

支持向量机在高压绝缘子污秽程度评定中的应用^{*}

焦尚彬¹, 刘 丁¹, 郑 岗¹, 张 青^{1,2}, 王洪江¹

(1 西安理工大学 信息与控制工程研究中心, 陕西 西安 710048;

2 杨凌职业技术学院 电力系, 陕西 杨凌 712100)

[摘 要] 高压绝缘子污秽程度的评定可以表述为多类模式识别问题, 由于影响绝缘子表面污秽状况的因素复杂, 并且实际获得的样本数据有限, 因此传统的智能方法往往达不到工程应用的精度要求。研究利用支持向量机在解决小样本、非线性及高维识别中的优势, 将可用于多类模式识别的DA G SVM s模型用于高压绝缘子污秽程度的评定。该模型易于实现, 且能够找到模式间的最优分类超平面, 泛化能力较高。用SVM评定高压绝缘子污秽程度时, 可以综合考虑不同污秽程度绝缘子泄漏电流的电气特性、环境参数与绝缘子污秽程度之间的非线性关系, 从而实现具有极大模糊性的绝缘子表面污秽程度的评定。结果表明, 此方法对解决绝缘子污秽程度的评定问题具有良好的适应性和实用性。

[关键词] 支持向量机; 高压绝缘子; 污秽程度评定; 模式识别

[中图分类号] TM 732

[文献标识码] A

[文章编号] 1671-9387(2006)01-0142-05

暴露在污秽条件下的绝缘子表面会沉积污秽, 污秽层受潮后使绝缘子的外绝缘能力下降, 故常引起污闪事故, 从而严重影响供电系统的可靠性。绝缘子泄漏电流在线监测系统通过监测绝缘子表面泄漏电流的电气特征量和环境参数及泄漏电流超过某一阈值作为污秽过度的判断依据。但实际研究表明, 泄漏电流的大小除了与绝缘子表面的污秽程度有关外, 同时还受到环境温度、湿度等因素的影响, 并且各因素间存在复杂的非线性关系。因此, 单一的阈值评定方法存在一定的缺陷。

焦尚彬等^[1]和解建军等^[2]分别利用模糊逻辑方法和模糊专家系统对绝缘子污秽程度进行了评定。但这2种方法的准确性受模糊逻辑规则完备性的影响较大, 且模糊逻辑规则的制定与修改需要大量的数据样本和现场经验。张寒等^[3]利用神经网络来预测基于气象因素的绝缘子等值附盐密度, 从而确定绝缘子表面的污秽程度。但是神经网络的结构过于复杂, 需要估计的参数相对于较少的数据样本显得太多, 导致所得的神经网络模型对数据过学习, 即泛化能力不够, 从而使预测精度不高。另外, 神经网络的结构难以选择, 也限制了它的应用。

支持向量机(Support Vector Machine, SVM)是在统计学习理论(Statistical Learning Theory,

SLT)基础上发展起来的一种新的分类和回归工具^[4]。其通过结构风险最小化原理来提高泛化能力, 较好地解决了小样本、非线性、高维数、局部极小点等实际问题^[5-6]。SVM这种非线性的数据处理工具在模式识别、函数逼近、回归估计和非线性控制中均能很好的应用。

针对高压绝缘子污秽程度评定问题所具有的样本有限、非线性、维数高的特点, 本研究将SVM用于高压绝缘子污秽程度的评定中, 综合考虑泄漏电流的电气特性和环境参数与绝缘子污秽程度之间的非线性关系, 以期能够实现绝缘子污秽程度的自动评定, 以避免单一阈值评定可能导致的误判。

1 支持向量机简介

经验风险最小化原则一直是解决统计模式识别等统计机器学习问题的基本思想, 在此思想的指导下, 人们主要解决的是如何更好地求取最小经验风险(训练误差最小)。但实践证明, 一味地追求训练误差最小并不能得到最好的泛化能力, 在有些情况下, 训练误差太小反而会导致泛化能力下降, 这在神经网络学习中表现得尤为突出(即过学习问题)。导致该问题出现的一个根本原因是因为传统统计学是一种渐进理论, 它的许多结论都是在样本数目趋向于

