SVM分类核函数及参数选择比较

奉国和
FENG Guohe

华南师范大学 经济管理学院 信息管理系，广州 510006
School of Economy & Manangement,South China Normal University,Guangzhou 510006,China
E-mail:ghfeng@163.com

FENG Guohe.Parameter optimizing for Support Vector Machines classification.Computer Engineering and Applications, 2011,47(3):123-124.

Abstract: Support Vector Machine(SVM) has good performance for classification,but the performance is restricted by the kernel function and its parameters.This paper discusses the problem,and uses cross validation,grid searching for optimizing the kernel function parameters.

Key words: Support Vector Machines(SVM);kernel function;classification

摘 要: 支持向量机(SVM)被证实在分类领域性能良好,但其分类性能受到核函数及参数影响。讨论核函数及参数对SVM分类性能的影响,并运用交叉验证与网格搜索法进行参数优化选择,为SVM分类核函数及参数选择提供借鉴。

关键词: 支持向量机; 核函数; 分类
DOI:10.3778/j.issn.1002-8331.2011.03.037 **文章编号:**1002-8331(2011)03-0123-02 **文献标识码:**A **中图分类号:**TP316

分类是数据挖掘的一项重要应用。分类方法很多,但研究表明支持向量机(SVM)的分类性能,尤其是泛化能力好于传统的分类方法[1]。SVM基于结构风险最小化原理,求解化为一个线性约束的凸二次规划(Quadratic Programming,QP)问题,解具有唯一性和全局最优性。SVM的分类性能受到诸多因素影响,其中下面两个因素较为关键:(1)误差惩罚参数C,对错分样本比例和算法复杂度折衷,即在确定的特征子空间中调节学习机器置信范围和经验风险比例,使学习机器的推广能力最好。其选取由具体的问题而定,并取决于数据中噪声的数量。在确定的特征子空间中C的取值小表示对经验误差的惩罚小,学习机器的复杂度小而经验风险值较大;C取无穷大,则所有的约束条件都必须满足,这意味着训练样本必须准确地分类。每个特征子空间至少存在一个合适的C使得SVM推广能力最好。当C超过一定值时,SVM的复杂度达到了特征子空间允许的最大值,此时经验风险和推广能力几乎不再变化。(2)核函数形式及其参数,不同核函数对分类性能有影响,相同核函数不同参数也有影响。在多项式核函数中参数是d,RBF核函数数是σ,在Sigmoid核函数参数是d[2]。如何选择核函数及参数直接影响到SVM分类好坏,本文研究核函数及参数对SVM分类性能影响,对于推广SVM应用有启示作用。

1 支持向量机原理

支持向量机是基于结构风险最小化原理(Structural Risk Mininization,SRM),为了控制泛化能力,需要控制两个因素,即经验风险和置信范围值。传统的神经网络是基于经验风险最小化原则,以训练误差最小化为优化目标,而支持向量机以训练误差作为优化问题的约束条件,以置信范围最小化为优化目标。它最终化为解决一个线性约束的凸二次规划(QP)求解问题,所以支持向量机的解具有唯一性,也是全局最优的[3-4]。应用核函数技术,将输入空间中的非线性问题,通过函数映射到高维特征空间中,在高维空间中构造线性判别函数,常用的核函数有如下三种。

(1)Polynomial核函数:$K(x,x_i)=[\gamma*(x\cdot x_i)+coef]^d$,其中$d$为多项式的阶,$coef$为偏置系数。

(2)RBF核函数:$K(x,x_i)=\exp(-\gamma*\|x-x_i\|^2)$,其中$\gamma$为核函数的宽度。

(3)Sigmoid核函数(两层神经网络):$K(x,x_i)=\tanh(\gamma(x\cdot x_i)+coef)$。

2 参数选取方法

SVM的参数选择问题,其实质就是一个优化问题。目前SVM参数选取方法主要有:经验选择法、实验试凑法、梯度下降法、交叉验证法、Bayesian法等。同时随着遗传算法、粒子群优化、人工免疫等智能优化方法的成功,陆续有学者采用这些方法来优化选择SVM参数。近几年,进化计算领域兴起了一类新型优化算法,即分布估计算法,并迅速成为进化计算领域

