一种支持向量机集成的核选择方法

王 敏,王文剑
WANG Min,WANG Wen-jian

山西大学 计算机与信息技术学院,计算智能与中文信息处理教育部重点实验室,太原 030006
School of Computer and Information Technology,Key Lab of Computational Intelligence and Chinese Information Processing of Ministry of Education,Shanxi University,Taiyuan 030006,China
E-mail:hanxiao-wm@163.com

WANG Min,WANG Wen-jian.Approach for kernel selection from SVM ensemble.Computer Engineering and Applications,2009,45(27):31-33.

Abstract: Kernel selection is an important issue in Support Vector Machine(SVM) modeling.SVM is a popular machine learning tool with good generalization ability,but its performance is often dependent on selected kernel function.For a given problem,it is difficult to choose an appropriate kernel function.This paper proposes an approach for kernel selection based on SVM ensemble.Some basic SVMs are constructed by adopting different kernel function or parameters and then the final prediction is obtained through aggregating the results of these basic SVMs.The proposed algorithm will integrate kernel selection with SVM learning.In so doing,not only the influence of kernel selection on a single SVM can be avoided,but also good generalization performance can be obtained.Simulation results on UCI benchmark datasets demonstrate the validity of the proposed approach.
Key words: Support Vector Machine(SVM);ensemble learning;kernel selection;heterogeneity SVM;homogeneity SVM

摘 要: 核选择问题是支持向量机(Support Vector Machine,SVM)建模中的一个关键问题,虽然支持向量机具有良好的泛化性能,但其性能受核函数的影响比较明显,而对于一个给定问题,选择合适的核函数及参数通常很困难。提出一种基于SVM集成的核选择方法,利用不同的核函数构造子SVM学习器,然后对子学习器的预测结果集成。提出的核选择方法将SVM集成学习与核选择同时进行,不仅避免了单个SVM的核选择对泛化能力的影响,而且可以获得良好的泛化能力。在UCI标准数据集上的结果说明了提出的方法的有效性。
关键词: 支持向量机;集成学习;核选择;异质SVM;同质SVM
DOI: 10.3778/j.issn.1002-8331.2009.27.010 **文章编号:** 1002-8331(2009)27-0031-03 **文献标识码:** A **中图分类号:** TP18

1 引言

由Vapnik等人提出的支持向量机[1]是一种通用有效的机器学习方法,已成为当今机器学习领域的一个研究热点,可以用于解决非线性分类、非线性回归及概率密度估计等问题,目前已经在许多智能信息获取与处理领域取得成功的应用[2-3]。然而在基于SVM的应用中还存在一些问题如虽然SVM具有良好的泛化性能,但其性能受核函数及参数的影响比较大,对于一个给定的问题, 要选择合适的核函数及参数通常很困难,这使得核选择成为支持向量机研究的核心问题之一,同时这也是设计SVM学习器时首先要面临的问题。目前,关于SVM的核选择方法研究已取得了一些成果[4-5],但这些方法或是时间复杂度较高,或是算法较为复杂,或是稳定性不强[6],而且对核选择方法的研究主要是面向单个SVM的。鉴于集成学习(Ensemble Learning)在机器学习领域的成功应用,该文将集成学习技术引入SVM中来解决核选择问题。

集成学习是机器学习的重要研究方向之一,它将一系列子学习器集成起来共同解决同一问题,并利用子学习器的差异性显著提高学习系统的泛化能力[7]。在集成学习中,子学习器的多样性是评价集成学习算法的重要标准[8],如何获取不同的子学习器对集成学习的效果有着重要影响,常用的子学习器的获取方式有:对训练数据进行处理如比较流行的bagging和boosting方法;对输入特征进行处理等。目前集成学习在神经网络、决策

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60673095);国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z165);教育部科学技术研究重点项目(the Key Project of Science Technology Research of Ministry of Education No.208021);国家教育部新世纪人才支持计划(the New Century Excellent Talent Foundation from MOE of China under Grant No.NCET-07-0525);山西省青年学术带头人支持计划(the Program for the Top Young Academic Leaders of Higher Learning Institutions);山西省留学归国人员基金(the Project for Returned Overseas of Shanxi Province No.2008-14)。
作者简介: 王敏(1985-),男,硕士研究生,主要研究方向:机器学习;王文剑(1968-),女,博士,教授,博士生导师,主要研究方向:机器学习,计算智能等。
收稿日期: 2009-03-09 **修回日期:** 2009-05-06

树等方面已经取得了显著的研究成果。由于SVM具有较强的理论背景,简洁的数学形式及良好的泛化能力,国内外学者也将集成学习技术与SVM的研究结合在一起,如比较Bagging和Boosting方法对支持向量机泛化能力的影响[9],利用集成学习的思想解决SVM的多分类问题[10],通过对训练数据进行处理构造SVM集成学习器[11-12],将经典集成算法应用于SVM来解决实际问题等[13-14]。

