一种回归SVM选择性集成方法*)

张 妤 王文剑 康向平

(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)

摘 要 泛化能力是机器学习关心的一个根本问题,采用集成学习技术可以有效地提高泛化能力。本文提出了一种将支持向量机(Support Vector Machine, SVM)进行选择性集成回归的方法。通过引入三个阈值,可以选择合适的子SVM,从而进一步提高了整个集成学习的效率。实验结果表明,本文提出的选择性集成方法可以在一定程度上解决SVM的模型选择问题和大规模数据集的学习问题,与传统的集成方法Bagging相比具有更高的泛化能力。

关键词 支持向量机,集成学习,回归,Bagging

A Regression SVM Selection Ensemble Approach

ZHANG Yu WANG Wen-Jian KANG Xiang-Ping

(School of Computer & Information Technology Key Laboratory of Computational Intelligence & Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006)

Abstract Generalization performance is a basic problem for machine learning, and ensemble learning technique can improve the generalization performance effectively. This paper presents a selective ensemble learning approach based on SVM regression. By introducing three threshold parameters, some component SVM can be chose, and the efficiency of the whole ensemble learning system can be further improved. Simulation results demonstrate the presented approach can solve model selection and learning for large-scale dates for SVM. Compared with the traditional ensemble approach like Bagging, the presented approach possesses better generalization performance.

Keywords Support vector machine, Ensemble learning, Regression, Bagging

1 引言

集成学习(Ensemble Learning)技术是利用基学习器的多个版本来解决同一个问题,可以显著地提高学习系统的泛化能力[1]。国际上,对集成技术的研究成为机器学习的一个热点,研究者对集成技术的理论分析、实现方法的设计和实际应用等几个方面,展开了深入研究[2~4],其中以Bagging和Boosting方法影响最大[5,6]。在实现方法的设计上,分为个体生成和个体结合两部分[7]。对于前者的实现方法,从策略上分为两种:一是直接生成子学习器,如最重要的技术Boosting和Bagging就是基于这种策略;二是先产生多个学习器,再从中选择部分学习器进行集成(overproduce and choose)[7]。对于后者的实现方法,分类器主要采用多数投票[8],回归学习器则采用简单平均和加权平均。目前集成学习的研究主要集中在神经网络方面。

支持向量机是一种基于统计学习理论的机器学习方法[9],目前已成功应用于模式识别、回归估计及概率密度估计等领域[10]。然而,SVM的模型选择问题(主要指核及相关参数的选择)和大规模数据的学习问题等影响了SVM在实际应用中的效率。本文提出了一种回归SVM选择性集成方法,将集成学习技术引入SVM回归预测,通过采用特定阈值选择合适的子SVM,从而进一步提高整个SVM的效率。实验结果表明本文提出的选择性集成方法可以在一定程度上解决SVM的模型选择的问题和大规模数据集的学习问题,与传统Bagging集成SVM方法相比具有更高的泛化能力。

2 SVM选择性集成回归方法

集成学习一般包含两个阶段,即个体学习器的生成阶段和个体学习器的结合阶段。本文方法在个体学习器的生成阶段采用的策略是:先产生多个学习器,再从中选择部分学习器进行集成。

令训练集为 $G=\{(x_i,y_i)\}_{i=1}^{l}$,测试集为 $G'=\{(x_i,y_i)\}_{i=1}^{p}$,其中 x_i 为输入,y_i 为输出。首先,利用经典的Bagging方法在训练集 G 上产生 T 个子SVM,然后,将这 T 个子SVM在测试集 G' 上进行回归预测。设第 i、j 两个SVM的预测结果为 Y_i 和 Y_j($Y_i,Y_j\in G'$均为 p 维向量),对于某个样本 $(x_k,y_k)\in G'$,令 $f_k=(Y_{ik}-Y_k)(Y_{jk}-Y_k)$,$Y_{ik}$、$Y_{jk}$ 分别表示第 k 个样本由第 i、j 两个SVM预测的结果,Y_k 表示第 k 个样本的期望输出 y_k。则 f_k 可以表示 i、j 两个SVM对第 k 个样本的预测结果对于第 k 样本的真实预测值的相对位置,即若 $f_k<0$,则表示 Y_{ik} 与 Y_{jk} 在 Y_k 的两侧;若 $f_k=0$,$f_k>0$,则表示 Y_{ik} 与 Y_{jk} 在 Y_k 的同侧;若 $f_k=0$,则表示 Y_{ik} 与 Y_{jk} 中至少有一个与 Y_k 相同。

