改进型支持向量机在水质分类中的应用研究*

刘 保
(淮南职业技术学院 网络中心,安徽 淮南 232001)

【摘 要】文章分别使用BP、RBF等神经网络和支持向量机等非线性方法对相同的水质数据建立分类模型。使用支持向量分类机建立水质分类模型过程中,选用RBF核函数,结合归一、降维等数据预处理手段,利用网格搜索算法对参数进行寻优,得出水质分类模型。实验结果证明在非线性方法中,采用支持向量机并结合相应的数据预处理手段这种方案得出的分类准确率更高,更加具有推广性。

【关键词】水质评价;分类;支持向量机;神经网络;核函数
【中图分类号】X824 【文献标志码】A 【文章编号】1673-1891(2015)01-0042-04
DOI:10.16104/j.cnki.xccxb.2015.03.013

1 神经网络在水质分类中的应用

1.1 RBF神经网络在水质分类中的应用

1.1.1 RBF神经网络概述

作为一种前向神经网络,RBF神经网络和大多数前向网络类似。其网络结构通常也分为输入层、隐含层、输出层。结构简单、训练学习速度快是这种神经网络的特点,能够在理论上逼近任何非线性函数。

1.1.2 利用RBF神经网络建立水质分类模型

文章共选取189个水质样本,从中选取89个水质样本作为训练样本,另外89个水质样本作为测试样本。利用Matlab10.0编程建立分类模型,首先对训练样本和测试样本做数据预处理,即将训练集和测试集合并在一起做归一操作。在数据预处理过程中并没有使用主成分分析这一步骤,原因是实验证明对其进行主成分分析不但没有提高其分类准确率反而降低了最终的准确率,由此可以看出主成分分析包括归一在数据预处理的过程中并不是必须的步骤。然后通过径向基神经元建立概率神经网络,最后使用测试集来验证该模型的分类准确率。

使用RBF神经网络建立分类模型,分类结果如图1所示:

从分类结果对比图1可以看出,利用RBF神经网络建立分类模型的测试结果中,测试集中3个样本本来属于1类,被模型误分类为2类;有6个样本本来属于3类,被模型误分类为2类。共有9个测试数据分类有误,准确率为0.898876。

1.2 BP神经网络在水质分类中的应用

1.2.1 BP神经网络概述

由鲁梅哈特等科学家领导的团队于20世纪80年代提出的一种神经网络模型叫做BP神经网络。作为一种典型的反向传播多层前馈网络模型之一在其出现后得到非常广泛的应用,RBF神经网络比较相似,BP神经网络模型的拓扑结构也是由BP输入层(输入),隐藏层(隐层)和输出层(输出)三部分组成。

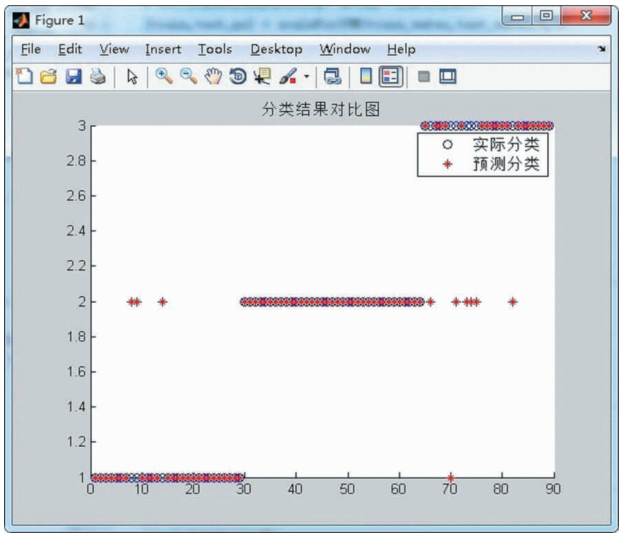

图1 分类效果图

1.2.2 利用BP神经网络建立水质分类模型

和通过RBF神经网络建立分类模型类似,在使用BP神经网络建立分类模型时也是选取89个水质样本作为训练样本,另外89个水质样本作为测试样本,首先对训练样本和测试样本做数据预处理,即将训练集和测试集合并在一起做归一操作。在数据预处理过程中也没有使用主成分分析这一步骤。建立BP神经网络分类模型,最后使用测试集来验证模型的分类准确率。

