DOI:10.11705/j.issn.1672-643X.2017.03.30

基于支持向量机决策树的泵站稳态工况运行下状态识别

洪 建,李臣明,高红民
(河海大学 计算机与信息学院, 南京 210098)

摘 要:根据统计学理论提出一种基于决策树支持向量机的泵站状态识别方法,支持向量机是基于统计学理论发展而来的学习方法,在处理小样本,非线性,高维数的问题上较为出色。泵站系统数据往往维度较高,通过主成分分析将数据压缩降维,利用处理过后的数据对三级支持向量机进行训练结合决策树建立泵站运行状态判别模型,进行泵站稳态运行下的状态识别。实验表明:该方法优点是训练时间短,识别准确度高,具有较强鲁棒性。

关键词:泵站; 状态识别; 支持向量机; 决策树
中图分类号:TV93; S277.9 **文献标识码**:A **文章编号**:1672-643X(2017)03-0163-05

State recognition of the running pump station based on the Decision – tree SVM classifier

HONG Jian, LI Chenming, GAO Hongmin
(*College of Computer and Information*, *Hohai University*, *Nanjing* 210098, *China*)

Abstract: A Decision – tree SVM classifier is applied to the state recognition of the running pump station based on statistical learning theory(SLT). SVM is a novel machine learning method based on SLT and powerful for the problems with small sample, nonlinear and high dimension. The data of pump station system tends to have higher dimension, and the data is dimensioned down by principal component analysis. The Decision – tree SVM classifier, trained with the sampling data from the above dealing process and forming an identification model, identifies the state of the pump station. The test results show that the proposed classifier has an excellent performance on correcting ratio and training speed.

Key words: pump station; state recognition; support vector machine (SVM); decision – tree

1 研究背景

泵站在抗洪、灌溉、供水和改善生态环境等方面发挥着极为重要的作用。目前,关于大型调水工程优化调度研究,往往侧重于工程运行成本,而忽略机组维修成本等其他涉及到工程运行的指标[1]。如果单纯强调经济目标,可能造成更大损失,比如,考虑时段电价成本和供水总量的优化调度,在可能采取更为频繁的机组启、停,或更为频繁的叶片调节等运行工况变化,增加部件耗损、缩短机组使用寿命、增加故障率,进而增加了维修成本[2]。泵站是一种随机、缓变/瞬变、多模态、多要素耦合、非线性相关的大型系统,对于这个复杂非线性系统,安全运行是调度运行过程中首先要考虑的问题,因此从泵站整体出发,利用测量数据的相关度和冗余度结合各方面技术对运行参数进行计算处理获得泵站系统级别实时状态信息,及时判断泵站运行安全状态对泵站后期系统运行、维护、管理以及科学合理调度至关重要。

解决识别问题的方法有很多,已经有了很多这方面的研究,文献[3-5]分别介绍了模糊技术、专家系统、遗传算法都在机械、电气等方面有较好的识别结果;文献[6]介绍了基于数据挖掘和模糊理论的电力设备状态评价方法;文献[7]中构造了电力系统状态评估参量模型。但是这些方法往往存在一些缺点,如神经网络训练时间比较长,存在局部最优问题[8];当系统的模型未知或者非常复杂时,参量

收稿日期:2017-02-27; **修回日期**:2017-03-24
作者简介:洪 建(1993-),男,安徽宣城人,硕士研究生,从事复杂系统建模研究。
通讯作者:李臣明(1969-),男,江苏南京人,博士,副教授,硕士导师,从事智能信息处理方面的教学与研究工作。

模型构建非常困难已不再适用[9];模糊技术在状态评估识别领域运用极为广泛,但模糊技术在样本学习上存在巨大缺陷[10];支持向量机分析过程中,如果不对大量的识别因素处理,会出现样本维度高,建立模型较为复杂,学习时间较长等缺点[11]。

支持向量机是在有限样本统计学习理论上衍生而来的学习方法,其克服了合理的网络结构难以确定等问题,在解决样本有限、维度很高、非线性等学习问题方面效果显著,很大程度地提高了方法的泛化能力。支持向量机在字体以及人脸识别、故障分类[12-14]、复杂非线性系统识别[15]等方面都取得了显著效果。本文将支持向量机运用到泵站运行状态识别中,考虑到泵站是多要素耦合的大型时变系统,测量数据存在较强关联性,而且数据维数较高[16]。本文通过 PCA(主成分分析),对原始数据进行数据压缩和数据特征提取,尽可能消除数据间的相关性,提取出包含样本数据足够的主要成分,解决对大量数据处理问题。

