尽可能大。

传统的覆盖方法不管是在原始特征空间还是在变换后的特征空间,形成的覆盖都是"球形覆盖",并且所有覆盖的描述形式是单一的(都是用覆盖中心、覆盖半径或覆盖边界去描述)。而样本的分布,不管是在原始特征空间还是在变换后的特征空间未必都成球状,如图1(a)所示,用传统的覆盖方法求得的覆盖必然类似如图1(b)(虚线)所示,而球型覆盖之间要么有重合,要么有空隙,这是球形覆盖固有的缺点,显然不是较优覆盖。直观地看,最优的覆盖应该如同图1(c)所示。而聚类方法形成的簇可以是任意形状(如DBSCAN等),如果把聚类中形成的簇与覆盖对应起来,这样就可以形成非"球覆盖"覆盖。由此,文献[7]把聚类引入到覆盖的求取之中,优化覆盖区域。

但是单纯用聚类的方法很难事先找到合适的相似度,所以无法求得全部"优簇"。所以文中把两者结合起来优化覆盖区域。并推广至其它分类方法形成基于聚类优化覆盖领域的集成学习方法。

收稿日期:2010-03-29;**修回日期**:2010-07-06
基金项目:欧盟项目TYPES资助(types project 29001)
作者简介:李文娟(1985-),女,硕士研究生,研究方向为智能信息处理。

a b c

图 1 覆盖比较图

1 覆盖算法

问题：给定一个训练样本集 $K = \{(x_1, y_1), \cdots, (x_p, y_p)\}, x_i \in Rn, y_i \in \{0,1\}k$，要求构造一个向神经网络 N，满足：$N(x_i) = y_i, i = 1, \cdots, p$。

覆盖算法：

(1) 利用变换：

$f: R^n \to S^{n+1}, f(x) = (x, \sqrt{R^2 - |x|^2})$

(2) 求一组覆盖：

$C = \{C_1^1, C_2^1, \cdots, C_{n_1}^1, C_1^2, C_2^2, \cdots, C_{n_2}^2, \cdots, C_1^k, \cdots, C_{n_k}^k\}$，其中每一覆盖子组 $\{C_1^i, C_2^i, \cdots, C_{n_k}^i\}$ 只覆盖第 i 类的样本点，则 C 就是覆盖算法得到的解。

2 聚类算法

“物以类聚，人以群分”。聚类[8](clustering)就是将数据对象按其性能分组成为多个簇(cluster)，使得同一个簇中的对象之间具有较高的相似性(similarity)，而不同簇中的对象具有较大的相异性(dissimilarity)。

简单的聚类(cluster)问题可以描述如下[9,10]。

聚类问题：设在 N 维空间中，给定一个有限点集 K，再给出一个量度相似性函数 D 和一个相似度阈值 d，要求将 K 进行分类，并满足下面条件：

(1) 将 K 分成子类 $\{K_1, K_2, \cdots, K_m\}$，并构成一个划分；

(2) 则 x, y 属于同一类。

可以看出，聚类预先不知道目标数据有关类别的信息，而是把那些相似性测度较大(比如欧氏距离较小)的样本点归为一类(簇)。而分类与其有很大的不同，分类模型中，存在着学习样本，这些学习样本的类标号是已知的，分类的目的是从训练样本集中探索每一类样本点的规律。从着色的角度更加直观形象地理解分类的本质：选定特征之后，样本集就被表示成特征空间中一群点，可以按照先验知识来给样本点着色，使得同一类样本点有同样的颜色，而不同类的样本点的颜色也不同，研究目标就是摸清楚：红色点有什么规律？绿色点又有什么规则可循？再来一个新的样本点，当投射到同一个特征空间之后，判断它应该染成红色、绿色还是其他某种颜色[11]？

所以，分类是“示例学习”，而聚类是一种无监督的学习方法，属于“观察学习”。

3 基于聚类思想的求覆盖方法(CbC)

现实世界中，样本的归类所依靠的标准应该是分层次而不一样的。例如：要把全校的学生进行分组归类，标准有很多，如年级、班级、性别、身高等等。若要求形成的每一组都属于同一班级同一性别，按照其中某一种标准都无法完成这个分组的任务，应该先以班级为标准来划分成若干子类，然后再在子类中按性别标准来划分，这是个多粒度分层次的概念。