利用BP神经网络建立分类模型并测试分类准确率,分类结果如图2所示:

收稿日期:2015-03-25
*基金项目:淮南职业技术学院基金项目"改进型支持向量机在水质分类中的应用研究"(项目编号:HKJ13-3)。
作者简介:刘保(1979-),男,河北沧州人,讲师,硕士,研究方向:计算机应用。

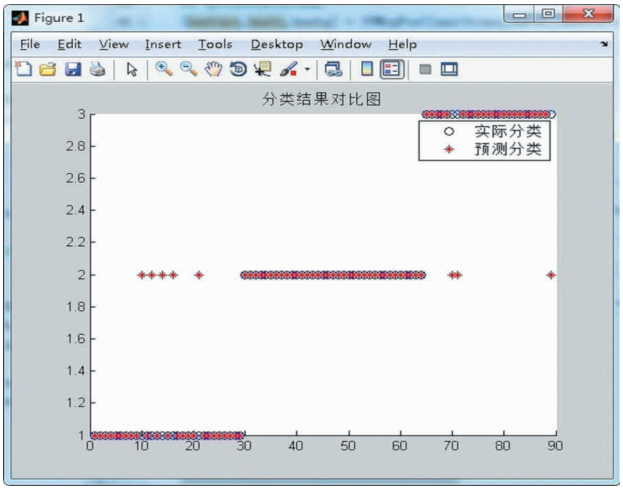

图2　分类结果对比图

从分类对比图6可以看出,利用BP神经网络建立分类模的测试结果中,测试集中有5个测试样本本来属于1类,被模型误分类为2类;有3个测试样本本来属于3类,被模型误分类为2类。共有8个样本分类有误,准确率为0.910112。

2 支持向量机在水质分类中的应用

2.1 支持向量机的概念

数据挖掘的有效工具包括统计模式识别、线性或非线性回归以及人工神经网络。这些数据挖掘工具目前随着计算机软硬件技术的高速发展已经得到了广泛的应用。有一个“小样本难题”摆在眼前,许多实际课题中已知样本数量较少,而传统的模式识别或人工神经网络方法都要求有较多的训练样本。如何解决这类“小样本难题”,既用较少的样本数量就可以得到推广能力较好的模型,在支持向量机技术没有出现之前是模式识别研究领域的一个难题。

支持向量机实现的是如下思想:将非线性可分的样本输入空间通过某种特定的非线性映射方法映射到一个高维特征空间并使其线性可分。最优分类超平面正是通过这样一个高维特征空间构造出来,从而实现分类。

2.2 利用支持向量机建立水质分类模型

2.2.1 数据选取

```
% 选定训练集和测试集
% 将第一类的1-30,第二类的60-95,第三类的131-153做为训练集
train_water = [water(1:30,:);water(60:95,:);water(131:153,:)];
% 相应的训练集的标签也要分离出来
train_water_labels = [water_labels(1:30);water_labels(60:95);water_labels(131:153)];
% 将第一类的31-59,第二类的96-130,第三类的154-178做为测试集
test_water = [water(31:59,:);water(96:130,:);water(154:178,:)];
% 相应的测试集的标签也要分离出来
test_water_labels = [water_labels(31:59);water_labels(96:130);water_labels(154:178)];
```

图3　训练集和测试集的选取

其中:train_water和train_water_labels分别是测试集和测试数据的标签,test_water和test_water_labels分别为测试集和测试集的标签。

2.2.2 数据的预处理

(1)数据归一

文章采用Matlab10.0自带的mapminmax函数对测试集和训练集进行归一操作。

```
%%
[mtrain,ntrain] = size(train_water);
[mtest,ntest] = size(test_water);
dataset = [train_water;test_water];
[dataset_scale,ps] = mapminmax(dataset',0,1);
dataset_scale = dataset_scale';
train_water = dataset_scale(1:mtrain,:);
test_water = dataset_scale( (mtrain+1):(mtrain+mtest),: );
```

图4　训练集和测试集合并归一

(2) 主成分分析

主成分分析是一种常见的统计分析方法,也是一种数学降维的方法,可以从众多变量中找出几个综合变量,最大程度地让这几个综合变量能够代表原来的众多变量,而这几个综合变量之间则关系不大或者没有任何关系。这种从众多变量提取出少量几乎没有关联的综合变量的方法叫做主成分分析。

