文章编号:0253-9993(2012)08-1418-05

露天煤矿卡车路段行程时间的实时动态预测新方法

薛　雪,孙　伟,梁　睿

(中国矿业大学 信息与电气工程学院,江苏 徐州　221116)

摘　要:针对露天煤矿卡车优化调度中重要的行程时间预测问题,考虑影响卡车行程时间的各种因素,建立卡车行程时间预测模型,利用最小二乘支持向量回归算法(LS-SVR)和选择性集成学习思想,提出一种基于最小二乘支持向量回归的选择性集成学习算法实现卡车行程时间的动态预测。并在实际采集的露天煤矿数据上进行实验,得到较高的预测精度,说明了算法的有效性。

关键词:露天煤矿;卡车;行程时间;动态预测;最小二乘支持向量回归;选择性集成学习

中图分类号:TD57　　**文献标志码**:A

A new method of real-time dynamic forecast of truck link travel time in open mines

XUE Xue, SUN Wei, LIANG Rui

(*School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou*　221116, *China*)

Abstract: Aiming at travel time prediction problem in optimal dispatching of truck in open coal mines, a truck travel time prediction model which considered various truck travel time influencing factors was built. Using least squares support vector regression (LS-SVR) algorithm and selectivity ensemble learning concept, this paper proposed a truck travel time dynamic prediction method realized by selectivity ensemble learning algorithm based on least squares support vector regression. Experiments were done using the practical data acquired from open coal mines. Higher prediction accuracy was obtained, and the effectiveness of the proposed algorithm was proved.

Key words: open mine; truck; travel time; real-time dynamic forecast; least squares support vector regression; ensemble selection learning

露天煤矿中,卡车在路径上的行驶时间是卡车实时优化调度的重要依据,对其预测准确程度的高低,将直接影响到调度决策的实时性和可靠性。所以如何相对准确地预测各种卡车在不同气候条件下,及在不同行车条件下路段上的行程时间就成为了露天煤矿卡车实时调度中的重点研究内容之一。在传统的卡车调度系统中,通常是对卡车的路径行程时间或运行速度进行统计观测,并将统计的平均值作为预测值。这种方法不可避免地带有较大的误差,如果将其用于实时调度决策,有可能会造成生产失衡,难以实现预先的生产计划。

路径行程时间是求解车辆路径问题的重要参数。国内外关于行程时间预测的研究方法包括了非参数回归模型[1]、时间序列预测方法[2]、卡尔曼滤波预测[3-4]、神经网络预测方法等[5]。目前,关于路径行程时间预测问题的研究主要集中于公交系统中,而对于露天煤矿实时调度中的路径行程时间预测问题却缺乏足够的研究。孙庆山[6]对卡车调度中道路运行时间的统计方法进行了研究,建立了多型号卡车运输的矿山路段运行时间统计的数学模型。白润才等[7]在 2005 年应用人工神经网络(ANN)对卡车路段时间进行实时动态预测,取得了良好的预测性能。

在文献[7]中,证明了行程时间预测函数是复杂的非线性关系,指出了单因子预测方法的不足,提出

收稿日期:2011-08-18　　**责任编辑**:许书阁
基金项目:中央高校基本科研业务费专项资金资助项目(2011QNA19)
作者简介:薛　雪(1980—),女,江苏徐州人,副教授,博士。E-mail:cumtxx@126.com。通迅作者:梁　睿(1981—),男,江苏连云港人,副教授,博士。E-mail:cumtlr@126.com

了多因子预测的神经网络行程时间模型,但该模型存在局部极小点,导致收敛速度慢、迭代次数多等问题。为使行程时间预测具有实时性、可靠性和更高的精度,笔者尝试结合集成学习和支持向量回归理论,提出了一种基于最小二乘支持向量回归的选择性集成学习的卡车行程时间预测算法建立卡车路段行程时间预测函数 $t=f(x_1,x_2,x_3,x_4)$ 模型,其中 x_i 为预测因子,进行了实例仿真,以取得满意结果,并与普通的集成学习算法和 BP 网络预测模型比较。