传统的覆盖方法不管是在原特征空间还是在变换后的特征空间，形成的覆盖都是“球形覆盖”；并且覆盖的描述形式是单一的，都是用覆盖中心、覆盖半径或覆盖边界去描述。而某些聚类方法形成的簇可以是任意形状(如 DBSCAN 等)，把聚类中形成的簇与覆盖对等起来，这样最后形成的覆盖可以形状多样的，而非单一的“球覆盖”。此外聚类簇是用相似度的概念去表示，所以可以通过不断地修改相似度去求簇而得到覆盖，最后求得的覆盖的表示方法在不同的覆盖里面是不一样的，是用各自簇的相似度去描述，从而达到分层次多粒度的效果，这样更有利于异类样本的区分，较容易地完成覆盖任务。

为了便于描述聚类的结果的好坏，先简单说明非优簇和优簇的概念。

定义：非优簇，如果 β = MAX(簇中样本按类别计数统计)/簇中样本个数 $< 1 - B$(B 为参数，称为抗噪声系数)，则称该簇是非优簇，反之，称为优簇，即覆盖。

当 $B = 0$ 时，表示该分类器不具备抗噪声的能力，即该簇中样本必须完全属于某一个类时才是优簇。如图 2UCI 标准数据集 iris 某次聚类结果的按类别统计所示。其中，形成簇数为 5 个(0、1、2、3、4)，Iris - setosa、Iris - versicolor、Iris - virginica 分别为三个类别的名称，例如簇 0 有 11 个样本全部属于 Iris - virginica 类；簇 1 有 28 个样本，27 个属于 Iris - versicolor，1 个属于 Iris - virginica，如果令 $B = 0$，则簇 0、2、3、4 的 β 为 1，属于优簇，而簇 1 的 $\beta = 27/28 < 1 - 0$，为非优簇。

```
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
  0  0  0  0  0 | Iris-setosa
  0 27  0  0 23 | Iris-versicolor
 11  1 22 16  0 | Iris-virginica
```

图 2 聚类结果统计

求覆盖的方法如下：

(1)根据某种聚类算法及相似度求簇,得到簇 $B_1 = \{C_1, C_2, \cdots, C_n\}$。

(2)保留优簇 $A_1 = \{C_{i_1}, C_{i_2}, \cdots, C_{i_x} \mid i_x \in 1,2, \cdots, n\}$ 的结果,放弃非优簇的聚类结果,放在一起成为新的学习样本。

(3)修改相似度(或者修改聚类方法,修改聚类方法即是在修改相似度度量函数)。

(4)goto(1),直到求得的都是优簇。

聚类方法形成的簇可以是任意形状(如 DBSCAN 等),如果把聚类中形成的簇与覆盖对应起来,这样就可以形成非"球状"覆盖。此外覆盖用各自簇的相似度去描述,在不同的覆盖里面也可以是不同的,从而达到分层次多粒度的效果,这样更有利于异类样本的区分,较容易地完成覆盖任务。

4　基于聚类优化覆盖的集成学习方法

传统的覆盖方法形成的覆盖都是"优簇",但是覆盖形状却是球状,无法形成非球状的覆盖;而聚类求覆盖的方法就可以得到非球状覆盖,但是由于很难事先找到合适的相似度,所以无法求得全部"优簇"。所以就可以把两者的优点结合起来,这样既可以求得非球状覆盖,也可以求得全优簇。并且不同的覆盖表示形式也是多样的,从而达到分层次多粒度的效果,这样更有利于异类样本的区分,较容易地完成覆盖任务。

4.1　聚类求覆盖与传统求覆盖组合求覆盖的方法

算法描述如下:

学习时:

输入:条件参数 B,为簇抗噪声系数,文中的实验取 2%,即一个优簇允许 2% 的噪声。

第一步:

(1) 根据某种聚类算法及相似度求簇,得到簇 $B_1 = \{C_1, C_2, \cdots, C_n\}$。

(2) 保留优簇 $A_1 = \{C_{i_1}, C_{i_2}, \cdots, C_{i_x} \mid i_x \in 1,2, \cdots, n\}$ 的结果,放弃非优簇的聚类结果,放在一起成为新的学习样本。

(3) 修改相似度(或者修改聚类方法,修改聚类方法即是在修改相似度度量函数)。

(4)goto(1),直到求得的都是优簇或者是无法用当前的相似度求得优簇。

第二步:在第一步留下的非优簇 $B_i = \{C_1, C_2, \cdots, C_{n(i)}\}$ 内,针对每一个 $C_k, k = 1, \cdots, n(i)$ 用传统的求覆盖算法来构造覆盖 $A_{C_k} = \{D_1, D_2, \cdots, D_{n(C_k)}\}$。