2 数据来源与方法

2.1 数据来源

南水北调东线输水系统主要以大型梯级泵站(群)为核心,从长江至东平湖,须新建和改造 50 余座大型泵站(其中江苏境内 18 座),分 13 级提水,提水扬程 65 m,年均运行时间达 5 000 h 以上。东线一期工程多年平均抽江水量 89×10^8 m^3(江苏、山东、安徽 3 省多年平均供水量分别约为 19.25×10^8、13.53×10^8、3.23×10^8 m^3),因此确保泵站安全稳定运行至关重要。

本文以南水北调洪泽湖泵站运行中的历史数据以及泵站(群)SIS 平台及实时状态评估系统仿真数据开展泵站运行状态评估关键技术,数据包括稳态工况下的扬程、流量、轴功率、效率、转速、叶片角度、电机输入功率、泵壳温度、定转子温度等上百监控量,图 1 为南水北调东线工程泵站实时监控。



图 1 南水北调东线工程泵站运行状态监测

2.2 研究方法

2.2.1 主成分分析 主成分分析是统计学的一种方法,假设有 $X=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 为 n 维随机矢量,并且每一维的均值为 0,即 $E(x_i)=0(i=1,2,\cdots,n)$ 它的主成分 Y 满足下面两个条件:

(1)Y 必须由正交变换 U 得到,$Y=UX$,对于 U 中的 u_{ij} 满足 $\sum_{j=1}^{p}u_{ij}^2=1$, $\sum_{j=1}^{p}u_{ij}u_{kj}=0$。

(2) 正交变换 U 一定是具有最大方差的,y_2 与 y_1 独立,并且在所有与 y_1 独立的随机变量中 y_2 具有最大方差,同理 y_3 等等。可以得到了 Y 的各各分量 $y_1,y_2,\cdots,y_n$ 可以称为 X 的第 i 个主分量。

假设 R 为随机矢量 X 的协方差矩阵,可以理解它的特征值与特征矢量对 (λ_1,l_1), (λ_2,l_2),$\cdots$, (λ_n,l_n),其中,$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n\geqslant0$。特征值的大小就可以反映对应的主成分样本的离差程度,并且可以确定主成分的个数,一般通过累计贡献率来确定。对于第 i 个元素的累计贡献率为:

$$\beta_i=\sum_{k=1}^{i}\lambda_k/\sum_{k=1}^{p}\lambda_k \tag{1}$$

在实际运用中根据原样本空间的绝大部分信息和实际情况决定要选取主成分的个数。

2.2.2 支持向量机 SVM 的原理是寻找一个满足分类要求的最优分类超平面通过对历史特征数据(即特征与目标值)不断地学习、结果对比、再学习,最终得到合理的判别规则,即产生分类模型[17]。分类模型产生后,即可利用分类模型完成所需要的分类任务,将需要分类的向量输入建立好的分类模型,实现分类。

假设训练样本为(x_i,y_i),$i=1,2,\cdots,l$, $x_i\in R^n$, $y_i\in\{\pm1\}$。二维空间中线性判别函数可以表示为:

$$g(x)=\omega^{\mathrm{T}}x+b \tag{2}$$

对应的超平面方程表示为:

$$\omega^{\mathrm{T}}x+b=0 \tag{3}$$

所有被正确分类且具备分类间隔的样本点都必须满足:

$$y_i\cdot(\omega^{\mathrm{T}}x_i+b)\geqslant1,\ i=1,2,\cdots,l,\ y_i\in\{\pm1\} \tag{4}$$

式中:y_i 为第 i 个样本的类别标签;ω 为权系数。

此时,寻找最优超平面问题变成在式(4) 的约束之下,寻求 $\max(1/\|\omega\|)$,问题转化为求解:

$$\min\phi(\omega)=\frac{1}{2}\|\omega\|^2=\frac{1}{2}(\omega^{\mathrm{T}}\omega) \tag{5}$$

引入 Lagrange 因子 α_i,对上式求解可得最优解 $\omega^*=\sum_{i=1}^{l}\alpha_i^*y_ix_i$,令 $\alpha_i^*\neq0$ 的训练样本即为支持向

量,最优分类超平面的权系数向量 ω^*为支持向量的线性组合。b^* 可由约束条件 $y_i(\omega^{\mathrm{T}}x_i + b) - 1 = 0$ 求解得到。

对于非线性的数据分类问题基本思想可以概括为:通过核函数实现非线性变换将输入空间映射到一个高维空间,然后在这个新空间求取最优分类平面。常用的核函数主要有以下 4 种:

(1) 线性(linear) 核函数:

$K(x_i,x_j) = (x_i \cdot x_j)$

(2) 多项式(polynomial) 核函数:

$K(x_i,x_j) = [(x_i \cdot x_j) + 1]^q$

(3) 径向基(RBF) 核函数:

$K(x_i,x_j) = \exp\{-|x_i - x_c|^2/(2\sigma)^2\}$

(4) S 形(sigmoid) 核函数:

$K(x_i,x_j) = \tanh(\upsilon(x_i \cdot x_j) + c)$

对于处理多分类问题有两种方法:直接法和间接法。直接法直接计算出用于多分类的分类函数,该方法计算量很大,求解过程复杂,花费时间多,实现起来困难,在解决实际问题中不适用。间接法主要通过组合多个二分类 SVM 来实现 SVM 的构建,最常用的两种方法为一对多和一对一法[18]。

2.2.3　支持向量机结合决策树的方法　K 分类问题($K > 2$) 可以最后按二分类问题解决,存在 K 类可分的问题,则这 K 类两两可分。反之,K 分类问题中,在知道其中类别两两可分情况下,通过组合,通过二分类来解决 K 分类问题。

设对 K 类样本问题构建决策树,一种类对应树的一个叶结点,一个支持向量机分类器为每个度为 2 的非叶结点。所以决策树共有 $2K-1$ 个结点,叶结点个数为 K,支持向量机分类器为 $2K-1$ 个。支持向量机决策树通过实验得出了以下的结论:

(1) 一对一方法的错误率最小,与支持向量机决策树方法的错误率相当。

(2) 一对一方法训练速度较快,支持向量机决策树稍慢,一对多速度最差。

(3) 支持向量机决策树测试时间最短。测试按照层级进行,对单个样本可能达不到决策树最底层,所以测试速度最快。

3　基于支持向量机决策树的泵站状态识别

3.1　数据降维

首先对各样本数据进行 PCA 处理,作为 SVM 的输入,首先对原来的 N 个量测量的样本级标准化,根据标准化求出其方差矩阵 R,以及特征向量和特征值,根据给定的 α 值和贡献率确定主成分。图 2 为前 10 个主成分贡献率对比图,由图 2 可见,第一主成分贡献率达到 0.7866,给定 $\alpha = 0.95$,经过计算 $\beta_7 = 0.9567 \geqslant 0.95$。

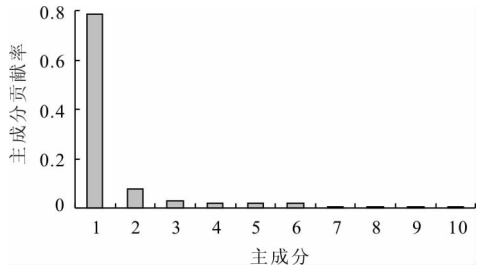


图 2　主成分贡献率

选择前 7 个主成分作为新的样本,作为 SVM 的输入即效率,扬程,流量,轴功率,转速,叶片角度,电机输入功率为输入量。

y^i 为处理过后的特征数据($i = 1,2,3...7$),这样组成了 1 个 7 维的特性向量 $Y = [y^1,y^2,y^3,y^4,y^5,y^6,y^7]^{\mathrm{T}}$。

3.2　网络训练

根据泵站运行状态特点与支持向量机结合决策树理论,本文依据南水北调东线工程泵站运行指标集采用 3 级支持向量机对泵站运行状态进行 4 类(优,良,中,差)划分,确定运行状态以便于后期维护检修。基于支持向量机分类器的状态识别基本原理如图 3。

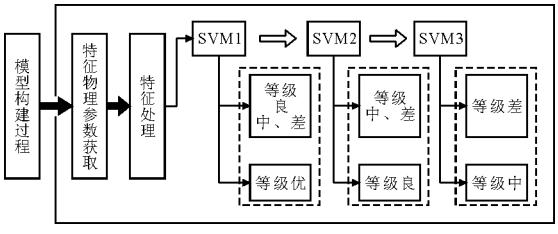


图 3　基于支持向量机决策树分类器泵站运行状态分类原理

对第一级支持向量机 SVM1 进行训练和分类,将系统运行状态为良好,中等,差 3 个级别看作一大类标记为 -1,运行状态为优的视为一类标记为 1,训练第一级模型时将样本数据全部输入,良,中,差,标记为 -1,优标记为 1 进行模型训练;第二级支持向量机 SVM2 进行训练和分类,将系统运行差,中等两个级别看作一类标记为 -1,运行为良好标记为 1,将样本里状态为中等、差的标记为 -1,良好样本标记为 1 全部输入进行模型训练;第三级支持向量机 SVM3 进行训练和分类,最后一级系统运行级别

