基于感知器的SVM自学习模型

宋营军,张化祥
SONG Ying-jun,ZHANG Hua-xiang

山东师范大学 信息科学与工程学院,济南 250014
College of Information Science and Engineering,Shandong Normal University,Jinan 250014,China
E-mail:syj089@163.com

SONG Ying-jun,ZHANG Hua-xiang.SVM self-learning model based on perceptron.Computer Engineering and Applications,2010,46(7):57-58.

Abstract: This paper proposes a kind of SVM classification,which is based on perceptron.This model in the classifier of training,has introduced the perceptron study thought,uses the support vector machines nuclear function to do nuclear calculation,then judges the classification of performance,classifying correctly does not make any revision,on the contrary,transforms to the perceptron study question.Experiments show that the model can not only improve the performance of SVM classification,but also can reduce the SVM classification performance to the nuclear function and parameters choice dependence.

Key words: Support Vector Machine(SVM);Kernel function;perceptron;kernel parameter select

摘 要:提出了一种基于感知器的SVM分类模型(PSVM)。该模型在对分类器的训练中,引入感知器分类思想,其先利用SVM的核函数进行核计算,判断其分类性能,分类正确则不作任何修改,反之则转化成感知器分类问题。实验结果表明该模型不但能提高SVM的分类性能,而且还可以降低SVM分类性能对核函数及参数选择的依赖。

关键词:支持向量机;核函数;感知器;核参数选取
DOI:10.3778/j.issn.1002-8331.2010.07.017 **文章编号:**1002-8331(2010)07-0057-02 **文献标识码:**A **中图分类号:**TP391

1 引言

20世纪90年代VaPnik[1]基于统计学理论提出了一种新的机器学习方法支持向量机(Support Vector Machine,SVM),该方法是建立在结构风险最小化原则基础之上的,其核心思想是引入核函数技巧,把低维空间线性不可分问题,映射成高维空间线性可分问题,它能较好地解决非线性、高维识别、小样本和局部极小点等问题。SVM的发展,不但丰富和发展了统计学理论,而且在很多应用领域得到推广和应用,如:文本分类[2],手写体识别[3],人脸识别[4],WEB挖掘,回归分析[5]等。如何进一步改进支持向量机的性能,一直以来都是模式识别和机器学习领域关注和研究的热点。对此已有一些学者在不同方面提出了一些相应的改进思想,如在SVM核函数参数的选择方面,周志华等,在最小核距离分类中,通过引入最优化目标函数来学习核参数[6];在SVM核函数的构造方面,张冰等根据不同核函数的优缺点,把全局核函数和局部核函数相加,形成新的核函数[7];在SVM的集成学习方面,张妤等提出了一种将支持向量机进行选择集成回归的方法[8],以上这些改进支持向量机的方法都已取得了很好的分类效果,但却忽视了对SVM错误分类数据的"二次"再利用。

感知器的训练算法的基本原理来源于著名的Hebb学习律,其基本思想是:依次将样本集中的训练数据输入到感知器网络中,根据实际输出结果和理想输出之间的差距来调整感知器相应的权重(即各属性对感知器正确分类的贡献度),直至达到理想的效果。感知器学习法则很好地利用了错分类的数据信息,来提高分类器的性能,但是感知器学习法则却存在只能对线性可分类数据进行学习的缺点。

针对分类错误数据的再利用问题,结合支持向量机对线性不可分数据处理和感知器针对分类错误数据学习的优点,提出了一种基于感知器的支持向量机分类模型,该模型有效地利用了错分数据的信息来提高分类器性能。

2 支持向量机理论基础

给定训练样本集$D=\{(x_i,y_i),i=1,2,\cdots,l\}$,其中,$x_i\in R^n$,$y_i\in\{1,-1\}$为类别标识符(这里仅考虑两类问题,多类问题可以转化为多个两类问题进行解决)。SVM存在线性可分和线性不可分两种情况,其判断依据是:

基金项目:山东省中青年科学家科研奖励基金(博士基金)(No.2006BS01020);山东省高新技术自主创新工程专项计划(No.2007ZZ17);山东省自然科学基金(the Natural Science Foundation of Shandong Province of China under Grant No.Y2007G16);山东省科技攻关计划(the Key Technologies R&D Program of Shandong Province,China under Grant No.2008GG10001015));山东省教育厅科技计划项目(No.J07YJ04)。

作者简介:宋营军(1983-),男,硕士在读,主要研究方向:数据挖掘,机器学习,模式识别;张化祥(1966-),男,博士,教授,主要研究方向:数据挖掘,模式识别,机器学习,人工智能及Web挖掘。

收稿日期:2008-10-23 **修回日期:**2008-12-26

若样本集 $D=\{(x_i,y_i),i=1,2,\cdots,l\}$,其中,$x_i\in R^n$,$y_i\in\{1,-1\}$,存在 $w\in R^n$,$\rho>0$,$b\in R$,使得式(1)成立。

$$y_i(w^{\mathrm{T}}x_i+b)-\rho\geqslant 0\quad i=1,2,\cdots,l \tag{1}$$

则称样本集 D 线性可分,反之,则称样本集 D 线性不可分。

对线性可分得情况,先在特征空间中寻找一个最优的分类平面 $w_0^{\mathrm{T}}\phi(x)+b_0=0$,要求两分类样本距分类面的最小间隔尽可能的大,对线性不可分的情况,SVM 先通过一个映射,将样本空间映射到高维特征空间,转化为线性可分的情况再进行求解。SVM 分类算法要求解下面有约束的优化问题:

$$\min_{w,b}\phi(w)=\frac{1}{2}w^{\mathrm{T}}w+C\sum_{i=1}^{l}\xi_i \tag{2}$$

其中:$y_i(w^{\mathrm{T}}\phi(x_i)+b)\geqslant 1-\xi_i$ (3)

$$\xi_i\geqslant 0\quad i=1,2,\cdots,l \tag{4}$$

引入 Lagrange 乘子转化为其相应的对偶问题:

$$\min_{a}w(a)=\sum_{i,j}a_ia_j,y_iy_jK(x_i,x_j)-\sum_j a_j \tag{5}$$

$$\sum_i y_ia_i=0,0\leqslant a_i\leqslant C\quad i=1,2,\cdots,l \tag{6}$$

其中 C 是预先设定的常数,称为惩罚因子,a_i 是 Lagrange 乘子,$K(x_i,x_j)\equiv\phi^{\mathrm{T}}(x_i)\phi(x_j)$是满足 Mercer 条件的核函数。

则 SVM 分类算法的决策函数是:

$$f(x)=\sum_{i=1}^{l}y_ia_iK(x_i,x)+b \tag{7}$$

令 $v_i=y_ia_i$,则方程(7)为:

$$f(x)=\sum_{i=1}^{l}v_iK(x_i,x)+b \tag{8}$$

$$sign[f(x)]=\begin{cases}1,f(x)\geqslant 0\\-1,f(x)<0\end{cases} \tag{9}$$

3 PSVM 算法思想

3.1 PSVM 基本思想和策略

支持向量机能很好地解决线性不可分的问题,并且能使结果达到全局最优,但 SVM 在对数据二次处理中,忽视了被错分类样例的信息,相反感知器分类恰恰是利用错分类样例来调整自己的权重(各属性对正确分类的贡献度),以求达到较好的分类效果。把 SVM 对线性不可分数据处理和感知器分类对错误数据的学习方法相结合,在 SVM 分类的基础上,感知器利用 SVM 分类错误的样例来重复修正其权重,以求达到更好的分类效果。

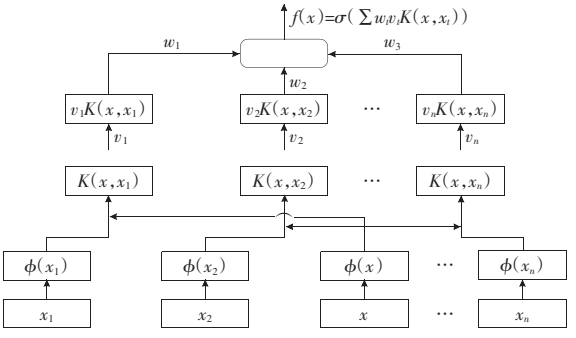

图 1 PSVM 流程图

3.2 支持向量感知器算法步骤

(1)选择核函数 $K(x,y)$,并初始化参数 w_i,v_i,C

(2)for 训练样本集 D 中每条记录

(3) 将其映射到相应的核空间 R'

(4) 计算其内积,并用决策函数(8)进行判断

(5) If 分类结果正确

(6) w_i不做任何修改,Return TURE;

(7) else

(8) 修改权重 w_i,并保证 $\sum_{i=1}^{l}w_i=1$;

(9)Until 条件满足 Break

4 实验

采用 UCI 机器学习数据库[9]中的 5 个数据集对上述 PSVM 算法进行了实验。所用的测试模式是十折交叉法,感知器初始权重都设为 $w_i=1/l$,其中 l 代表样本集数据个数,degree 代表多项式核函数的维数。这 5 个数据集的具体信息见表 1。

表 1 实验所用数据集

Data Set	No. Attributes	No. Class	No. Instances
Iris	4	3	150
segment	20	7	2 310
anneal	39	7	898
segment-test	20	7	810
Glass	10	7	214

实验中,将提出的 PSVM 和 SVM 算法相比较(对同一数据集,保证 SVM 和 PSVM 的参数设置一致)。由表 2,3,4 中 SVM 和 PSVM 对各数据集的分类正确率可以看出 PSVM 模型的分类性能相对于 SVM 要好,特别是对 SVM 分类能力较差的数据集。由表 2 和表 4 中选择不同的核函数后 SVM 和 PSVM 对各数据集的分类正确率对比,可以看出核函数的选取对 PSVM 的分类正确率影响不大,由表 2 和表 3 中同一核函数选不同参数情况下 SVM 和 PSVM 的分类正确率对比,可以看出 PSVM 对核函数及参数选择的依赖性有所降低。因此该模型不但能很好地提高 SVM 的分类性能,而且还可以降低 SVM 分类性能对核参数和核函数选择的依赖性。

表 2 在 degree=3 时,SVM 和 PSVM 的多项式核函数分类正确率 (%)

	Iris	segment	anneal	Glass	Segment-test
SVM	96.667	84.672	94.382	66.355	95.185
PSVM	95.012	85.365	95.107	71.253	95.192

表 3 在 degree=5 时,SVM 和 PSVM 的多项式核函数分类正确率 (%)

	Iris	segment	anneal	Glass	Segment-test
SVM	95.333	86.368	92.089	69.626	94.567
PSVM	95.365	84.239	93.107	75.426	95.089

表 4 在核函数选择 RBF 核时 SVM 和 PSVM 的分类正确率 (%)

	Iris	segment	anneal	Glass	Segment-test
SVM	96.667	65.368	90.089	68.692	41.111
PSVM	95.326	76.362	92.107	74.253	79.231

但该模型也存在不足,如有可能出现过拟合现象。目前,对 SVM 的研究主要针对统计学理论和各种应用领域的研究,而

(下转 61 页)