多核小波支持向量机在 Carrousel 氧化沟系统的应用①

何渊淘, 刘超慧

(郑州航空工业管理学院 计算机科学与应用系, 郑州 450000)

摘 要: Carrousel 氧化沟广泛应用于城市污水处理, 但污水处理的效果受到水质和环境因素影响很大, 难以建立精确的预测模型. 现有的机器学习方法普遍预测效果较差, 为了准确预报污水处理的效果, 本文采用多核小波支持向量机进行建模, 实验表明该方法提高了预报的精确度, 适合用于氧化沟系统的实时在线预测.

关键词: 多核支持向量机; SILP 算法; SimpleMKL 算法; 小波核函数; 氧化沟系统

Application of Wavelet Multi Kernel Learning on Carrousel Oxidation Ditch System

HE Yuan-Tao, LIU Chao-Hui

(Department of Computer Science and Application, Zhengzhou Institute of Aeronautical Industry Management, Zhengzhou 450000, China)

Abstract: Carrousel oxidation ditch has been widely used in sewage disposal system, however the result of sewage disposal was affected by water quality and environmental factors, so it is difficult to build a precise prediction model. The existing method of machine learning algorithm usually gets a poor result in prediction. In order to precisely predict the result, this essay uses the multi-kernel wavelet support vector machine when build a model. The outcome of this experiment demonstrates that the new method improve the degree of definition in forecasting, and it is suitable for actually online prediction.

Key words: multi-kernel svm; SILP algorithm; SimpleMKL algorithm; wavelet kernel; oxidation ditch system

1 引言

1967 年, 第一代 Carrousel 氧化沟在荷兰 DHV 公司研制成功, 满足了城市污水脱氮, 脱磷的要求. 时至今日氧化沟发展到了第三代, 但是其基本的原理都是利用嗜氧细菌来降解水中的营养物质. 由于生物反应过程本身的复杂性, 现有的数学模型难以满足多变的反应环境.

多核支持向量机是单核支持向量机的新发展, 由于多核的引入算法的泛化能力、训练速度和目标函数可解释性进一步提高, 本文将小波函数应用于多核支持向量机中从而为氧化沟反应建立更为准确的模型.

2 多核支持向量机

多核支持向量机是单核支持向量机的扩展, 使用多个同质不同属性或不同质的核函数的组合来替代传统支持向量机中单一的核函数. 近几年的研究发现将多核、合成核引入支持向量机能进一步提高支持向量机的性能. 文献[1-2]表明将多个性质差别较大的核函数进行一定规则的组合较好解决了文本、DNA 等较为复杂的分类问题, 进一步提高了此类问题的泛化能力. 多核函数最常见的形式为:

$$k(x,z)=\sum_{i=1}^{M}\beta_i k_i(x,z),\beta_i>0 \tag{1}$$

将核函数替换后, 支持向量机原问题转化为:

$$\begin{aligned}&\min: && \frac{1}{2}\left(\sum_{k=1}^{K}\|w_k\|_2\right)^2+C\sum_{i=1}^{N}\xi_i\\ &w.r.t. && w_k\in R^{D_k}\quad \xi\in R^N\quad b\in R\\ &s.t. && \xi_i\ge 0\ and\ y_i\left(\sum_{k=1}^{K}\langle w_k,\phi_k(x_i)\rangle+b\right)\ge 1-\xi_i\\ & && \forall i=1,\dots N\end{aligned} \tag{2}$$

① 基金项目:国家自然科学基金(41171341)

收稿时间:2013-03-19;收到修改稿时间:2013-04-27

经过对偶变换[3]，原问题转变为公式(3):

$$\begin{aligned}&\max_{\beta}\min_{\alpha}\ \sum_{k=1}^{K}\beta_k S_k(a)\\&w.r.t.\quad a\in R^N\quad \beta\in R^K\\&s.t.\quad 0\le a\le C\ \ 0\le\beta\\&\qquad \sum_{i=1}^{N}a_i y_i=0\ and\ \sum_{k=1}^{K}\beta_k=1\end{aligned}\tag{3}$$

Sonnenburg 在论文[3]提出了此类问题的求解算法，即 SILP 算法，使用外循环和内循环的方式来求解. 外循环使用传统 SVM 算法计算出最优的乘子 a_i ，而内循环使用线性规划计算出最佳的 β_k 值. 由于 SILP 面临迭代次数多的问题，Rakotomamonjy 在[4,5]中提出了 SimpleMKL 算法，将目标函数中 f_m 的范数改为其范数的二次方，此时原问题转变为公式(4).

