文章编号:1000-582X(2008)02-0179-04

支撑向量机在给水管网故障诊断中的应用

朱晓红[1],朱　丹[2]
(1. 重庆大学 计算机学院,重庆 400030; 2. 中国市政工程西南设计研究院,四川 成都 610081)

摘　要:针对城市给水管网运行故障诊断问题,构造一个给水管网室内实验模型,通过测试节点水压变化,利用建立在结构风险最小原理基础上的支撑向量机(SVM)模式分类功能,确定相应的特征向量和核函数参数;选取样本进行训练和测试,在小样本情况下对管网故障点进行判决。在多次实验模型测试数据的基础上,对基于 SVM 的给水管网故障诊断方法进行测试,并在相同实验测试数据的前提下与神经网络(ANN)法进行比较,实际算例表明,SVM 方法的诊断精度优于神经网络法。

关键词:支撑向量机;模式识别;故障诊断;给水管网
中图分类号:TU821.3　　文献标志码:A

Application of Support Vector Machine in Water Distribution Systems' Fault Diagnosis

ZHU Xiao-hong[1], *ZHU Dan*[2]
(1. College of Computer Science and Engineering, Chongqing University, Chongqing 400030, P. R. China;
2. South west Municipal Engineering Design and Research Institute of China, Chengdu Sichuan 610081, P. R. China)

Abstract: Aimming at the fault diagnosis of urban water supply pipeline network, an indoor experiment model of water supply pipeline network was designed. After the feature vector and nuclear function parameter were confirmed, the method for locating the bursting point of water supply pipeline network was discussed with small swatch data through testing the change of hydraulic pressure and utilizing the pattern recognition function of support vector machine (SVM) based on the theory of structural risk minimization. On the basis of the same experiment testing data, the fault diagnosis method of water supply pipeline network based on SVM was tested and compared with the method based on artificial neural network(ANN). The numerical example indicates that the fault diagnosis method based on SVM is more precise than the method based on ANN.

Key words: support vector machine; pattern recognition; fault diagnosis; water distribution systems

智能诊断是当前工程诊断领域的一个重要发展方向,研究从测试数据出发寻找规律,利用这些规律对故障状态进行识别和预测。现有的基于数据的机器学习,包括神经网络在内,共同的重要理论基础之一是统计学。传统统计学研究的是样本数目趋于无穷大时的渐近理论,但在实际问题中,样本数往往是有限的。统计学习理论(statistical learning theory 或 SLT)是一种专门研究小样本情况下机器学习规律的理论。Vapnik 等人从 20 世纪 60～70 年代开始致力于此方面研究,到 90 年代中期,统计学习理论开始受到越来越广泛的重视。支撑向量机[1-2](support vector machine 或 SVM)是建立在 SLT 的 VC 维理论

收稿日期:2007-10-21
基金项目:重庆市自然科学基金资助项目(2006BB2232)
作者简介:朱晓红(1970-),女,重庆大学副教授,博士后,主要从事计算机应用技术、信息安全等方面的研究,(Tel)13908374663;(E-mail)xhzhu@cqu.edu.cn。

和结构风险最小原理基础上,相比较神经网络而言,具有良好的推广性能,SVM 在解决小样本、非线性及高维问题中表现许多优势,并能推广到其它的机器学习问题中,将有力地推动机器学习理论和技术的发展[3-4]。

1 支撑向量机

给定样本集,$x_i \in R^n, y_i \in \{-1,1\}, i=1,2,\cdots,l$ 和核函数 $k(x_i,y_i)$,k 对应某特征空间 Z 中的内积,即 $k(x_i,y_i)$ 小于 $\varphi(x_i)$,大于 $\varphi(x_j)$。变换 $\varphi: x \to Z$ 将样本从输入空间映射到特征空间。设计基于SVM的二分类器,就是在 Z 中寻找一定意义下的最优分类面 $<w,\varphi(x)> -b=0$。当样本集在 Z 中线性可分时,使分类间隔最大,即求解

$$\min_{w,b} \frac{1}{2}\|w\|^2$$
$$\text{Subject to } y_i(<w,\varphi(x_i)>-b) \geqslant 1, \quad i=1,2,\cdots,l。 \tag{1}$$