基金项目:国家社科基金(the National Social Science Foundation of china under Grant No.08CTQ003);广东省哲学社会科学规划项目(No.06M03)。
作者简介:奉国和(1971—),男,博士,副教授,研究方向:数据挖掘、数字图书馆。
收稿日期:2009-09-04 **修回日期:**2009-11-05

的研究热点和解决工程问题的有效方法,也有学者将该方法引入到SVM参数选择问题中来。交叉验证是目前应用较为普遍的一种方法,该方法易于实现,但是计算量大,尤其对于大样本问题。运用支持向量机进行参数优化要根据实际问题具体解决,本文主要阐述交叉验证法与网格搜索法进行SVM参数选择。

(1)K交叉验证(k-Cross Validation)

将训练数据集分成k份相等的子集,每次将其中$k-1$份数据作为训练数据,而将另外一份数据作为测试数据。这样重复k次,根据k次迭代后得到的MSE平均值来估计期望泛化误差,最后选择一组最优的参数。留一法是k-交叉验证的特例,即每次用$n-1$个数据(n为训练数据集大小)训练,而用另一个数据测试。

(2)网格搜索法(Grid Search)

比如在用SVM分类时,采用RBF核函数,此时需要确定两个参数即惩罚因子C与核函数参数σ。基于网格法将$C\in[C_1,C_2]$,变化步长为C_s,而$\sigma\in[\sigma_1,\sigma_2]$,变化步长为$\sigma_s$。这样,针对每对参数$(c',\sigma')$进行训练,取效果最好的一对参数作为模型参数。网格搜索看似愚蠢但直接,利用网格搜索有两个主要优点:一是心理上的,通常的启发式或逼近方法由于不是穷尽搜索而感觉不很保险;二是运用网格搜索法寻找好参数所需时间并不比其他高级方法多,因为其基本涉及两个参数;三是网格搜索算法可以平行进行计算,节省时间开销[5]。

3 实验分析

3.1 实验步骤

实验基于开源软件LIBSVM与Matlab平台进行,LIBSVM是台湾大学林智仁教授等开发设计的一个简单、易于使用和快速有效的SVM模式识别与回归的软件包。该软件不仅提供编译好的可在Windows系列系统的执行文件,还提供了源代码,方便改进、修改以及在其他操作系统上应用。软件具有如下特点:(1)支持不同的支持向量机算法。(2)高效的多分类。(3)交叉验证的模型选择。(4)概率估计。(5)支持不平衡数据的权SVM。(6)同时提供C++与Java源代码。(7)提供用户图形界面的的分类与回归示范。(8)提供与Matlab、Python、R(also Splus)、Ruby、Weka、Common LISP、LabVIEW、C#.NET等接口。通过该软件可以解决C-SVM分类、ν-SVM分类、ν-SVM回归和ε-SVM回归等问题,也包括对一类问题的分布估计[6]。实验利用LibSVM与Matlab的接口,在Matlab环境下进行。实验步骤如下:

(1)对数据预处理(如缺失值的处理,简单办法是直接将这些记录删除,文中即采用该方法)。

(2)对数据进行归一化处理,减少大属性值对小属性值的影响,同时降低数值计算困难。

(3)数据分析基于LIBSVM开源软件,所以还必须将数据转化为LIBSVM格式。

(4)确定一些优化方法对参数进行寻优计算(交叉验证与网格法组合法,程序代码见表1)。

①依据网格法初步设定参数变化范围,针对参数不同组合运用交叉验证求得分类正确率。

②根据前述参数范围进一步细分网格,得到更精确的参数值,根据交叉验证平均正确率排序,选择分类正确率最高的参数组合作为模型的最优参数。

③将数据重新分成训练集与测试集,利用最优化参数模型训练模型,利用测试数据测试模型性能。

(5)结束。

```
function svm_cross(infile,n)
%infile is the input data which is according with the libsvm data format.
%n is cross validation set numbers.
[label_vector,instance_matrix]=libsvmread(infile);
bestcv=0;bestc=0;bestg=0;
for lb c=-5:12,
  for lb g=-5:5
    for d=1:6
      cmd=['-s 0 -t 2 -v 5 -c',num2str(2^lb c),'-g', num2str(2^lb g),'-d',num2str(d)];
      cv=svmtrain(label_vector,instance_matrix,cmd);
      if(cv>=bestcv),
        bestcv=cv;bestc=2^lb c;bestg=2^lb g;bestd=d;
      end
      fprintf('%g %g %g %g(best c=%g,g=%g,d=%g,rate=%g)\n',lb c,lb g,d,cv,bestc,bestg,bestd,bestcv);
    end
end
end
```

图1 交叉验证网格优选参数Matlab程序

3.2 实验结果

(1)数据集1

数据集1-Wisconsin Diagnostic Breast Cancer Data Set(可在http://archive.ics.uci.edu/ml/下载),来自威斯康星州医院实际乳腺肿瘤数字化图像数据,总共包含10个特征,699个数据,数据分良性、恶性两类,其中16个数据存在属性缺失值现象。数据分析结果见表1,表1中核函数及参数一栏所列参数值是各核函数最优参数值,正确率是通过交叉验证得到的平均分类正确率。

表1 各种核函数对Breast Cancer最优参数表

	Polynomial(缺省$coef$=0)(c=0.125,g=0.125,d=1)	RBF(c=0.125,g=0.125)	Sigmoid(c=1,g=0.062 5)
交叉验证平均正确率/(%)	97.511 0	97.218 2	96.925 3

将数据集Wisconsin Diagnostic Breast Cancer Data Set分成两部分,其中500个样本数据作为训练集,另外183个数据作为测试集。运用上面得到的最优化参数训练模型,结果如表2。

表2 各种核函数对Breast Cancer分类性能比较

	Polynomial	RBF	Sigmoid
训练集正确率/(%)	95.8	96.400 0	96.2
测试集正确率/(%)	100.0	99.453 6	100.0

(2)数据集2

数据集2-Heart Disease样本数据集包含2个类别,总共有270个样本数据,属性13个(http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#heart)。该数据集是心脏病数据。根据前述实验方法,计算得到优化参数如表3。

(下转128页)

延长左右两端的数据长度,待EMD分解完成后,再截去数据两端端点效应最为明显的部分。以周期谐波函数为例,通过数值实验,将通过数学的方式拟合延长两端数据的方法(以极值点的镜像延拓为例)与直接截取两端数据的方法分别和理想分解结果进行比较。将实际的EMD分解结果与理想分解结果分别作为矩阵,计算它们之间的相关系数来衡量EMD的分解效果。并得出以下结论:

(1)通过延长采样数据的方法进行端点延拓,同样可以起到抑制端点效应的作用;

(2)若左右截取半个信号周期长度的数据信号,则得到的分解结果优于通过端点延拓方法得到的EMD分解结果,且截取的点数越多,得到的结果越接近理想的分解结果。

但是延长采样数据,本身也存在一定的局限性:由于EMD本身是一个迭代过程,在此过程中数据长度的增加必然导致计算量的增加;同时由于在IMF筛分停止条件中,端点处的异常情况会影响IMF的筛分次数。因此延长采样数据,进行EMD后,再截取中间部分的信号,也会存在端点处的误差由端点处向内逐渐传播,同样会产生端点效应,影响IMF的分解质量。

参考文献:

[1] 刘慧婷,张旻,程家兴.基于多项式拟合算法的EMD端点问题的处理[J].计算机工程与应用,2004,40(16):84-86.

[2] Huang N E,Shen Z,Long S R.The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis[C]//Proc Conference on Mathematical,Physical and Engineering Sciences.London:The Royal Society,1998:903-995.