输出:$\{A_1 \cup A_2 \cdots \cup A_m\}, \{B_1, B_2, \cdots, B_i\}$ 以及 $\{A_{C_1}, A_{C_2}, \cdots, A_{C_{n(i)}}\}$,最后得到优簇 $\{A_1 \cup A_2 \cup \cdots \cup A_m \cup A_{C_1}, A_{C_2}, \cdots, A_{C_{n(i)}}\}$ 及候选簇 $\{B_1, B_2, \cdots, B_i\}$,其中优簇 $\{A_1 \cup A_2 \cup \cdots \cup A_m\} \cup \{A_{C_1}, A_{C_2}, \cdots, A_{C_{n(i)}}\}$ 即是所求的覆盖。

测试时:根据学习时用到的相似度度量方法依次验证。

第一步:

for $B_j = B_1$ to B_i

求 $B_j = \{C_1, C_2, \cdots, C_{n(j)}\}$ 中与 x 最相似的簇 C_x

if $C_x \in A_j$

则 x 属于 C_x 对应的类别,return;

else

如果不在,继续

end for

第二步:进入 $A_{C_k} = \{D_1, D_2, \cdots, D_{n(C_k)}\}$ 中用传统覆盖方法判断 x 所属的覆盖 D_x,其中 C_k 为第一步最后一个候选簇 B_i 中与 x 最相似的簇。

4.2　基于聚类优化覆盖的集成学习方法

基于机器学习的分类算法与传统的覆盖方法一样都是完成对空间的划分,上述方法中第二步传统覆盖算法也可以替换为其它的分类算法,所以可以对上述方法作一推广,基于聚类优化覆盖的集成学习算法组合成很多形式,以满足不同的数据集,如聚类求覆盖与 SVM 结合起来(聚类求覆盖 + SVM[12]),聚类求覆盖与 Bayes 结合起来(聚类求覆盖 + Bayes)等等集成学习方法,算法只是把覆盖 $A_{C_k} = \{D_1, D_2, \cdots, D_{n(C_k)}\}$ 换成对应的分类器 A_{C_k} 即可。

此方法相对单纯的分类器而言,相当于多了个预划分的过程。把容易划分的区域先分开,对交错复杂的区域采取逐步分解的解决方式。"保证解决容易的,力争解决难的",而不期望采用同一个标准、同一模式来完成对空间的划分。这样划分后的空间具有多粒度、分层次的结构。

4.3　实验结果与分析

文中的实验随机选择 80% 样本作为学习样本,20% 作为测试样本,准确率是预测正确的样本在测试样本中所占的比例,取多次实验平均值。文中的实验中采用 DBSCAN、EM 聚类方法组合起来求优化覆盖。

实验一:标准数据集合上的比较。

选用 UCI 数据库上的 Iris, Diabetes, Mushroom 三组真实数据作为实验数据。从表 1 可以看出,基于聚类优化集成方法的准确率都比没有用聚类优化覆盖的方法准确率都要高。并且从神经元个数来看,如表 2 所示:Iris 和 Mushroom 数据库上用聚类方法就可以直

接求得覆盖,算法没有进入第二步,而 Diabetes 数据集上先用聚类求覆盖的方法得到两个覆盖,然后对剩下的样本进行划分,形成 8 个子集,分别形成 8 个分类器,并且聚类求覆盖 + 交叉覆盖的方法中每个分类器的覆盖数是 195 个,比普通覆盖方法形成的覆盖数(322 个)要少很多。

表 1 不同集成方法的准确率比较

算法 / 准确率 / 数据集	聚类求覆盖 + SVM	聚类求覆盖 + Bayers	聚类求覆盖 + 交叉覆盖
Iris	99.9%	99.9%	99.9%
Diabetes	97.4%	85.06%	70.58%
Mushroom	99.9%	99.8%	99.8%

表 2 不同集成方法对应的神经元个数比较

算法 / 准确率 / 数据集	聚类求覆盖 + SVM	聚类求覆盖 + Bayers	聚类求覆盖 + 交叉覆盖
Iris	16 + 0	16 + 0	16 + 0
Diabetes	2 + 6	2 + 6	2 + 6(195)
Mushroom	22 + 0	22 + 0	22 + 0

实验二:噪声数据的识别。

选取 UCI 数据库中的 Waveform 数据,Waveform 数据有 41 个属性,其中有 21 个含有噪声的属性和一个类别决策属性,共 5000 个样本,同样选取 80% 作为学习,20% 作为测试。该数据集上记载的方法的识别率,优化 Bayes 是 86%,Nearest Neighbor Alogorithm 是 78%。