$$\begin{aligned}&\min_{\{f_m\},b,\xi,\beta}\ \frac{1}{2}\sum_m\frac{1}{\beta_m}\|f_m\|_{H_m}^2+C\sum_i\xi_i\\&s.t.\quad y_i\sum_m f_m(x_i)+y_i b\ge 1-\xi_i\ \ and\ \xi_i\ge 0\quad\forall i\\&\qquad \sum_m\beta_m=1\quad \beta_m\ge 0\quad\forall m\end{aligned}\tag{4}$$

经过对偶变换后公式(4)转变为公式(5). 此时目标函数 $J(\beta)$ 对于 β_k 值可微分，内循环可使用梯度法来加快收敛速度，以求得最优的 β_k 值. SimpleMKL 相对于 SILP 算法，其收敛速度得到了显著的提高，更适合在线学习.

$$\min_{\beta}J(\beta)\quad such\ that\ \sum_{m=1}^{M}\beta_m=1\ \ \beta_m\ge 0$$

$$J(\beta)=\begin{cases}\min\limits_{\{f\},b,\xi}\ \frac{1}{2}\sum\limits_m\frac{1}{\beta_m}\|f_m\|_{H_m}^2+C\sum\limits_i\xi_i\\ s.t.\quad y_i\sum\limits_m f_m(x_i)+y_i b\ge 1-\xi\\ \qquad and\ \ \xi_i\ge 0\ \ \forall i\end{cases}\tag{5}$$

3 小波核函数

小波是现代信息处理领域的一个强大的工具，通过小波函数族可以逼近任意的信号. 其中小波母函数起到最重要的作用，通过对小波母函数的缩放和平移可以生成一族小波函数. 假定 $h(x)$ 为小波母函数，小波函数族可以通过如下的方式构成公式(6)(7)

$$h_{a,c}(x)=|a|^{-1/2}h(\frac{x-c}{a})\ a,c,x\in R\tag{6}$$

在公式(6)中，其中 a 是缩放因子，c 是平移因子. 对于任何一个函数 $f(x)\in L_2(R)$ 的小波变换可以写成

$$W_{a,c}(f)=\langle f(x),h_{a,c}(x)\rangle\tag{7}$$

通过反向小波变换可得到原函数

$$f(x)=\sum_{i=1}^{l}W_i\cdot h_{a_i,c_i}(x)\tag{8}$$

由公式(7)(8)可以看出，小波族函数有着良好的逼近性质，适合做为多核支持向量机的核函数[8-13]. 在本文中采用如下形式的小波核函数

$$K(x,x')=\prod_{i=1}^{N}h\left(\frac{x_i-c_i}{a}\right)h\left(\frac{x_i'-c_i'}{a}\right)\tag{9}$$

4 实验结果及其分析

在本文中我们使用了河南省漯河市的污水净化中心采集到的数据. 该数据是在 2008 到 2012 年间采集的生产记录，污水来源主要是食品加工企业污水和城市生活污水，样本集共有十个属性，其中如水指标为: 水温 T、进水 SS、进水 COD、进水 TN、进水 TP、MLSS、MLVS、SV30，输出数据为出水 TN，出水 TP，总数据共 2080 组[14-16].

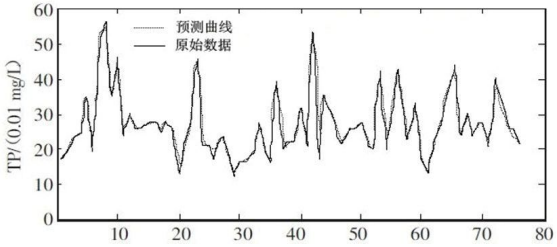
为了消除个别样本对实验结果的影响，本文采用五次交叉验证来对样本进行随机分类. 共平均分成五份数据，四份数据作为训练集，一份作为测试集，并通过多次实验来对结果取平均值. 在本实验中采用了公式(10)的母函数，并通过使用公式(11)来生成一族小波函数，然后用小波函数的线性组合来生成多核小波核函数，见公式(12).