PCA就是这样的一种分析方法。PCA一般用来对数据进行降维,文章就是利用PCA降维这种方法来提取这种线性组合,目的是为了去除高维空间中的冗余数据信息和噪声信息,通过降维算法来寻找内部数据的本质结构特征,在某种情况下对最终的分类准确率和训练时间有较大改善。文章采用MATLAB中princomp函数来实现降维护,降维效果分别如图5所示:

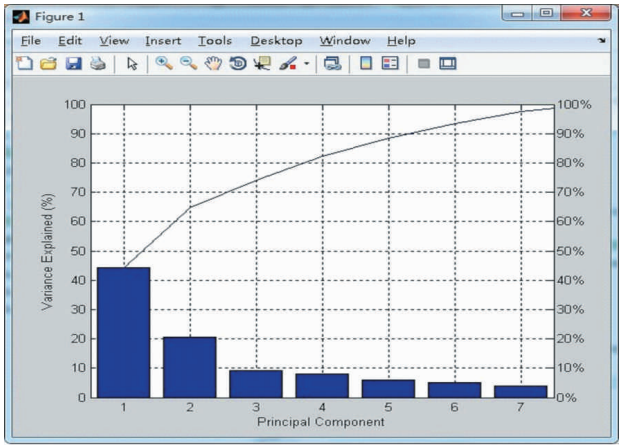

图5　PCA降维效果图

(3) 参数寻优

交叉验证是一种统计评估方法,分析机器学习方法对独立数据集的泛化能力(推广能力),过拟合问题得到解决(为了得到一致假设而使假设变得过

度复杂)。参数蚁群算法、网格搜索算法、遗传算法等都是参数寻优的算法,各种寻优算法都有其优缺点,在文章中采用各种算法对最终的结果影响很小。所以文章选用网格搜索算法对参数c和g进行寻优。

所谓网格搜索算法,也就是说遍历各种可能的c和g的值,对每一组c和g的值进行交叉验证,找出一组能够产生最高精确度的c和g的值,这就是网格搜索算法的原理。网格算法其实就是使用多循环的方法来建立程序。为了更加形象的说明参数c和g寻优的效果,文章以3D图的方式展示寻优的结果。

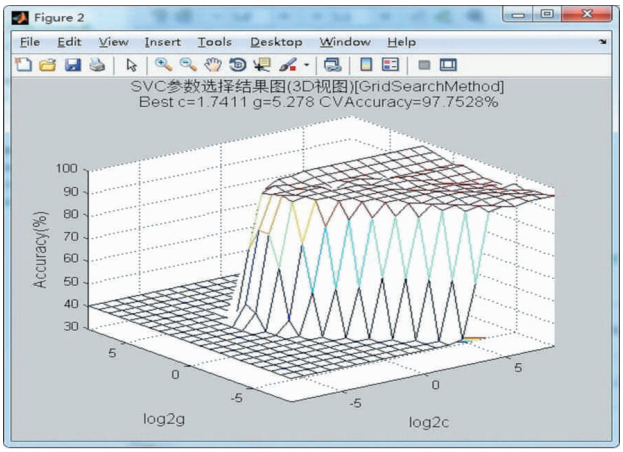

图6 寻优效果3D图

2.2.3 创建模型和模型模型评测

通过上述数据准备、数据预处理、参数寻优、创建模型以及验证等步骤,得出比较符合预期设想的分类准确率。

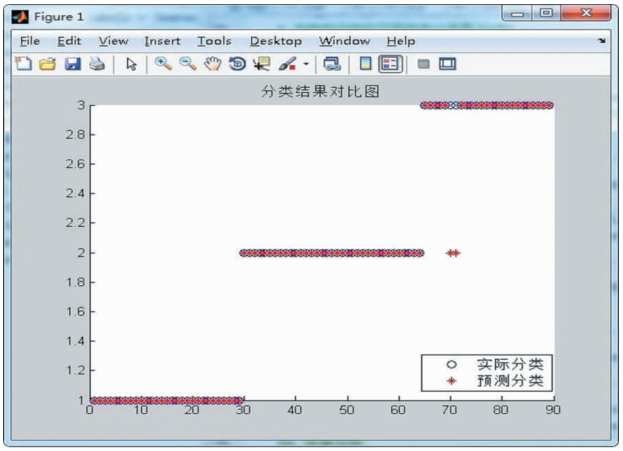

图7 分类结果对比图

从分类结果对比图中可以看出,测试集89个测试数据中有两个测试数据预测分类不准确。经查验原本属于3类的两个样本数据,测试时误将其分类成2类。此模型分类准确率达到0.977528。