表 3 噪声数据集 Waveform 上的比较

聚类求覆盖 + SVM	识别率	90.7%
	神经元个数	6 + 8
聚类求覆盖 + NaiveBayes	识别率	6 + 8
	神经元个数	81.9%
聚类求覆盖 + 交叉覆盖	识别率	81.16%
	神经元个数	6 + 8(670)
SVM	识别率	86.9%
NaiveBayes	识别率	80.6%
覆盖	神经元个数	1967
	识别率	77.4%

图 3 和表 3 显示基于聚类优化的集成学习方法在识别率上比没有用覆盖优化方法的识别率要高,如聚类求覆盖 + SVM 比 SVM 的识别率要高,聚类求覆盖 + NaiveBayers 比 NaiveBayers 要高等等,所以从一定的程度上证明 4.3 节讲述的基于聚类优化的方法的抗噪能力。从神经元个数来看,基于聚类优化的集成学习方法首先用聚类求覆盖的方法得到 6 个覆盖,然后对剩下的样本进行划分,形成 8 个子集,分别形成 8 个分类器,并且聚类求覆盖 + 交叉覆盖的方法中每个分类器的覆盖数是 670 个,而传统的覆盖方法形成 1967 个覆盖,神经元个数远远少于传统的覆盖方法。

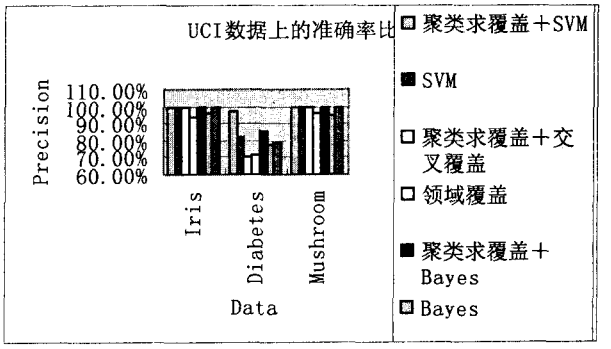


图 3 标准数据集上的准确率比较图

5 结束语

基于聚类优化覆盖的集成方法,相当于在其它分类方法的基础上做了初始的划分。从学习的角度看,该方法秉着"先学习容易的"的思想,并且在学习的过程中按照不同的学习标准(相似度不同)来观察学习,更好地完成样本空间的划分,并且使得划分后的空间具有多粒度、分层次的结构,实验证明方法的可行性。

参考文献:

[1] Zhang Ling, Zhang Bo. A geometrical representation of McCullochr Pitts neural model and its applications[J]. IEEE Transactions on Neural Networks, 1999, 10(4): 925 - 929.

[2] 张燕平,张 铃,吴 涛,等.基于覆盖的构造性学习算法 SLA 及在股票预测中的应用[J].计算机研究与发展,2004(6):979 - 984.

[3] Wu Gao - Wei, Tao Qing, Wang Jue. Kernel covering algorithm and a design principle for feed - forward neural networks[C]//In: Proceedings of the 9th International Conference on Neural Information Processing. Singapore: [s. n.], 2002: 1064 - 1068.

[4] 吴 涛,张 铃,张燕平.机器学习中的核覆盖算法[J].计算机学报,2005(8):1295 - 1301.

[5] 张 铃,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):737 - 742.

[6] 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.

[7] 田福生,张燕平.基于聚类优化覆盖的学习方法[J].计算机科学,2008(增刊):196 - 201.

[8] 方开泰,潘恩沛.聚类分析[M].北京:地质出版社,2006:89 - 112.

(下转第 58 页)

$Z'^{(t+1)}_j$ 将从输出层神经元 j 的输出值 $Z'^{(t)}_j$ 中获得，即 $Z'^{(t+1)}_j = Z'^{(t)}_j$。神经元的状态将进行重复变化，直到整个系统达到平衡状态。

4 系统性能指数

峰值信噪比 PSNR 的计值是评价系统性能的标准。

$$\mathrm{PSNR}=10\times\log\left(\frac{B^2}{\mathrm{MSE}}\right) \tag{5}$$

上式中，B 为灰度级最大值，取 255。MSE 是原图像与处理图像之间的均方误差。

$$\mathrm{MSE}=\sqrt{\frac{\sum_{i=1}^{N}(Z_i^{\mathrm{true}}-Z'_i)^2}{N}} \tag{6}$$