$$h(x)=\cos(1.75x)\exp\left(-\frac{x^2}{2}\right)\tag{10}$$

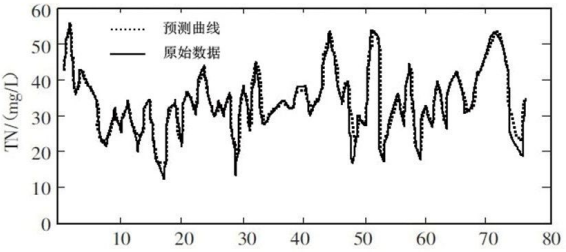
$$K_l(x,x')=\prod_{i=1}^{N}h\left(\frac{x_i-x_i'}{a_l}\right)\tag{11}$$

$$K(x,x')=\sum_{l=1}^{K}\beta_l K_l(x,x')\tag{12}$$

本实验在 Matlab 下进行仿真，使用了单核多核支持向量机工具包，实验结果如下. 图一、图二，表一为多核小波支持向量机对出水 TP、出水 TN 的预报曲线及算法迭代次数对应的出水 TP 预报误差.



出水 TP 预报



出水 TN 预报

从中可以看出基于多核小波支持向量机的预测模型很好得预测了氧化沟系统的出水 TN 和 TP 指标. 同时在表 1 中的数据表明多核支持向量机的最优测试误差达到了 0.0966%, 与单核支持向量机的分类误差相比有了一定的提高. 这主要是由于在本实验中多核小波支持向量机使用了多个不同参数小波函数, 这些不同参数的小波核函数能较好适应氧化沟数据在不同环境下得到的出水数据的变换. 从表 2 的数据可以看出, 多核支持向量机选用了较多的核函数. 在 SILP 和 SimpleMKL 分别选取了 23 和 40 个核来构成决策函数, 相比于采用单一内核的支持向量机更好地适应了氧化沟环境的变换. 由于多核支持向量机需要额外的迭代来寻找最优的线性核函数 $\sum_{k=1}^{K}\beta_k S_k(a)$ 的组合系数 β_k, 因此可以得出其整体的迭代次数必然高于单核支持向量机, 算法整体所耗费的时间也较高. 但从表 2 的数据得出, 虽然整体迭代次数增加了很多, 但是算法的总时间耗费并没有增加. 这就表明了寻找线性组合的迭代并没有增加太多的时间耗费, 同时也减少了支持向量机训练的次数, 所以整体的时间耗费反而减少了.

表 1 算法运行次数及分类误差(TP 指标)

运行次数	10	20	50	100
训练误差	0.0603%	0.0513%	0.0514%	0.0512%
测试误差	0.1066%	0.0966%	0.0967%	0.0966%

表 2 污水处理数据(运行 50 次)

	训练误差	测试误差	核函数个数	训练运行时间（秒）
单核支持向量机	0.0513%	0.0966%	1	192
多核 (SILP)	0.0611%	0.0967%	23	130
多核 (SimpleMKL)	0.0611%	0.0897%	40	141

5 结束语

在上述实验中, 多核小波支持向量机算法被用于氧化沟水质预测中. 从实验结果的整体来看多核小波支持向量机对氧化沟系统数据的收敛速度较快, 从与单核支持向量机的对比来看, 多核小波支持向量机在预测准确度上进一步得提高, 因此可以得出多核支持向量机对于氧化沟水样数据变化的适应性更强, 更易于进行水质的在线预测.

参考文献

1 Bach F. Consistency of the group Lasso and multiple kernel learning. Journal of Machine Learning Research, 2008, 9: 1179–1225.

2 Bach F, Lanckriet G, Jordan M. Multiple kernel learning, conic duality, and the SMO algorithm. Proc. of the 21st International Conference on Machine Learning, 2008, 41–48, 2004a.

3 Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B. Large scale multiple kernel learning. Journal of Machine Learning Research I, 2006: 1–18.

4 Rakotomamonjy A, Bach R, Canu S, Grandvalet Y. Simple MKL. Journal of Machine Learning Research X, 2008: 1–34.

5 Rakotomamonjy A, Bach R, Canu S, Grandvalet Y. More efficiency in multiple kernel learning. Proc. of the 24th International Conference on Machine Learning, Corvallis, OR, 2007.