3 SVM、RBF和BP分类模型准确率对比

通过表1可以看出不同的数据预处理,在支持向量机模型中得出的分类准确率不同,通过表2可以看出利用支持向量机和利用RBF、BP神经网络建立分类模型的分类准确率的不同。

表1 支持向量机不同预处理分类准确率

序号	方法	合并归一	分开归一	降维	分类准确率
1	支持向量机	是	否	是	0.977528
2	支持向量机	是	否	否	0.906292
3	支持向量机	否	是	是	0.876773
4	支持向量机	否	是	否	0.834532

表2 各种分类模型对比

序号	方法	分类准确率
1	支持向量机	0.977528
2	BP神经网络	0.910112
3	RBF神经网络	0.898876

通过表1可以看出,对于支持向量机来说,必要的数据预处理可以提高其模型的分类准确率。通过表2可以看出对于支持向量机、BP神经网络和RBF神经网络来说,由于支持向量机采用了适当的数据预处理和优化,支持向量机分类模型其分类准确率更高。

4 不足之处

文章所进行的研究探索以及实践工作尚处于初步阶段,在很多方面还有待进一步完善:

(1)文章中的数据集中的各类样本数目基本均衡,即各类所含的样本数大致相当,而且测试集样本和训练集样本数量基本相同。即文章中所做出的模型是基于均衡数据集的情况下得出的,这种建立模型的方法在某些场合下并不适合。而且在现实生活中还有很多不均衡数据集的情况存在,所以今后还要对不均衡数据集进行研究。

(2)在支持向量机方法中,直接影响支持向量机方法的性能是核函数选择的好坏,因此核函数的选取非常关键。核函数及其参数的确定,完全依赖于使用者的经验或者是通过实验在一定范围内进行最优选择的,是因为目前对于核函数及其参数的确定,尚没有一个明确的方法指导。因此,支持向量机应用中一个有待研究的问题是如何结合具体的应用选择最优核函数及参数的取值。对于最终核函数的合理选取和确定,需进一步深入研究以提高其算法的实用性。

注释及参考文献:

[1] 储岳中,徐波.基于流行分析与AP算法RBF神经网络分类器[J].华中科技大学学报,2012(8):98-102.

[2] 陈诚.基于GA、BP神经网络和多元回归的集成算法研究[J].计算技术与自动化,2011(2):91-97.

[3] 胡新和.基于BP神经网络自适应控制系统的改进与优化[J].船电技术,2011(5):50-56.

[4] 梅玲.支持向量机模型的相关研究[D]. 开封:河南大学,2011.

[5] 奉和国.SVM分类核函数及参数选择比较 [J].计算机工程与应用, 2011(3):127-128.

Research on the Application of Supportive Vector Machine in the Classification of Water Quality

LIU Bao

(The Computer and Internet Work Center, Huainan Vocational Technical College, Huainan, Anhui 232001)

Abstract: My paper intends to build a model based on the application of artificial neural networks such as BP, RBF and non-linear method such as supportive vector machine in classifying the data on the same water quality. In such a process, using supportive vector machine, adopted radial basic function (RBF), methodologies such as normalization, dimension reduction, and grid search algorithm to get optimization out of relevant parameter to classify the water quality. the results of my experiment suggest that among non-linear methods, combining the use of supportive vector machine with the relevant pre-processing data methods has proved more accurate in the classification, thus making it worth further promotion.

Key words: assessment of water quality; classification; supportive vector machine; artificial neural networks; radial basis function

(上接第41页)

parts, combined with the actual geometric features of parts, compared with each method of drawing the characteristics, providing design ideas of 3D part drawing.

Key words: Auto CAD; 3D part drawing; coordinate system; the rotary geometries; the square geometries