^{*} [收稿日期] 2005-07-19

[基金项目] 陕西省教育厅重大产业化资助项目(04jc13)

[作者简介] 焦尚彬(1974-), 男, 四川宣汉人, 讲师, 在读博士, 主要从事电力系统智能状态检测研究。E-mail: jsbzq@163.com

无穷大的条件下得出来的,而在小样本条件下,以传统渐进统计学为理论基础的统计风险最小化原则,并不能很好地实现由贝叶斯决策理论导出的期望统计风险最小化原则。为了克服传统渐进统计理论在小样本统计学习中的不足,Vapnik^[4]建立了统计学习理论。统计学习理论指出,在小样本条件下,只有同时控制经验风险和学习机容量(用VC维衡量),才能获得具有良好泛化能力的学习机。支持向量机正是在这一理论基础上发展起来的,能有效解决有限样本、非线性和高维数的模式识别问题。

1.1 最优分类超平面

支持向量机方法最初是由线性可分类情况下的最优分类超平面(Optimal Hyperplane)提出的。最优分类超平面是指该分类面不但能正确分类,而且能使各类别的分类间隔最大^[7]。

设样本集 (x_i, y_i) 线性可分, $i = 1, 2, \dots, l$, 其中 $x_i \in R^n$ 为输入向量; $y_i \in \{+1, -1\}$ 为输出类别; l 为样本数, n 为输入维数。则判别函数为: $g(x) = (\omega \cdot x) + b$, 分类面方程为: $(\omega \cdot x) + b = 0$, 其中, “ $\cdot$ ”为向量点积, ω 为权向量, b 为阈值。

求最优分类面的问题可以表示为以下优化问题:

$$\begin{cases} \min \psi(\omega) = \frac{1}{2} \|\omega\|^2 = \frac{1}{2} (\omega \cdot \omega) \\ \text{s.t. } y_i(\omega \cdot x_i + b) - 1 \geq 0 \end{cases} \quad (1)$$

该问题可以先转换为其对偶问题再求解:

$$\begin{cases} \max Q(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ \text{s.t. } \sum \alpha_i y_i = 0 \quad \alpha_i \geq 0 \end{cases} \quad (2)$$

这是一个不等式约束条件下的二次函数极值问题,存在唯一解 α^* 。对其求解可以得到最优分类函数 $f(x)$,进而可得到最优分类面方程:

$$\begin{cases} f(x) = \text{sgn} \left[\sum \alpha_i^* y_i (x \cdot x_i) + b^* \right] \\ \sum \alpha_i^* y_i (x \cdot x_i) + b^* = 0 \end{cases} \quad (3)$$

根据Kuhn-Tucker条件,满足 $\alpha^* \neq 0$ 的样本对分类结果起重要作用,称为支持向量。

1.2 支持向量机

对线性不可分的情况,可通过某种事先选择的非线性映射,将原空间的输入向量映射到一个高维特征空间 z ,在这个高维空间中构造最优分类超平

面。但这样容易陷入维数灾难。

根据式(2)和式(3),求解最优分类面只涉及特征空间的内积运算,如果以函数 $K(x, x_i) = \psi(x) \cdot \psi(x_i)$ 来表示特征空间的内积运算,并可以用原空间的输入向量 x 直接计算得到,则最优分类面的求解与特征空间维数无关,避免了维数灾难。

统计学习理论指出, $K(x, x_i)$ 只要满足Mercer条件就可以展开成 $\psi(x) \cdot \psi(x_i)$,即可以作为内积函数(亦称核函数)使用。用 $K(x_i, x_j)$ 代替式(2)中的点积 $(x_i \cdot x_j)$,相当于将原空间变换到某一新的特征空间,对应的最优分类函数为:

$$f(x) = \text{sgn} \left[\sum \alpha_i^* y_i K(x \cdot x_i) + b^* \right] \quad (4)$$

以上的分类函数学习机称为支持向量机。

1.3 支持向量机的模式识别模型

支持向量机的模式识别模型如图1所示。模式识别过程主要包括2个阶段:第一阶段为训练阶段,即根据输入、输出构造训练样本,然后对支持向量机进行训练;第二阶段为识别阶段,即对那些输入已知、输出未知的样本,利用已经训练好的支持向量机模型,根据样本的输入得到其识别输出。