当样本集中在 Z 中线性不可分时,使分类间隔和分类错误达到某种折中,即求解

$$\min_{w,b,\xi}\left[\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i\right]$$
$$\text{Subject to } y_i(<w,\varphi(x_i)>-b) \geqslant 1-\xi_i, \quad \xi_i \geqslant 0, i=1,2,\cdots,l, \tag{2}$$

其中,ξ_i 是松驰变量,C 为正则化参数。

大多数方法不直接求解式(1)和(2),而是求解其对偶问题

$$\min_{\alpha}(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{\alpha} - e^{\mathrm{T}}\boldsymbol{\alpha}$$
$$\text{Subject to } \quad y^{\mathrm{T}}\boldsymbol{\alpha} = 0 \tag{3}$$
$$0 \leqslant \alpha_i \leqslant c, i=1,2,\cdots,l,$$

其中:$\boldsymbol{\alpha}=(a_1,a_2,\cdots,a_l)^{\mathrm{T}}$,$\alpha_i$ 是公式(1)中不等式约束 $y_i(<w,\varphi(x_i)>-b)\geqslant 1$ 对应的拉格朗日乘积因子;Hessian 矩阵 $\boldsymbol{Q}$ 是半正定的,

$$Q_{ij} = y_i y_j <\varphi(x_i),\varphi(x_j)> \geqslant y_i y_j k(x_i,x_j);$$
$$e=(1,1,\cdots,1)^{\mathrm{T}}。$$

求解上述规划问题,得到一个二分类器:

$$u(x) = \sum_{i=1}^{l} a_i y_i k(x_i,x) - b, \tag{4}$$

其中,$y(x)=\mathrm{sign}(u(x))$。

若 $a_i=0$,样本 $\boldsymbol{x}_i$ 称为非支撑向量;若 $a_i>0$,$\boldsymbol{x}_i$ 称为支撑向量;若 $a_i=C$,$\boldsymbol{x}_i$ 称为有界支撑向量;若 $0<a_i<C$,$\boldsymbol{x}_i$ 称为非有界支撑向量。

式(1)-(3)都是凸规划问题,而凸规划的局部最优解即全局最优解。因此,SVM 方法避免了神经网络等方法存在的局部最优解问题。SVM 方法具有清晰的几何含义,在式(1)和(2)分别等价于求解特征空间中2类训练样本形成的2个凸包或者缩小的凸包之间的距离。几何方法就是利用这种解释,将 SVM 的训练问题转化为经典的几何问题。作为一个 QP 问题,式(4)的求解在理论上并不困难。对于一些特殊情况,通常通过迭代逐渐逼近最优解。当样本较小时,可直接利用已有的算法,如内点算法,梯度投影法等[5]。

对于非线性问题,可在高维内积空间构造最优超平面,进行内积计算。根据泛函的有关理论,只要一种核函数 $R(x_i,x_j)$ 满足 Merce 条件,它就对应某一变换空间的内积。因此,采用适当的核函数 $R(x_i,x_j)$ 就可实现某一非线性变换后的线性分类。简单地说,支撑向量机就是首先通过内积核函数将输入空间变换到一个高维空间,然后在这个空间求最优分类面[6]。

SVM 中不同的内积核函数形成不同的算法,常用的核函数有:

径向基核函数

$$k(x_i,x_j) = \exp\left\{-\frac{\|x_i-x_j\|^2}{\sigma^2}\right\}, \tag{5}$$

多项式核函数 $k(x_i,x_j)=[a(x_i\cdot x_j)+1]^q$,

神经网络核函数 $k(x_i,x_j)=s(a(x_i\cdot x_j)+t)$,其中 s 是 Sigmiod 函数,a、t 为常数。

2 SVM 在给水管网实时故障诊断中的应用

因管体腐蚀、气温变化、管道周围土体变形及产品质量等因素,城市给水管网常常发生管道渗漏和爆裂事故,造成资源的巨大浪费[7]。因此,水管爆裂预测与诊断一直是给水管网安全运行研究的重要课题。目前常采用的诊断方法是通过测试相邻水网节点的水压和流量数据,运用人工神经网络技术来诊断水管爆裂的位置、故障程度和故障影响范围[8],但是,人工神经网络方法在解决水网运行中非线性数据的映射与收敛问题中存在缺陷。对此,结合支撑向量机在解决非线性问题中的优势,通过设计实验模型对水网节点运行水压与流量数据进行测试,选取适当的特征向量,利用支撑向量机 SVM 进行水管爆裂位置的诊断,通过算例对方法进行验证。

2.1 实验模型的设计

对于一个给定的实际给水管网,可充分利用现有的 SCADA 系统采集实际的监测数据,进行分析研究。但对故障问题来讲,由于采集到分布比较好、范围足够大的实际数据是非常困难的,全部采用实际数据也不便于分析研究,因此通过构建实验模型取得数据来补充实际数据的不足,无论对于研究和实际应用都是非常必要的。作为水网故障诊断方法的研究,本文着重于介绍方法,因此全部采用实验数据,即采用给水管网局部破坏状态下水力分析来产生所需数据。

构建一平面管网,1 个水源, 25 个节点, 40 个管段,整个管网在一个水平面上,以方便试验数据的测量,在每根管段上设置阀门用来调节过水流量。节点上采用水龙头来模拟用户用水,接点处设一个水压的测点,节点水压用压力传感器量测,为了加快传感器的读数速度及提高读数精度,在传感器前采用稳压器将不断脉动的水压平稳下来,节点流量采用体积时间法测量。平面管网如图 1 所示。

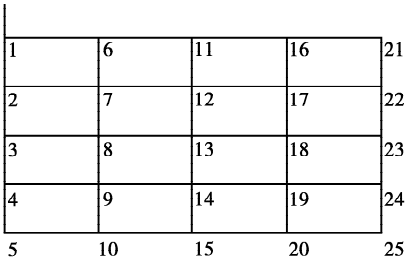

图 1　实验管网平面布置图

首先进行管网正常情况下的水力分析, 可以求得管网所有节点的水压。当管网某点出现爆裂现象前后,随着故障程度不同,节点水压出现变化,通过传感器可以测出。由于在实际给水管网故障诊断时主要需要找出故障位置,从而进行及时的修复,因此在研究时主要针对故障位置的确定进行研究。实验测试时,记录了较多的测试数据,表 1 给出水网局部漏水后 1 次测试数据。

表 1　管网局部故障后水网运行工况数据表

节点号	爆管后水头/cm	基准水头/cm	水头差值/cm	水头差值/基准水头
1	46	71	25	0. 357
2	54	68	14	0. 206
3	57	66	9	0. 136
4	61	67	6	0. 089
5	58	65	7	0. 108
6	53	65	12	0. 185
7	47	58	11	0. 189
8	52	65	13	0. 200
9	49	63	14	0. 222
10	64	73	9	0. 123
11	61	69	8	0. 116
12	62	70	8	0. 114
13	57	69	12	0. 174
14	58	76	18	0. 237
15	53	74	21	0. 284
16	49	69	20	0. 290
17	50	65	15	0. 231
18	57	71	14	0. 197
19	55	68	13	0. 191
20	58	72	14	0. 194
21	49	65	16	0. 246
21	52	67	15	0. 224
23	54	64	10	0. 156
24	57	66	9	0. 136
25	53	68	15	0. 221