在选择阶段,先将这 T 个子SVM在测试集上的回归预测的结果依其精度进行重新排序,得到的结果记为 $Y_1,\cdots,Y_T$,然后相应的子SVM分别记为 $SVM_1,SVM_2,\cdots,SVM_T$。依次考察最精确的子 SVM_1 与剩余的 $T-1$ 个子 SVM_t($t=2,\cdots T-1$),对于每一个样本 (x_k,y_k),$k=1,\cdots,p$,计算 $f_k=$

*)本文受到国家自然科学基金(60673095),山西省高校科技研究开发项目(20061101),山西省留学人员科技活动择优资助项目,山西省高校青年学术带头人基金资助;张 妤 硕士研究生,主要研究方向:机器学习;王文剑 教授,博士生导师,主要研究方向:计算智能与机器学习。

$(Y_{1k}-Y_k)(Y_{jk}-Y_k)$。令 $d_k=|Y_{1k}-Y_{jk}|$，视 $\lambda_1\leqslant f_k\leqslant\lambda_2$ 且 $d_k\leqslant\lambda_3$ 为选择条件(其中 $\lambda_1,\lambda_2,\lambda_3$ 为设定的常数)，将满足上述条件的所有子 SVM_j 的序号 j 存储在集合 C_k 中。依次考查测试集 G'中的 p 个样本，将得到 p 个集合 $C_1,\cdots,C_p$，将在上述集合中，出现概率在 50%以上的那些 SVM_t 作为最后用于集成的子 SVM。

在个体学习器的结合阶段，本文采用了简单平均和加权平均两种方法。其中加权平均中的权值定义为：

$$\omega_i=\frac{1-E_i}{\sum_N(1-E_N)} \tag{1}$$

$$E_i=\sqrt{\sum_{i=1}^{n}\frac{(a_i-p_i)^2}{n}} \tag{2}$$

其中 a_i 为测试值，p_i 为期望输出值。

基于上述分析，本文提出的 SVM 集成回归算法归纳如下：

Step1：设定 $\lambda_1,\lambda_2,\lambda_3$ 的值；从训练集 G 中用 Bootstrap Sampling 的方法产生出 T 个子训练集 $G_1,\cdots,G_T$；在每个训练集 $G_i(i=1,\cdots,T)$上学习得到 T 个个体模型 $SVM_i(i=1,\cdots T)$；

Step2：将 T 个子 SVM_i 在测试集上测试，依其泛化精度对每个子 SVM 进行重新排序，记为：$SVM_1,SVM_2,\cdots,SVM_T$；

Step3：依次考察最精确的子 SVM_1 与剩余的 $T-1$ 个子 SVM，对于每一个样本$(x_k,y_k)\in G'$的预测结果，计算：

$$f_k=(Y_{ik}-Y_k)(Y_{jk}-Y_k) \tag{3}$$

及

$$d_k=|Y_{ik}-Y_{jk}| \tag{4}$$

将满足 $\lambda_1\leqslant f_k\leqslant\lambda_2$ 且 $d_k\leqslant\lambda_3$ 的所有的子 SVM_j 的序号 j 存储在集合 C_k 中；

Step4：将在 $C_1,\cdots,C_p$ 该整体中，出现概率在 50%以上的那些 SVM_j 作为最后用于集成的子 SVM；

Step5：依式(1)计算每个子 SVM 学习器的权值，并可得到集成学习的最终结果：

$$F_1(x_l)=\sum_{l=1}^{N}\omega_l\cdot F_l(x_l) \tag{5}$$

$$F_2(x_l)=\frac{1}{N}\sum_{l=1}^{N}F_l(x_l) \tag{6}$$

其中式(5)、(6)分别为加权平均和简单平均的集成预测结果。

3 实验结果及分析

3.1 数据集及误差评价指标

本文采用了三个常用的数据集对本文提出的 SVM 集成回归方法进行测试，它们分别定义如表 1 所示。

表 1 本文实验所采用的数据集

Data set	Function	Variable
2−d Mexican Hat	$y=\sin c\lvert x\rvert=\frac{\sin\lvert x\rvert}{\lvert x\rvert}$	$x\sim U\lvert-2\pi,2\pi\rvert$
3−d Mexican Hat	$y=\frac{\sin\sqrt{x_1^2+x_2^2}}{\sqrt{x_1^2+x_2^2}}$	$x\sim U\lvert-4\pi,4\pi\rvert$
Sin C	$y=\frac{\sin\lvert x\rvert}{x}$	$x\sim U\lvert0,2\pi\rvert$