1　基于 LS-SVR 的选择性集成学习的卡车行程时间预测

1.1　最小二乘支持向量回归

支持向量机(Support Vector Machine)[8]是 Vapnik 等根据统计学理论提出的一种新的通用学习方法,它是建立在统计学理论的 VC 维理论和结构风险最小原理基础上的,能较好地解决小样本、非线性、高维数和局部极小点等实际问题,已成为机器学习界的研究热点之一,并成功地应用于分类、函数逼近和时间序列预测等方面。

最小二乘支持向量机(Least Squares Support Vector Machine,LS-SVM)[9]是支持向量机的改进,与标准 SVM 模型比较,该方法优势明显:① 用等式约束代替标准 SVM 算法中的不等式约束;② 将求解二次规划问题转化为直接求解线性方程组避免了不敏感损失函数,大大降低了计算复杂度,且运算速度高于一般的支持向量机。对最小二乘支持向量回归的理论进行介绍。

设有训练样本集为:$L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_{|L|},y_{|L|})\}$,则有回归函数为

$$y(x)=\boldsymbol{\omega}\varphi(x)+b \tag{1}$$

式中,$\varphi(x)$ 为特征映射;$\boldsymbol{\omega}$ 为权值向量;b 为偏置项。

LS-SVRM 算法的目标为求解以下的最小值问题:

$$\begin{cases}\min Q(\boldsymbol{\omega},e)=\dfrac{1}{2}\|\boldsymbol{\omega}\|^2+\dfrac{\gamma}{2}\sum\limits_{i=1}^{l}e_i^2\\ y_i=\boldsymbol{\omega}\varphi(x_i)+b+e_i,i=1,2,\cdots,l\end{cases} \tag{2}$$

其中,γ 为正则化参数,引入 Lagrange 函数

$$L(\omega,b,e,\boldsymbol{a})=\frac{1}{2}\|\boldsymbol{\omega}\|^2+\frac{\lambda}{2}\sum_{i=1}^{l}e_i^2-\sum a_i(\boldsymbol{\omega}\varphi(x_i)+b+e_i-y_i) \tag{3}$$

式中,$\boldsymbol{a}$ 为 Lagrange 乘子,$\boldsymbol{a}=[a_1,a_2,\cdots,a_{|L|}]^{\mathrm{T}}$。

根据 KKT(Karush-Kuhn-Tucker)条件,对 $\boldsymbol{\omega}$,b,e_i,a_i 求偏导数,并令其为零,得到

$$\begin{cases}\dfrac{\partial L}{\partial\boldsymbol{\omega}}=\boldsymbol{\omega}-\sum\limits_{i=1}^{l}a_i\varphi(x_i)=0\\ \dfrac{\partial L}{\partial b}=-\sum\limits_{i=1}^{l}a_i=0\\ \dfrac{\partial L}{\partial e_i}=\gamma e_i-a_i=0\\ \dfrac{\partial L}{\partial a_i}=\boldsymbol{\omega}\varphi(x_i)+b+e_i-y_i=0\end{cases} \tag{4}$$

根据式(4),优化问题可以变成求解线性方程的问题,最后可得最小二乘支持向量机的回归模型为

$$y(x)=\boldsymbol{\omega}\varphi(x)+b=\sum_{i=1}^{l}a_iK(x_i,x_j)+b \tag{5}$$

其中,将确定的回归函数的参数 a,b 统称为回归参数;$K(x_i,x_j)$ 为核函数,本文采用的是径向基函数:

$$K(x_i,x_j)=\exp\left(-\frac{|x_i-x_j|^2}{\sigma^2}\right)$$

其中,σ^2 为核函数的核宽度。对最小二乘支持向量回归机进行训练,最后得到的模型参数 $M_j=(a_j,b_j)$,其中 $j=1,2,\cdots,T$。

1.2　选择性集成

集成学习(Ensemble Learnnig)使用多个(通常是同质的)学习器来解决同一个问题。可以有效地提高学习系统的泛化能力。如何将多个学习器进行结合,得到一个好的集成系统的输出结果是集成学习关心的重要问题。对于回归问题,当每个学习器的重要程度相同时,则采用简单平均的方法将各子学习器的输出值结合。当每个学习器的重要程度不同时,则给每个学习器赋上相应的权值再加权平均[10]。