在获取给水管网正常运行和局部破坏后水网运行工况数据,运用支撑向量机(SVM)进行故障位置的诊断。

2.2　特征相量的选取

水网故障诊断过程是一个模式识别过程,其诊断精度和可靠性很大程度上取决于故障特征量的选择,对诊断特征量的基本要求是能够很好的区分诊断对象的正常状态和所有可能的故障状态,即识别能力比较强。在给水管网故障诊断中, 任一单元(管道)或节点的故障(破坏)会影响给水管网其它所有单元或节点的水压,故障程度不同,该影响不同;而且同一故障对不同位置单元或节点的影响不同。基于这一出发点,以对给水管网故障最灵敏的水压为参数,根据节点水压在管网局部破坏前后变化来诊断给水管网的故障位置实际上是一非线性识别问题,需要建立故障与节点水压变化之间的对应关系。因此,选取识别能力较强的节点水压参数作为特征相量。

2.3　核函数参数的确定

支撑向量机在实际应用中关于参数选择的问题还没有很好的解决, 比如多项式学习机器的阶数问题,径向基学习机器中的函数宽度问题, 以及 Sigmiod 机器中函数的宽度和偏移问题等, 统计学理论目前对这些问题只是给出了一些建议和解释。例如对于径向基核函数

$$K(x_i,x_j) = \exp\{-\gamma \| x_i - x_j \|^2\}。 \qquad (6)$$

其中 x_i,x_j 分别表示两个训练矢量;γ 表示径向基的宽度, 是惟一需要人为设定的参数。由于该核函数是负指数函数, 因此其指数值不可以太大, 否则核函数对指数的变化不敏感。为此, 令

$$E(\gamma \| x_i - x_j \|^2 = 1,$$

可取

$$\gamma = \frac{1}{E(\| x_i - x_j \|^2},$$

E 表示数学期望,上述公式只是经验公式。实验表明, 据此得到径向基核函数宽度的效果基本上不错, 通常需根据实际数据来仔细调节这些参数,以使系统的性能更好。

利用 SVM 方法进行故障诊断时,为了得到较好的精度,必须精心选择 SVM 的超参数。在采用径向基核函数的 SVM 中,超参数包括如下几项:1)正则化参数 C 决定着模型复杂度与训练误差在目标函数中的比重;2)径向基核函数的宽度参数 γ 隐含的定义了从输入空间到高维特征空间的非线性映射, 决定了特征空间的结构,因而控制了最终解的复杂性。笔者采用交叉有效性验证方法来确定它们。

2.4　训练与测试

在选取训练样本时,采用以下方法:

首先,测量出正常运行(即流量分配相对平均,稳定, 无任何爆管点) 时节点水力参数向量 $\boldsymbol{Q}_m(m=1,2,\cdots,M)$ 作为基准向量. 然后相应提取 $r(r=1,2,\cdots,R)$ 点故障后的节点水力参数向量 $\boldsymbol{Q}'_m$

求偏差系数向量

$\boldsymbol{d}_m = (\boldsymbol{Q}_m - \boldsymbol{Q}'_m)/\boldsymbol{Q}_m](m = 1,2,\cdots,M)$,

向量 $\boldsymbol{d}_m$ 作为样本输入,r 点故障作为样本输出。训练样本和测试样本按以上方法由实验测试和计算后取得。

选取给水管网正常运行和故障状态的数据各100组作为训练样本,80组作为测试样本,计算出相应的指标数值作为特征矢量送入分类器中进行训练或测试。

表2 样本的输入与输出关系表

样本状态	样本数据
输出(r 点故障)	$P_1,P_2,\cdots,P_r$
输入(r 点故障前后节点水力差)	$d_1, d_2,\cdots,d_m$

选用径向基核函数分类器,相应的参数 $\sigma = 8.7, C = 100$,采用的程序为:

SVMTorch(http://www.idiap.ch/learning/SVMTorch.html)。

3 算例分析

以多次实验模型测试数据为基础,对基于SVM的给水管网故障诊断方法进行测试,并与神经网络(ANN)法在相同实验测试数据的前提下进行比较,以证明其有效性。

选取平均相对误差(Δ_{MRE})和均方根误差(Δ_{RMSE})为性能指标,定义如下

$$\Delta_{\mathrm{MRE}} = \frac{1}{n}\frac{1}{r}\sum_{i=1}^{n}\sum_{j=1}^{r}|x_{i,j} - \bar{x}_{i,j}|, \quad (7)$$

$$\Delta_{\mathrm{RMSE}} = \left(\frac{1}{n}\frac{1}{r}\sum_{i=1}^{n}\sum_{j=1}^{r}(x_{i,j} - \bar{x}_{i,j})^2\right)^{\frac{1}{2}}, \quad (8)$$

其中 n 为诊断次数,r 为管网节点数,$x_{i,j}$ 表示第 i 次第 j 点处故障诊断值,$\bar{x}_{i,j}$ 表示实际故障值。发生爆管故障时 $x_{i,j}$,$\bar{x}_{i,j}$ 的值为1,未发生爆管故障 $x_{i,j}$,$\bar{x}_{i,j}$ 的值为0。

在相同实验数据基础上,运用神经网络(ANN)方法进行给水管网故障诊断,通过大量计算比较得知,SVM在实验数据诊断的 Δ_{MRE} 为1.3%,SVM的故障诊断精度平均比ANN提高了0.6%左右。

4 结 论

由于在城市的实际给水管网中采用传统的水力学方法存在着相当多的困难,如管网运行工况的不确定性、管段的水力特性特殊不易测定等等,都对管网的运行工况分析形成了障碍。而笔者采用支撑向量机的方法则可以避开这些障碍,只对运行工况监测信息所表示出来的模式进行辨别,从而可以发挥其一定的功能。

支撑向量机是在统计学理论的基础上发展起来的一种新的学习算法,它是专门研究小样本情况下的学习规律,解决了实际问题中样本有限的问题,比神经网络能更好的解决小样本情况下过学习的问题。将支撑向量机方法用于给水管网故障诊断,实际算例表明,在相同实验测试数据前提下,SVM方法的诊断精度优于神经网络法。

笔者通过构建实验模型对所有节点水力数据进行测试,在实际城市给水管网中,管网范围大,不可能对每个节点设置传感器采集数据,如何对整个城市给水网进行故障监测区域划分以及在一个故障监测区域内如何设置监测点,是笔者将本方法进行实际应用的关键所在。

参考文献:

[1] 许建强. 基于遗传算法的支撑向量机的特征选取[J]. 计算机工程, 2004, 30(24): 1-3.
XU J Q. Feature selection for svm based on genetic algorithm[J]. Computer Engineering, 2004, 30(24): 1-3.

[2] ANGULO C, PARRA X, CATALA A. A support vector machine for multi-class classification [J]. Neurocomputing Volume, 2003, 55: 57-77.

[3] VAPNIK V N. The nature of statistical learning theory[M]. New York: Springer Verlag, 2001.

[4] GUO G D, LI S Z. Content based audio classification and retrieval by support vector machines[J]. IEEE Transon Neural Network, 2003, 14 (1): 109-115.

[5] 徐勋华. 支撑向量机的多类分类方法[J]. 微电子学与计算机, 2004, 21(10): 149-152.
XU X H. Support vector machines for multi-class classification[J]. Microelectronics and Computer, 2004, 21(10): 149-152.

[6] 郭明. 一种基于ICA-SVM的故障诊断方法[J]. 中南工业大学学报, 2003, 34(4): 447-449.
GUO M. The method of independent component analysis and support vector machine based fault diagnosis[J]. Journal of Central South University of Technology, 2003, 34(4): 447-449.

[7] LIGGETT J A, CHEN L C. Inverse transient analysis in pipe networks [J]. Journal of Hydraulic Engineering, ASCE, 1994, 120: 934 - 955.

[8] 梁建文. 给水管网故障实时诊断方法[J]. 水利学报, 2001 (12):4 0-46.
LIANG J W. On-line fault diagnosis of water distribution systems[J]. Journal of Hydraulic Engineering, 2001 (12): 40-46.

(编辑 赵 静)