从每个数据集中随机选取 200 个数据作训练集，200 个数据作为测试集。

为了评价模型的回归效果，本文采用 MAE(Mean Absolute Error)、RMSE(Root Mean Square Error)和 WIA(Willmott's index of agreement)来度量 SVM 集成的预测结果。它们分别定义如下：

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|a_i-p_i| \tag{7}$$

$$RMSE=\sqrt{\sum_{i=1}^{n}\frac{(a_i-p_i)^2}{n}} \tag{8}$$

$$WIA=\frac{\sum_{i=1}^{n}(a_i-p_i)^2}{\sum_{i=1}^{n}(|a'_i|+|p'_i|)} \tag{9}$$

其中 a_i 为测试值，p_i 为期望输出值，$a'_i=a_i-\bar{a}$，$p'_i=p_i-\bar{a}$，$\bar{a}=\frac{1}{n}\sum_{i=1}^{n}a_i$

3.2 SVM 选择性集成学习方法的回归结果

实验中，首先利用 Bagging 方法在训练集上训练得到 20 个子 SVM 个体模型，然后，对其进行简单平均和加权平均，可得最后的集成结果。本实验中，各子 SVM 采用高斯核($k(x_i,y_j)=\exp(-\|x-y\|^2/2p^2)$)，其中 $p=1.0$、$C=10$、$\varepsilon=0.001$。

在上述实验的基础上，采用本文提出的选择性集成方法可以降低集成规模。为了提高个体的多样性，子训练集的大小为原始数据集的一半。实验中所使用的参数设置见表 2。

表 2 算法的参数设置

Data set	λ_1	λ_2	λ_3
2-d Mexican Hat	−0.85e−006	0.85e−006	0.0002
3-d Mexican Hat	−2.5e−006	2.5e−006	0.009

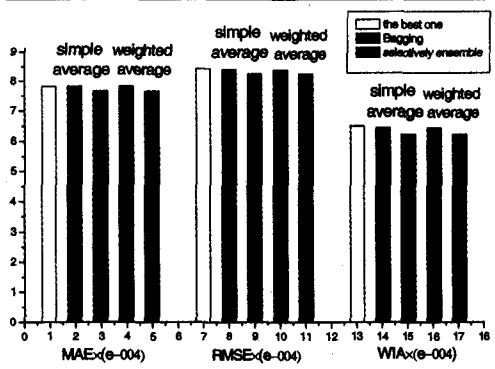

图 1 在数据集 2-d Mexican Hat 上的回归结果图

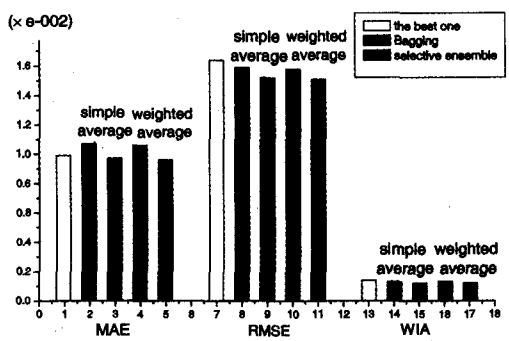

图 2 在数据集 3-d Mexican Hat 上的回归结果图

算法在数据集 2-d Mexican Hat 上，选出了在 Bagging 集成实验中的 8 个个体用作最后的集成；在数据集 3-d Mexican Hat 上，选出了在 Bagging 集成实验中的 12 个个体用作最后

的集成。图1和图2分别给出了在数据集2-d Mexican Hat和3-d Mexican Hat上本文提出的算法与最好个体及未选择前(Bagging方法)的预测结果的比较。

从图1和2中可以看出,采用本文提出的算法可以有效地缩小集成规模,集成效果明显好于单个个体,而且经过选择的SVM集成比未经选择的Bagging方法具有更高的预测精确度。用Bagging方法集成后得到的MAE值略坏于最好个体;对于RMSE、WIA,无论简单平均还是加权平均,集成效果均要好于最好个体。在任何情况下,加权平均略好于简单平均。