为差标记为 -1,运行状态中等标记为 1,将样本状态为中等标记为 1,样本中状态为差的标记为 -1 全部输入进行最后一级模型训练,完成三级支持向量机模型训练。

3.3 实验结果分析

采用泵站运行时包括流量,扬程,轴功率等 7 维特征向量总共 50 组作为历史数据作为训练集,其中 25 组为优,15 组为良,5 组为差(洪泽站年开机时间较短,故障数据缺乏,根据泵站运行特性曲线,模拟出 5 组故障数据)。

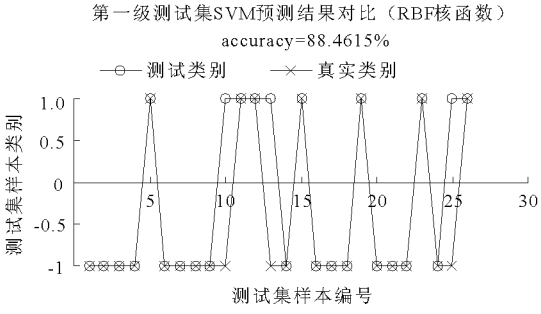


图 4 第一级分类结果

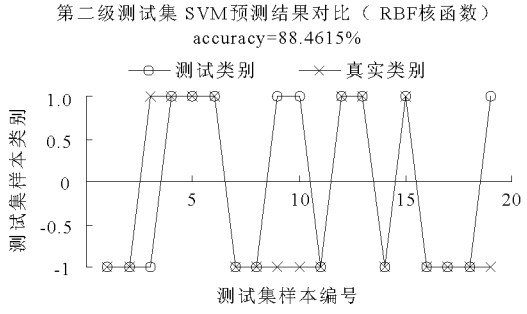


图 5 第二级分类结果

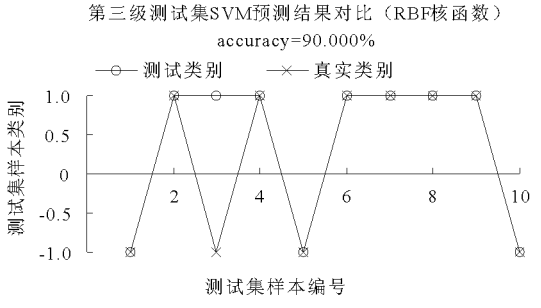


图 6 第三级分类结果

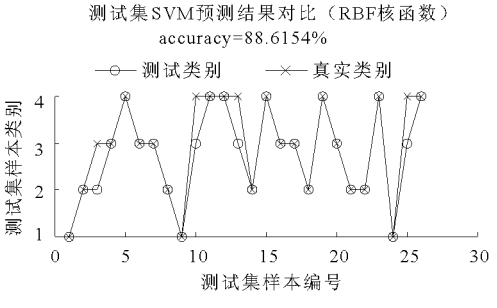


图 7 基于二叉树的多级支持向量分类器分类结果

实验结果表明:三级支持向量机识别了 4 种状态,对于泵站复杂系统识别正确率达到 88.6%,具有较高的识别率。

另外通过神经网络的识别方法与支持向量机识别方法进行对比,更好地评价多级支持向量机分类器的泵站系统状态识别,采用某单隐含层 BP 神经网络,包括 7 个输入节点、46 个隐藏节点和 4 个输出节点,对泵站运行状态识别问题进行对比,并且采用完全相同的数据处理方法、训练样本和测试数据,对于训练采用动态法改进的快速学习算法并且目标误差分别选取 0.05、0.01 和 0.005 与提出来的多级支持向量机进行对比实验结果见表 1。

由表 1 可知,与 BP 神经网络识别方法相比,支持向量机方法从分类精度上优于 BP 神经网络并且训练时间显著小于 BP 神经网络。另外在 BP 神经网络试验中随着训练目标误差从 0.05 到 0.005,识别正确率有一定提高,当减小到一定层度后再减小,反而识别精度出现恶化,这就是 BP 神经网络过拟合而引起的问题,同时训练时间过度增加,BP 神经网络会出现较大的干扰性的识别结果。本文提出的基于主成分分析与多级支持向量机结合的识别方法则不会存在这样的问题。

经过上述描述,训练出 3 组支持向量机(SVM1,SVM2,SVM3),采用另外 26 组历史数据作为测试数据,其中:7 组为优秀,7 组为良好,6 组为中等,6 组为差。