[3] Rilling G,Flandrin P.On the influence of sampling on the empirical mode decomposition[C]//Proc 2006 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP 2006),Toulouse,France,2006,3:444-447.

[4] 邓拥军,王伟,钱成春,等.EMD方法及Hilbert变换中边界问题的处理[J].科学通报,2001,46(3):257-263.

[5] 黄大吉,赵进平,苏纪兰,等.黄变换的端点延拓[J].海洋学报,2003,25(1):1-2.

[6] Chen Q H,Huangn E,Riemenschneider S,et al.A B spline approach for empirical mode decompositions[J].Advances in Computational Mathematics,2006(24):171-195.

[7] 黎洪生,吴小娟,葛源.EMD信号分析方法端点问题的处理[J].电力自动化设备,2005,25(9):47-49.

[8] 郑天翔 杨力华.经验模式分解算法的探讨和改进[J].中山大学学报:自然科学版,2007,46:1-6.

[9] 胡维平,莫家铃,龚英姬,等.经验模态分解中多种边界处理方法的比较研究[J].电子与信息学报,2007,29(6).

[10] 朱金龙,邱晓晖.正交多项式拟合在EMD算法端点问题中的应用[J].计算机工程与应用,2006,42(23):72-74.

[11] 许宝杰,张建民,徐小力,等.抑制EMD端点效应方法的研究[J].北京理工大学学报,2006,26(3):96-200.

[12] 陈忠,郑时雄.EMD信号分析方法边缘效应的分析[J].数据采集与理,2003,18(3):114-118.

[13] 李舜酩,李香莲.振动信号的现代分析技术与应用[M].北京:国防工业出版社,2008.

[14] 胡广书,数字信号处理—理论、算法与实现[M].北京:清华大学出版社,1997.

(上接124页)

表3 各种核函数对Heart Disease最优参数表

	Polynomial(c=0.062 5, g=0.25,d=1,$coef$=0)	RBF(c=256, g=0.000 976 563)	Sigmoid(c=512, g=0.015 625)
正确率/(%)	81.851 9	84.074 1	85.555 6

将数据集Heart Disease Data Set分成两部分,其中200个样本数据作为训练集,另外70个数据作为测试集。运用上面得到的最优化参数训练模型,结果如表4。

表4 各种核函数对Heart Disease分类性能比较

	Polynomial	RBF	Sigmoid
训练集正确率/(%)	85.000 0	87.500 0	84.500 0
测试集正确率/(%)	82.857 1	84.285 7	84.285 7

3.4 实验结论

实验表明:结合交叉验证与网格搜索法进行SVM核函数参数寻优是有效的。三种核函数中,RBF核函数表现相对稳定,而Polynomial核函数与Sigmoid核函数稳定性要差。运用支持向量机分类时,可优先考虑RBF核函数。

4 结语

支持向量机已经广泛应用在分类领域,取得了良好的效果,但其分类性能受到核函数及参数影响。讨论了核函数及参数对SVM分类性能的影响,并运用交叉验证与网格搜索法进行参数优化选择,为SVM的应用提供借鉴。

参考文献:

[1] Sholkopf B,Sung K,Burges C J C,et al.Comparing support vector machine with Gaussian Kernels to radial basis function classifiers[J].IEEE Trans,Signal Processing,1997,45:2758-2765.

[2] Burges C J C.A tutorial on support vector machines for pattern recognition[J].Data Mining and Knowledge Discovery,1998(2):121-167.

[3] 奉国和.基于聚类的大样本支持向量研究[J].计算机科学,2006,33(4):145-147.

[4] Vapnik V N.The nature of statistical learning theory[M].New York:Springer,1999.

[5] Hsu C W.A practical guide to support vector classification[EB/OL].[2009-06-20].http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[6] LIBSVM—A library for support vector machines[EB/OL].[2009-06-07].http://www.csie.ntu.edu.tw/~cjlin/libsvm/.