对于集成学习来讲,当子学习模型具有较高的正确率且具有显著的差异性时,集成学习系统的泛化能力将明显提高。而SVM可以通过不同的核选择改变自身的结构,使子学习器具有多样性,从而满足集成学习的基本要求。提出一种基于SVM集成的核选择方法,将模型选择与SVM学习同时进行,而不像通常的SVM学习方法,即先进行核选择后进行训练,从而可以避免直接解决SVM的核选择问题,同时,该方法也具备集成学习的优点,有望获得良好的泛化能力。在UCI标准数据集上的实验说明,提出的SVM集成学习方法可以很好地处理SVM的核选择问题。

2 基于SVM集成的核选择方法

2.1 基本原理

支持向量机通过非线性映射Φ将输入空间映射到一个高维特征空间,然后在这个高维特征空间中构造最优超平面,使训练集中的点距离该超平面尽可能远。核函数是输入空间到高维特征空间映射Φ的内积,不同的核函数表明输入空间到高维空间的核映射不同,产生的支持向量(Support Vector,SV)和分类超平面也不同。集成学习是一种通过子学习器之间的差异性来提高泛化能力的学习方法,其有效的条件是子学习器的错误率不能高于50%,且相互之间要有较大的差异性。虽然集成学习对子学习器的精度要求并不严格,但在保证各子学习器之间具较大的差异性的情况下,集成学习器的精度往往高于各子学习器。对于SVM的核选择问题,可以利用集成学习的这一特点,不必要求子SVM的核选择效果好,只要使核选择具有较大的差异性就可以获得满意的学习结果。一般支持向量机具有良好的泛化性能,可以满足集成学习有效的第一个条件。此外,对支持向量机来讲,不同的核选择表明SVM产生的分类超平面不同,通过核函数及参数的改变可以使子SVM之间产生较大的差异性,满足集成学习有效的第二个条件。

基于SVM集成的核选择方法的基本思想是通过设置不同的核函数及参数构造具有差异性的子SVM学习器,然后将各子SVM的预测结果集成。在SVM分类器的构造过程中,主要涉及到两个自由参数的选取,分别为核函数Ker和惩罚系数C。不同核函数的SVM称为异质SVM,核函数相同但参数不同的SVM称为同质SVM。在子学习器的构造过程中,可以通过改变核函数和参数的方法产生异质SVM学习器,也可以以最常用的核函数为主,在不同的惩罚系数、核参数区间取值产生同质SVM学习器,对子学习器预测结果的集成可采用简单投票方法。从集成学习的角度来讲,该思想直接改变学习器的构成使其产生较大的差异性,避免了基于改变样本分布的集成学习算法在SVM应用中效果不明显的问题[15],具有很好的泛化能力。另外,值得说明的是,通过集成学习的方法来解决SVM的核选择问题,可有效地避免传统SVM必须先进行核选择再学习的难题,直接将模型选择与SVM的学习同时进行。

2.2 基于核选择的子学习器构造算法

根据集成系统中子学习器构造方法的不同,提出了基于异质SVM的集成学习算法(算法1)和基于同质SVM的集成学习算法(算法2)。异质SVM集成学习算法主要是利用不同的核函数来构造子学习器,由于常用的核函数有限,为了增加集成规模,同一核函数的参数可在常用的参数区间取不同值,也即将同质SVM与异质SVM共同加入集成系统中。算法在实现过程中,将设定的不同核函数及参数保存在核函数向量$\boldsymbol{Kv}$和参数向量$\boldsymbol{Pv}$中,对向量$\boldsymbol{Kv}$中的每一个元素kv_i,其参数值依次设置为$\boldsymbol{Pv}$中的各个元素,然后训练生成子学习器。需要说明的是,线性核函数不必设置核参数,所以不能加入向量$\boldsymbol{Kv}$中,需要单独训练。同质SVM集成学习算法中各子学习器的核函数相同,主要通过在不同的核参数区间取值构造子学习器,为了进一步加大各子学习器之间的多样性,可对子学习器在不同核参数区间的惩罚系数C进行适当调节。算法在实现过程中,将子学习的核参数与惩罚系数成对存入一个参数矩阵$\boldsymbol{ParaM}$中,然后根据每一对参数设置训练生成子学习器。所提出的集成学习算法直接对子学习器进行学习,因此算法的时间复杂度为$O(n)$,其中n为集成规模。