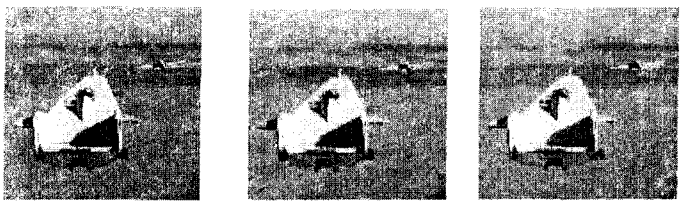
其中，Z^{true} 为初始图像中像素 i 的灰度级；Z'_i 为解压缩图像中像素 i 的灰度级；N 为图像中所有像素的个数。PSNR 值越大，代表图像失真越少。

5 实验分析

文中实验采用图 4 为样本图像，在 Matlab 环境下[9-12]，实现神经网络的过程基于四叉分块方法对分形图像的压缩，同时采用改进后的最速下降法，即使用梯度下降动量和自适应学习算法，目标误差设为 0.001，最大迭代次数为 500，在神经网络输入节点为 16 时，取隐含层神经元数为 8 来训练网络，训练结果见表 1。其中实验 1 为神经网络对样本图像的压缩，实验 2 是神经网络对样本的分形图像进行压缩。

表 1 基于神经网络的分形图像压缩性能

实验	迭代次数	Time(s)	PSNR(dB)
1	19	29.0830	78.1804
2	12	18.9070	79.3908



样本图像 实验 1 实验 2

图 4 样本图像

6 结束语

介绍了神经网络在分形图像压缩中的应用，在实验中结合非线性网络和最速下降法实现了对分形图像的压缩，获得较高质量的解压缩图像。使用神经网络方法进行并行方式计算，使得图像压缩以该并行方式快速执行，从而缩短了压缩时间。

文中的创新点在于，在详细研究了各种图像压缩技术的基础上，着重研究分形图像压缩技术和神经网络技术，将两者进行合理的结合，提出基于神经网络的分形图像压缩技术，提高压缩质量，缩短压缩时间，取得了良好的压缩效果。

参考文献：

[1] Stark J. A Neural Network to Compute the Hutchinson Metric in Fractal Image Processing[J]. IEEE Trans Neural Network, 1991(1):156-158.

[2] 王曙光. 分形图像压缩编码的原理与进展趋势[J]. 福建电脑, 2004(9):9-10.

[3] 黄贤武, 王加俊, 李家华. 数字图像处理与压缩编码技术[M]. 西安: 西安电子科技大学出版社, 2000.

[4] Fisher Y. Fractal Image Compression: theory and application [M]. New York: Springer-Verlag, 1995:49-51.

[5] 高瀚昭, 王俊生, 谢 立. 引入非线性变换的分形图像压缩编码[J]. 通信学报, 2000, 21(4): 89-92.

[6] Popescu D C, Dimca A, Yan Hong. A nonlinear model for fractal image coding[J]. IEEE Trans. Image Processing, 1997, 6(3):372-382.

[7] FISHERY. Fractal Image Compression[M]. New York: Springer, 1994.

[8] Shusterman E, Feder M. Image Compression Via Improved Quadtree Decompression Algorithm[J]. IEEE Trans: Image Process, 1994(6):207-215.

[9] 王爱玲, 叶明生, 邓秋香. MATLAB R2007 图像处理技术与应用[M]. 北京: 电子工业出版社, 2008:35-39.

[10] 薛定宇. 控制系统计算机辅助设计-MATLAB 语言及应用[M]. 北京: 清华大学出版社, 1996.

[11] 薛定宇. 反馈控制系统设计与分析-MATLAB 语言及应用[M]. 北京: 清华大学出版社, 2000.

[12] 楼顺天, 于 卫. 基于 MATLAB 的系统分析与设计控制系统[M]. 西安: 西安电子科技大学出版社, 1998.

(上接第 54 页)

[9] 张 铃, 张 拨. 多层反馈神经网络的 FP 学习和综合算法[J]. 软件学报, 1997(4):252-258.

[10] Chen Q C. Generating-shrinking algorithm for learning arbitrary classification[J]. Neural Networks, 1994, 5(7):1477-1489.

[11] 卜东波, 白 硕, 李国杰. 聚类/分类中的粒度原理[J]. 计算机学报, 2002(8):150-154.

[12] Vapnik V N. 统计学习理论的本质(中文版)[M]. 北京: 清华大学出版社, 2000.