根据 SVM 基本分类过程,在构造第一级分类器模型时,认为优和良中差分为两类各标记为 +1,-1。当测试数据输入第一级分类器时结果如图 4,真实类别曲线用叉线代表,测试曲线用圈线代表;第二级、第三级结果如图 5、图 6,最后对 3 级分类结果汇总(1,2,3,4 分别代表上述 4 类数据)如图 7。

表 1 神经网络与多级支持向量机分类器训练与识别结果对比

识别方法	训练时间/s	目标误差	识别正确率/%	干扰识别率/%
SVMs	小于 1	0.01	88.6	0
BP1	约 100	0.05	76.0	5.0
BP2	约 200	0.01	84.6	7.5
BP3	约 400	0.005	80.7	5.0

4 结 论

支持向量机能够在样本数量一定的条件下通过学习,较好地解决小样本,高纬度,非线性问题,本文讨论了基于支持向量机结合二叉树分类器的泵站稳态运行下的状态识别方法。实验结果表明该方法相比其他分类方法有以下优点:

(1)训练时间短,实时性较好;

(2)不存在局部极小值问题,支持向量机方法还具有对训练样本要求较低的优势,这样就加强了该方法的实时性;

(3)鲁棒性强,识别率正确率相对较高。因此可以很好地应用于泵站系统状态的识别,判断泵站运行安全状态对泵站后期系统运行、维护、管理以及科学合理调度至关重要,避免了重大故障对泵站运行生产造成的严重后果。本文将支持向量机结合二叉树的方法运用在泵站状态识别,目前还有一些问题需要研究包括核函数的选择和参数的优化。

参考文献:

[1] 梁 兴,刘梅清,燕 浩,等. 基于 Pareto 最优解的梯级泵站双目标优化调度[J]. 武汉大学学报(工学版),2015,48(2):156 - 159 + 165.

[2] 梁 兴. 基于瞬变流计算与安全性分析的大型梯级泵站优化调度研究[D]. 武汉:武汉大学,2014.

[3] 张 勇,李 昕,刘君华,等. 数据融合新技术在识别变压器油中四种特征气体的研究[J]. 中国电机工程学报,2001,21(8):10 - 14.

[4] 王由华,刘振娟,李宏光. 混合型集成神经网络故障诊断方法研究[J]. 北京化工大学学报(自然科学版),2003,30(1): 65 - 68.

[5] Li Jun, Zhang Qiang, Wang kai, et al. Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on support vector machine[J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2016,23(2):1198 - 1206.

[6] Zhong Wenhui, Sun Yixue, Xu Min, et al. State assessment system of power transformer equipments based on data mining and fuzzy theory[C]//. 2010 International Conference on Intelligent Computation Technology and Automation, Changsha, 2010: 372 - 375.

[7] 曹建东,吴 姜,蔡泽祥,等. 电力系统二次设备状态参量模型的构建[J]. 南方电网技术,2012,6(3): 107 - 110.

[8] 杨 宇. 基于 EMD 和支持向量机的旋转机械故障诊断方法研究[D]. 长沙:湖南大学,2005.

[9] 朱霄珣. 基于支持向量机的旋转机械故障诊断与预测方法研究[D]. 保定:华北电力大学,2013.

[10] 李清泉,高德荃,杨必胜. 基于模糊支持向量机的城市道路交通状态分类[J]. 吉林大学学报(工学版),2009,(S2):131 - 134.

[11] 张学工. 模式识别[M]. 北京:清华大学出版社,2002.

[12] Lü Xiaoyu, Zhang Zhiqing. A new method of shot boundary detection using support vector machine[C]//. Proceedings of International Conference of China Communication and Technology (ICCCT2010), Beijing, China, 2010.

[13] Xue Xinhua, Yang Xingguo, Chen Xin. Application of a support vector machine for predictionofslopestability[J]. ScienceChina (TechnologicalSciences), 2014, 57 (12): 2379 - 2386.

[14] 时文刚,刘树林,张嘉钟,等. 基于支持向量机的往复泵泵阀故障诊断方法[J]. 机械强度,2002,24(3):362 - 364.

[15] 段新生. 证据决策[M]. 北京:经济科学出版社. 1996.

[16] Jenkins L, Khincha H P, Shivakumar S, et al. An application of functional dependencies to the topological analysis of protection schemes[J]. IEEE Transactions on Power Delivery, 1992, 7(1):77 - 83.

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

[18] 杨 雪. 支持向量机多类分类方法的研究[D]. 哈尔滨:哈尔滨工程大学,2006.