算法 1 基于异质SVM集成的学习算法

步骤 1 给定训练集Tr,初始化核函数向量$\boldsymbol{Kv}=(kv_1,kv_2,\cdots,kv_{n1})$,参数向量$\boldsymbol{Pv}=(p_1,p_2,\cdots,p_{n2})$以及惩罚系数$C$;其中$kv_i$为常用的核函数,$p_i$是核参数$p_1$在不同区间的取值;

步骤 2 对$\boldsymbol{Kv}$中的每一个元素kv_i,令变量$j=1$,然后执行如下操作:

步骤 2.1 如果j不大于$n2$,执行步骤2.2,其中$n2$为向量$\boldsymbol{Pv}$的模;

步骤 2.2 以kv_i为核函数,参数p_1设置为p_j,构造支持向量机学习模型,训练得到子学习器$svm_{n2\times(i-1)+j}$,并在测试集上预测得到子输出$Y_{n2\times(i-1)+j}$;

步骤 2.3 j值加1,执行步骤2.1;

步骤 3 通过线性核函数训练得到子$svm_{n1\times n2+1}$,在测试上预测得到子输出$Y_{n1\times n2+1}$;

步骤 4 计算集成学习的最终输出$Y=\text{sgn}\sum_{l=1}^{n1\times n2+1}Y_l$。

算法 2 基于同质SVM集成的学习算法

步骤 1 给定训练集Tr,初始化核函数Ker,参数矩阵$\boldsymbol{ParaM}=(P,C)$,其中其中P,C分别为核参数p_1和惩罚参数C在常用区间取不同值组成的向量。

步骤 2 对参数矩阵$\boldsymbol{ParaM}$的每个行向量(p_i,c_i)执行如下操作:

将支持向量机参数p_1、C分别设置为p_i,c_i构造支持向量机学习模型,训练得到子学习器svm_i,并在测试集上预测得到子输出Y_i。

步骤 3 计算集成学习的最终输出:$Y=\text{sgn}\sum_{i=1}^{m}Y_i$,$m$为参数矩阵$\boldsymbol{ParaM}$的行数。

3 实验结果与分析

在UCI标准数据集上对所提出的两种集成学习算法的有效性进行验证,实验中用到的数据集见表1。在异质SVM集成学习中,除了算法中提到的线性核Linear,还采用了多项式核

表 1　实验数据集

数据集	样本数	属性	训练集	测试集	损失数据
Haberman	306	3	229	77	0
Breast-cancer	699	10	512	171	16
Bupa	345	6	259	86	0

Poly，高斯径向基核 RBF 和指数型径向基核 ERBF 三种核函数,并对这三种核函数的参数 $P1$ 设置了两个不同的值。在同质 SVM 集成系统中,也分别针对上述四种核函数进行了测试。

首先以 Haberman 数据集为例,验证提出的两个算法的有效性。在异质 SVM 集成算法实验中,核函数向量 $\boldsymbol{Kv}$=($poly$,rbf,$erbf$),核参数向量 $\boldsymbol{Pv}$=(5,7),根据算法 1,则集成系统中子学习器对应的核函数及核参数设置见表 2。集成学习系统中各子学习器的惩罚系数相同,分别设置为 10 和 100,进行了两组实验。

表 2　异质集成子学习器参数设置(Haberman)

参数	svm1	svm2	svm3	svm4	svm5	svm6	svm7
核函数 ker	poly	poly	rbf	rbf	erbf	erbf	linear
核参数 $P1$	5	7	5	7	5	7	

在 Haberman 数据集上的同质 SVM 集成算法实验中,子学习器的核参数 $P1$ 的取值区间是[2,10],并在常用参数区间对惩罚系数 C 做相应调节,由于 Linear 核无核参数,仅通过改变惩罚系数 C 来构造子学习器。各子学习器的具体参数设置见表3。

表 3　同质集成子学习器参数设置(Haberman)

核函数	svm1		svm2		svm3		svm4		svm5		svm6		svm7		svm8		svm9	
	$P1$	C	$P1$	C	$P1$	C	$P1$	C	$P1$	C	$P1$	C	$P1$	C	$P1$	C	$P1$	C
Poly	2	10	2	100	3	100	3.5	100	4	100	5	100	6	100	8	100	8	10
RBF	3	100	5	100	6	100	7	100	7	10	8	100	8	10	9	10	10	10
ERBF	2	10	3	10	4	10	5	10	6	10	7	10	8	10	9	10	10	10
Linear		10		20		30		50		80		100		150		200		300