由于集成学习可以利用多个学习器获得比仅使用单一学习器更强的泛化能力,因此很多算法都是通过集成大量的个体学习器来获得更好的性能。例如著名的 Boosting 算法[11]和 Bagging 算法[12]。但是随着个体学习器数目的增加,一方面,算法的计算和存储开销越来越大;另一方面,学习器之间的差异性越来越难以获得。Zhou 等[13-14]提出了“选择性集成”思想,并设计了选择性集成算法 GASEN。他们的研究结果表明,从已有的学习器集合中选择部分学习器来集成可以获得更好的性能。而且其理论分析和实验结果均表明,该方法性能优于传统集成学习方法。

选择性集成的基本思想就是利用多个学习器,并对学习器进行适当的选择来剔除对学习系统整体性能有负作用的学习器,最后将所选择的结果进行结合从而得到更好的学习器集合。文献[15]证明了可以大幅缩小集成的规模但却不损失泛化能力。回归问题中,当训练出多个学习器之后,从中选择一部分进

行集成,可望比使用所有学习器进行集成更好。

在本文中,对于通过训练得到的 T 个子学习器 $f_1,f_2,\cdots,f_T$,在训练样本集 S 上分别计算其对应的均方误差 MSE($m_1,m_2,\cdots,m_T$)。对 MSE 进行排序,找到最小均方误差 $m_{\min}$ 对应的学习器 $f_{\min}$。对于第 k 个学习器 $f_k(1\leqslant k\leqslant T-1)$,计算学习器 $f_{\min}$ 与 f_k 集成学习后的均方误差 m_k,如果满足 $m_k\geqslant m_{\min}$,则认为利用 f_k 进行集成学习,对回归预测起负面影响,所以去除学习器 f_k。

图 1 为基于 LS-SVR 的选择性集成学习的卡车行程时间预测算法的算法流程。本文算法利用集成学习思想对训练样本 S 利用 bootstrap 方法进行采样得到 T 个训练样本子集 $S'(S'_1,S'_2,\cdots,S'_T)$;之后使用训练子集对 LS-SVR 进行训练,得到 LS-SVR 的模型参数 $M(M_1,M_2,\cdots,M_T)$,即得到 T 个子学习器;然后对 T 个子学习器进行选择,去除 k 个对回归结果起负面作用的子学习器,再对剩余的($T-k$)个子学习器进行集成,得到最后的学习模型,利用这个模型对卡车行程时间进行预测。

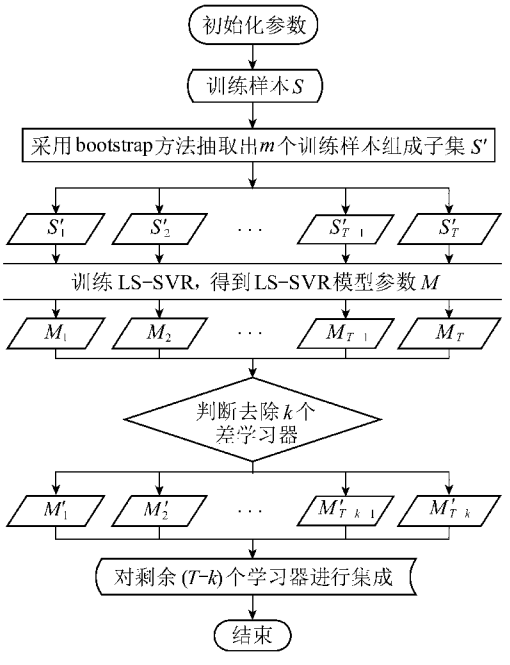

图 1 基于 LS-SVR 的选择性集成学习的卡车行程时间预测算法

Fig. 1 Algorithm of truck travel time prediction based on LS-SVR ensemble selection learning