由于在SVM中自由参数p、惩罚因子C和误差ε对SVM的回归效果有直接影响,本文还考察了这三个参数对集成结果的影响情况。实验结果分别见表3、表4。

表4 参数取不同值时在数据集SinC上的集成效果

	误差	最好个体	简单平均	加权平均
C、ε、p 取不同值	MAE	2.9502e−004	1.5626e−004	1.5454e−004
	RMSE	1.0632e−007	3.5567e−008	3.4869e−008
	WIA	1.1006e−007	3.6822e−008	3.6099e−008

其中,当C、ε不变,p取不同值时,在2-d Mexican Hat上最好个体的参数设置为p=1.5、C=10、ε=0.001;在3-d Mexican Hat上最好个体的参数设置为:

表3 参数取不同值时在数据集2-d Mexican Hat和3-d Mexican Hat上的集成效果

		2-d Mexican Hat			3-d Mexican Hat		
p 取不同值	误差	最好个体	简单平均	加权平均	最好个体	简单平均	加权平均
	MAE	6.7589e−004	6.1494e−004	6.1425e−004	8.1818e−004	7.4715e−004	7.4791e−004
	RMSE	5.7111e−007	5.1126e−007	5.1068e−007	7.4452e−007	9.1952e−007	9.1952e−007
	WIA	5.2062e−007	4.6607e−007	4.6554e−007	5.5101e−006	4.6876e−006	4.6876e−006
p、C 取不同值	误差	最好个体	简单平均	加权平均	最好个体	简单平均	加权平均
	MAE	7.1040e−004	6.5997e−004	6.5606e−004	7.5130e−004	7.0362e−004	7.0277e−004
	RMSE	5.9282e−007	5.4049e−007	5.3944e−007	9.0350e−007	8.3035e−007	8.2471e−007
	WIA	5.4043e−007	4.9273e−007	4.9177e−007	4.6052e−006	4.2324e−006	4.2037e−006

p=2.6、C=10、ε=0.001。当ε不变,p、C取不同值时,在2-d Mexican Hat上最好个体的参数设置为p=2.4、C=15、ε=0.001;在3-d Mexican Hat上最好个体的参数设置为p=2.9、C=11、ε=0.0009。当ε、p、C均取不同值时,在SinC上最好个体的参数设置为p=1.5、C=12、ε=0.00045。

从表3、表4可以看出,当子SVM取不同的参数值时,回归集成效果要比取最佳参数值时的最好个体更好。这种方法也为核参数的选取提供了一定的实验支持。

结束语 本文提出了一种选择性支持向量机的回归集成方法。通过引入f、d两个量来衡量子SVM的预测能力,以选择Bagging方法中所使用个体模型中的一部分进行最终的集成,这样不仅降低集成规模,而且比未选择的集成方法Bagging具有更强的泛化能力。得到的预测结果明显好于单个SVM,这种方法可以在一定程度上解决SVM模型选择的问题,而且可以将大规模数据分解为小数据集进行学习。本文提出的回归SVM选择性集成方法关于算法性能的理论分析及不同类型核作为子SVM的集成效果值得进一步研究。

参考文献

1 唐伟,周志华.基于Bagging的选择性聚类集成.软件学报,2005,16(4):496～502

2 Jimenez D. Dynamically weighted ensemble Neural networks for classification. In: Proc. IJCNN-98, Vol. 1, Anchorage, AK, IEEE Computer Society Press, Los Alamitos, CA, 1998. 753～756

3 Hansen L K, Liisberg L, Salamon P. Ensemble methods for handwritten digit recognition. In: Proc. IEEE Workshop on Neural Networks for Signal Processing, Helsinoer, Denmark, IEEE Press, Piscataway, NJ, 1992. 333～342

4 German S, Bienenstock E, Dourest R. Neural networks and the bias/variance dilemma. Neural Comput, 1992, 4(1): 1～58

5 Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123～140

6 Freund Y, Schapire R E. Experiments with a new boosting algorithm, In: Proc. ICML-96, Bari, Italy Morgan Kaufmann. San Mateo, Ca, 1996. 148～156

7 Giorgio, Giacinto, Roli F. Design of effective neural network ensembles for image classification purposes. Image and Vision Computing, 2001, 19: 699～707

8 Bauer E, Kohavi R. An empirical comparison of voting classification Algorithms: Bagging, boosting and variants. Machine Learning, 1999, 36(1-2): 105～139

9 Vapnik V N. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000

10 Vapnik V, Golowich S, Smola A. Support Vector method for function approximation, regression estimation, and signal processing. In: M. Mozer, M. Jordan, T. Petsche, eds. Neural Information Processing Systems, Vol. 9, MIT Press, Cambridge, MA, 1997