在 Haberman 数据集上的各组实验中集成学习系统的集成规模和实验结果见表 4。从表中可以看出:在异质 SVM 集成学习实验中，虽然惩罚系数 C 取不同时，集成学习器的性能不同,但在两组实验中集成学习器的正确率均远高于子学习器的平均正确率,当 C 取值为 10 时,集成学习器的正确率与最优个体正确率相同,C 取值为 100 时，集成学习器的正确率高于最优个体的正确率。在四组基于不同核函数的同质 SVM 集成学习实验中,基于高斯径向基核(RBF)和指数型高斯径向基核(ERBF)的集成学习器的正确率也远高于各子学习器的平均正确率,且核函数为 RBF 时,集成学习器的正确率与最优个体的正确率相同,核函数为 ERBF 时,集成学习器的正确率高于最优个体的正确率。核函数为线性核和多项式核时,各子学习器的之间的差异性并不明显,在测试集的中的正确率相同,所以没能体现出集成学习的优点。值得说明是文中集成学习器的集成规模都比较小,但是取得了满意的集成效果。

表 4　Haberman 数据集上实验集成规模及实验结果

核选择	集成规模	集成学习器	最优个体	最差个体	个体平均
异质 C=10	7	0.766 2	0.766 2	0.688 3	0.723 6
异质 C=100	7	0.727 3	0.714 3	0.662 3	0.686 5
同质 Ploy	9	0.688 3	0.688 3	0.688 3	0.688 3
同质 RBF	9	0.766 2	76.62 3	0.623 4	0.721 5
同质 ERBF	9	0.779 2	0.766 2	0.662 3	0.727 3
同质 Linear	9	0.701 3	0.701 3	0.701 3	0.701 3

在 Breast-cancer 和 Bupa 数据集上的异质学习算法实验中,核函数向量 $\boldsymbol{Kv}$ 的设置与在 Harberman 上的设置相同,参数向量 $\boldsymbol{Pv}$ 的设置分别为(3,5)和(3,8),集成规模也都为 7,各子学习器对应的具体核函数及核参数同表 2,惩罚系数 C 为 100。同质集成学习算法也分别基于上述 4 个核函数进行实验,实验中的核参数 $P1$ 的取值区间也都为[2,10],惩罚系数 C 的设置与在 Haberman 数据集上的类似，也在不同的参数区间做简单的调节。表 5 是基于不同核函数的同质集成学习算法的实验集成规模。

表 5　同质集成学习实验集成规模

数据集	Poly	RBF	ERBF	Linear
Breast-cancer	22	17	28	7
Bupa	11	11	11	7

为了更加直观分析实验结果,将 Breast-cancer 和 Bupa 数据集上各组实验中的子学习器错误率,集成错误率,子学习平均错误率进行归一化处理，以最差子学习器的错误率为单位 1,求得其相对误差,实验结果见图 1。

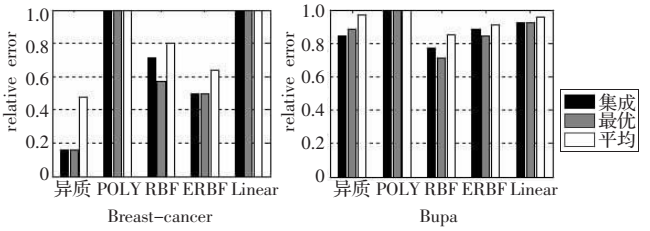

图 1　相对错误率

从以上的实验结果可以看出:基于异质的 SVM 集成算法在不同的数据集上都取得了良好的效果,集成后的性能远优于子学习器的平均性能,且等于或者优于最优个体性能。基于同质 SVM 的集成学习算法则与核函数类型有关，当核函数为 RBF 和 ERBF 时,在三个数据集上取得了满意的结果,集成后的性能接近或优于最优个体的性能。但是核函数为 Ploy 和 Linear 时,有些数据集上集成效果并不明显,其原因主要是由于多项式核与线性核比较简单,改变核参数和惩罚系数对各子学习器间的差异性影响较小。所以对于同质 SVM 集成算法来讲,最好选择比较复杂的核函数。

4　结语

提出了基于 SVM 集成的核选择方法，通过设定不同核函数及参数构造子学习器,并将集成学习器的集成结果作为最终的学习结果。该方法通过较小的集成规模可取得满意的效果,并且实现了模型选择与 SVM 学习的同时进行，而不像通常那样先进行核选择然后对 SVM 学习。提出的方法对核函数及参数的初始设定不需做限定，从而降低了对核选择的依赖性,同时为任一给定问题提供了核选择的一般原则，即基于异质 SVM 集成的核选择算法能够产生具有较大差异性的子学习器,在各个数据集上都能取得很好的效果,基于同质 SVM 集成的核选择算法虽受核函数类型的影响,但是多数情况也可以取得很好的效果。

参考文献:

[1] Vapnik V.Statistical learning theory[M].New York:Wiley,1998.

(下转 55 页)