2 算法步骤

基于 LS-SVR 选择性集成学习的卡车行程时间预测算法步骤如下:

(1)初始化参数。包括训练样本 S,训练子集的个数 T,训练子集大小 m。LS-SVR 的核函数类型 $K(x_i,x_j)$、正则化参数 γ 和核宽度参数 σ^2。

(2)从训练样本 S 中采用 bootstrap 方法抽取出 m 个训练样本组成训练子集 $S'_1,S'_2,\cdots,S'_T$。

(3)分别利用训练子集 $S'_1,S'_2,\cdots,S'_T$ 训练 LS-SVR,得到 T 个 LS-SVR 的模型参数 $M_1,M_2,\cdots,M_T$,即 T 个学习器 $f_1,f_2,\cdots,f_T$。

(4)对 T 个学习器进行选择,去除对回归结果起负面影响的 k 个学习器,得到 $T-k$ 个新的子学习器 $f'_1,f'_2,\cdots,f'_{T-k}$。

(5)对子学习器 $f'_1,f'_2,\cdots,f'_{T-k}$ 的回归结果进行平均,得到最后的回归预测值:

$$f_s=(f'_1+f'_2+\cdots+f'_{T-k})/(T-k) \qquad (6)$$

3 实验仿真

以某一露天煤矿的卡车行程时间预测为研究对象,使用 MTALAB 7.8.0(R2009a)进行仿真实验。对本文所提算法、基于 LS-SVR 的集成学习算法、LS-SVR、人工神经网络 BP 算法分别进行实验,验证基于 LS-SVR 的选择性集成学习算法在回归精度上的良好效果。

从实际的露天煤矿卡车行程时间数据库中选取了 312 个数据作为实验样本。实验中,分别随机选取 50,100,150,200 个样本作为训练样本进行训练;随机选取除训练样本以外 63 个样本作为测试样本。集成学习实验中,设定子学习器的个数 $T=10$,每个子学习器选取样本数为 $m=$(训练样本数/3)。所有类型 LS-SVR 的核函数均取为径向基核函数,通过交叉验证法来选择正则化参数 γ 和核宽度 σ^2。对于 BP 网络学习,输入层节点数为 5 个;隐藏层节点数为 13 个,激励函数采用正切 S 形函数("tansig");输出节点数为 1 个,激励函数采用线性函数("purelin")。采用测试样本的均方误差(MSE)来衡量各算法的的回归预测性能。

对每组数据集重复 100 次实验,取均方误差的平均值作为最后输出,得到表 1 和图 2。

表 1 测试样本的平均均方误差(MSE)

Table 1 Average square error of testing samples

算 法	训练样本数			
	50	100	150	200
基于 LS-SVR 的选择性集成学习算法	0.090 49	0.055 62	0.030 00	0.021 59
基于 LS-SVR 的集成学习算法	0.113 18	0.072 43	0.052 25	0.033 22
LS-SVR	0.150 23	0.107 57	0.066 50	0.040 47
人工神经网络 BP 算法	0.237 18	0.169 43	0.166 57	0.160 05