6 Antoniadis A, Fan J. Regularization by wavelet approxi-

(下转第 197 页)

4 结语

车载计算机系统作为指控系统的重要组成部分，通过对已存储的数据和所采集到的信息进行加工，以产生各种控制信息和显示信息，从而进行作战决策和控制，因此对其可用性进行评价和改进性能至关重要. 本文所提出的评价指标体系和评价方法在某装备的车载计算机系统上进行了应用，在一定程度上解决了目前车载计算机系统的可用性评价问题. 但具体的评价属性测量和定性数据定量化还有许多工作要做.

研究可用性评估的目的不是具体评价系统的可用性分数，而是评估系统目前所存在的主要问题，以便提出改进的方法与措施，提高车载计算机系统的质量，进而完善指控系统的功能和性能，达到提升装备的作战效能的目的.

参考文献

1 ISO 9421-10. Ergonomic requirements for office work with visua1display terminals(VDT's)-Part 10: Dialogue principles. Internationa1 Organization of Standardization, 1994.

2 IS0 9241-l1. Ergonomic requirements for office work with visua1display terminals(VDT's)-Part l1: Guidance on usability.Internationa1 Organization of Standardization,1997.

3 Gabbard JL, Swan JE II.Usability engineering for augmented reality: Employing user-based studies to inform design. IEEE Trans. on visualization and Computer Graphics, 2008, 14(3): 513–525.

4 尼尔森,刘正捷等译.可用性工程.北京:机械工业出版社,2004.

5 李乐山.人机界面设计(实践篇).北京:科学出版社,2009.

6 贺桂和,谭春辉,邓艳华.C2C 电子商务网站可用性评价指标体系设计研究.荆楚理工学院学报,2010,25(12):60–64.

7 唐琼,张新鹤.基于可用性的电子资源质量评价指标体系研究.图书馆理论与实践,2007(5):7–10.

8 任忠斌,孙庆珍.电子地图可用性评估指标体系问题研究.测绘与空间地理信息,2010,33(3):14–17.

9 刘胧,刘虎沉.手机可用性工程生命周期与评价方法.工业工程,2009,13(3):97–101.

10 Upadhyay N, Deshpande BM, Agrawal VP. Concurrent usability evaluation and design of software component:a digraph and matrix approach. IET Softw., 2011, 5(2): 188–200.

11 Kostaras N, Xenos M, Skodras AN. Evaluating usability in a distance digital systems laboratory class. IEEE Trans. on Education, 2011, 54(2): 308–313.

12 Garrido A, Rossi G, Distante D. Refactoring for usability in web applications. IEEE Software, 2011, 28(5): 60–67.

13 Mitsopoulos-Rubens E, Trotter MJ, Lenne MG. Usability evaluation as part of iterative design of an invehicle information system. IET Intell. Transp. Syst., 2011, 5(2): 112–119.

14 陈建明,王洪艳,宣亚克.指控软件可用性工程生命周期模型.指挥控制与仿真,2012,34(4):61–64.

15 李建光,申利民,赵承霞.面向用户的软件柔点可用性评估方法的研究.计算机应用与软件,2011,28(1):61–64.

16 梁保松.模糊数学及其应用.北京:科学出版社,2007.

(上接第 205 页)

mations. American Statistical Association, 2001,96:939–967.

7 Zhang L, Zhou WD, Jiao LC. Wavelet support vector machine. IEEE Trans. on System, Man and Cybernetics–Part B: CybernaticsYBERNATICS, Feb.2004, 34, (1).

8 Vapnik V. The Nature of Statistical Learning Theory. New York, Springer-Verlag, 2000.

9 Vapnik V. Statistical Learning Theory. New York, NY: John Wiley, 1998.

10 Burges CJC. A Tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121–167.

11 Almeida MB, Braga A, Braga JP. SVM-KM: Speeding SVMs learning with a priori cluster selection and k-means. Proc. of the 6th Brazilian Symposium on Neural Networks. Washington, DC: IEEE Computer Society, 2000: 162–167.

12 Hart PE. The condensed nearest neighbor rule. IEEE Trans. on Information Theory, 1968, 14(3): 515–516.

13 Krantz SG, ed.Wavelet: Mathematics and Application. Boca Raton, FL: CRC, 1994.

14 陈安.Carrousel 氧化沟系统水质特征动态分析的人工神经网络模型研究.武汉:武汉理工大学,2003:60–63.

15 陈学群,俞爱媚,吕斌.Carrousel 氧化沟技术发展研究.煤矿环境保护,2002,16(4):46–49.

16 邓乃扬,田英杰.数据挖掘中的新方法-支持向量.北京:科学出版社,2004.