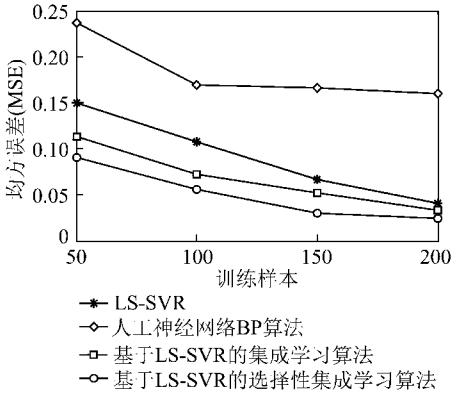

图 2　测试样本的平均均方误差对比

Fig. 2　Average square error comparison of test samples

可以看出,基于 LS-SVR 的选择性集成学习算法的预测精度明显好于其他 3 种方法,说明了本文算法可以很好地解决露天煤矿卡车行程预测问题,得到更精确的预测值。图 2 中,对 LS-SVR 进行集成后的效果要比单纯的 LS-SVR 算法的均方误差小,说明了经过集成的学习器优于单个学习器性能,这和集成学习的理论是相符的;同时,通过对集成学习器进行选择,可以得到比普通的集成学习更好的性能。在训练样本较少的时候,利用 LS-SVR 的 3 种算法可以得到比人工神经网络 BP 算法更好的性能,尤其是基于 LS-SVR 的选择性集成学习算法的均方误差最小,这是由于 LS-SVR 对小样本的学习更有效,而神经网络则需要足够多的训练样本进行训练才能得到满意的结果;这在露天煤矿的实际应用中有非常重大的意义,在采集数据的过程中,受各种条件的约束,需要花费大量的人力物力进行样本采集,所以只能采集到有限训练样本,这就需要能用最少的训练样本进行训练得到回归预测性能最好的学习器,本文算法可以在很少的训练样本上得到相对满意的效果。

表 2 为在不同训练样本情况下,在规定的精度范围内,重复进行 100 次实验的各种算法的平均运行时间。从表中数据可以看出,人工神经网络 BP 算法花费的时间最多,这是由于训练样本有限,神经网络需要多次重复训练来训练网络参数。基于 LS-SVR 的选择性集成学习算法比基于 LS-SVR 的集成学习算法花费的时间稍多,这是因为前者需要花费更多的时间对学习器进行筛选,通过花费了较少的时间却得到了较高的回归精度,所以采用选择性集成对算法是有意义的。随着训练样本的增加,可以看到 LS-SVR 算法花费的时间越来越多,这是由于 LS-SVR 算法是基于样本的算法,伴随着训练样本的增多,需要花费更多的时间进行支持向量选择;而集成学习算法的时间却增加较少,这是因为通过 bootstrap 方法抽取训练子集进行训练时,子学习器的训练样本数是较少的,所以训练速度是很快的,这也反映出采用集成学习算法可以在利用很少的时间得到很好的回归预测性能。

表 2　100 次实验的平均运行时间

Table 2　Average running time of 100 experiments　s

算　法	训练样本数			
	50	100	150	200
基于 LS-SVR 的选择性集成学习算法	0.187 2	0.234 0	0.483 6	0.858 0
基于 LS-SVR 的集成学习算法	0.109 2	0.156 0	0.374 4	0.655 2
LS-SVR	0.374 4	6.786 0	22.479 7	53.835 9
人工神经网络 BP 算法	56.368 2	43.192 9	75.624 8	62.562 7

图 3 表示在 100 个训练样本情况下,选取 100 次重复实验中的前 50 次实验的运行时间,可以看出,人工神经网络 BP 算法的运行时间随机性很大,其他 3 种算法则相对很平稳。在露天煤矿实时调度中,需要系统稳定运行,并且响应时间要尽量短,基于 LS-SVR 的选择性集成学习算法可以很好的满足这些条件,可以很好的解决卡车行程预测问题。

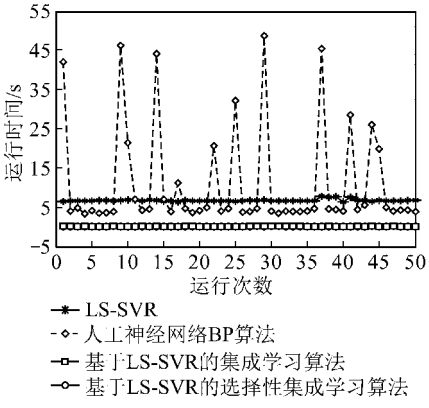

图 3　100 个训练样本重复进行 50 次实验运行时间

Fig. 3　Running times of 50 experiments carried out repeatedly using 100 training samples

4　结　　语

应用多个学习器,并对学习器进行适当的选择,将所选择的结果进行结合从而得到更好的学习器集合,从中选择一部分进行集成,得到基于 LS-SVR 的选择性集成学习的算法。将该算法用在卡车行程时间实时动态预测上,在采集的实际数据样本上进行实验仿真,获得了比 LS-SVR 和 BP 网络更好的回归预测性能,可以满足实际露天煤矿的调度精度要求,具有实际应用价值。同时,认识到算法中有较多的参数

需要进行人为设置,对于参数的优化学习将是今后的研究重点。

参考文献:

[1] 达庆东,段里仁. 交通流非参数回归模型[J]. 数理统计与管理,2003,22(4):41-46.
Da Qingdong,Duan Liren. Non-parameter regression model of traffic flow[J]. Application of Statistics and Management,2003,22(4):41-46.

[2] Mohammadi S,Keivani H,Bakhshi M,et al. Demand forecasting using time series modelling and ANFIS estimator [A]. 41st International Universities Power Engineering Conference [C]. Piscataway:IEEE,2006:333-337.

[3] 温惠英,徐建闽,傅 惠. 基于灰色关联分析的路段行程时间卡尔曼滤波预测算法[J]. 华南理工大学学报(自然科学版),2006,34(9):66-70.
Wen Huiying,Xu Jianmin,Fu Hui. Estimation algorithm with kalman filtering for road travel time based on grey relation analysis[J]. Journal of South China University of Technology(Natural Science Edition),2006,34(9):66-70.

[4] Najjar M E,Bonnifait P. A road-matching method for precise vehicle localization using belief theory and Kalman filtering[J]. Autonomous Robots,2005,19(2):173-191.

[5] 丁 涛,周惠成. 基于径向基函数神经网络的预测方法研究[J]. 哈尔滨工业大学学报,2005,37(2):272-275.
Ding Tao,Zhou Huicheng. Prediction method research based on radial basis function neural network[J]. Journal of Harbin Institute of Technology,2005,37(2):272-275.

[6] 孙庆山. 卡车调度中道路运行时间统计方法[J]. 露天采煤技术,1998(1):33-35.
Sun Qingshan. Statistical method of road running time in truck scheduling[J]. Opencast Coal Mining Technology,1998(1):33-35.

[7] 白润才,李建刚,徐建华. 卡车路段行程时间的实时动态预测[J]. 辽宁工程技术大学学报,2005,24(1):12-14.
Bai Runcai,Li Jiangang,Xu Jianhua. Real-time dynamic forecast of truck link travel time[J]. Journal of Liaoning Technical University,2005,24(1):12-14.

[8] Cortes C,Vapnik V. Support-vector network[J]. Machine Learning,1995,20(3):273-297.

[9] Suykens J A K,Vandewale J. Least Squares support vector machine classifiers[J]. Neural Processing Letters,1999,9(3):293-300.

[10] Perrone M,Cooper L N. When networks disagree:ensemble method for Hybrid neural networks[M]. London:Chapman & Hall,1993.

[11] Seha Pire R E. The strength of weak liability[J]. Machine Learning,1990,5(2):197-227.

[12] Breiman L. Bagging perdictors[J]. Machine Learning,1996,24(2):123-140.

[13] Zhou Zhihua,Wu Jianxin,Tang Wei,et al. Combining regression estimators:GA-based selective neural network ensemble[J]. International Journal of Computational Intelligence and Applications,2001,1(4):341-356.

[14] Zhou Zhihua,Wu Jianxin,Tang Wei. Ensembling neural networks:Many could be better than all[J]. Artificial Intelligence,2002,137(1/2):239-263.

[15] 盛高斌. 基于半监督回归的选择性集成算法及其应用研究[D]. 杭州:浙江工业大学,2009.
Sheng Gaobin. Research of ensemble selection algorithm based on semi-supervised regression and its application[D]. Hangzhou:Zhejiang Polytechnical University,2009.