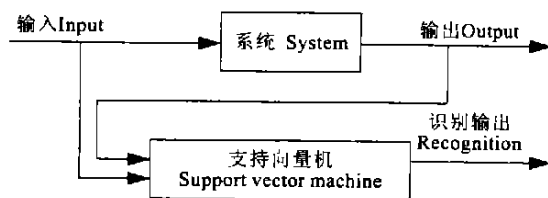


图1 支持向量机的模式识别模型

Fig 1 SVM model for common pattern recognition

1.4 多类模式识别问题

支持向量机直接处理的是两类模式的识别问题,对于多类模式的识别问题,有多种技术可以用来在二分类支持向量机的基础上构造多分类学习器,比如一对一、一对多、有向无环图(Directed Acyclic Graph, DAG)等,从训练速度、分类速度、结构选择与推广性能等方面综合考虑,DAG方法有较好的效果^[8],所以本研究采用DAG方法在二分类基础上构造多分类的支持向量机。

DAG多类SVM分类方法在训练阶段与一对一投票一样,也要构造出每两类间的分类面,即对于类问题有个分类器。但是在分类阶段,该方法将所有分类器构造成一种双向有向无环图^[8]:包括 $K(K-1)/2$ 个节点和 K 个“叶”;其中每个节点为一个分类器,并与下一层的两个节点(或者叶)相连。当对一

个未知样本进行分类时,首先从顶部的根节点开始,根据根节点的分类结果用下一层中的左节点或者右节点继续分类,直到达到底层某个叶为止,该叶所表示的类别即为未知样本的类别。

2 基于支持向量机的绝缘子污秽程度评定

在高压绝缘子污秽程度评定中,当绝缘子表面的污秽程度一定时,对于不同的环境条件,流过受污绝缘子表面泄漏电流的电气特征量是不同的。根据实验室和现场数据,结合专家经验,选用泄漏电流有效值(F_i)、泄漏电流峰值(F_p)、泄漏电流脉冲频度(F_f)、环境湿度(H)、温度(T)等 5 个变量作为模型的输入参数,模型输出为绝缘子的污秽程度。由于污秽程度被划分为污秽正常(NL)、一般污秽(CM)、较严重污秽(MS)和严重污秽(SR)4 个等级,因此需要构造 6 个分类器,由 6 个分类器构造的两向有向无

环图中包括 6 个节点和 4 个叶。

2.1 样本数据来源

样本数据一部分来源于实验室人工污秽实验,另一部分来源于现场。实验室样本数据的获取方法为:根据文献[9]规定的 4 个污秽等级,在每 2 个等级规定的等值附盐密度(ESDD)量之间选择 3 个点,每个点按文献[10]规定的方法在不同的温度、湿度条件下测其泄漏电流的各种电气特征量。现场数据中,泄漏电流的各种电气特征量和环境温湿度数据,来源于已投运到現場的多套变电站和输电线路绝缘子泄漏电流在线监测系统;ESDD 值来源于绝缘子表面的实测结果,部分样本见表 1。表 1 中,各峰值区段泄漏电流脉冲数是指在 5 min 内泄漏电流峰值分别处于 $[20 \text{ mA}, 100 \text{ mA}]$ 、 $[100 \text{ mA}, 250 \text{ mA}]$ 、 $[250 \text{ mA}, 350 \text{ mA}]$ 、 $[350 \text{ mA}, 450 \text{ mA}]$ 、 $[450 \text{ mA}, \infty]$ 范围的脉冲个数。

表 1 绝缘子污秽程度评定的部分样本数据

Table 1 Sample assessment data for insulator contamination condition

绝缘子 Insulator	等值附盐密度/ ($\text{mg} \cdot \text{cm}^{-2}$) Equal salt deposit density	相对湿度/% Relative humidity	温度/ Temperature	泄漏电流 有效值/mA R. M. S of leakage current	泄漏电流 峰值/mA Peak value of leakage current	各峰值区段 泄漏电流 脉冲数 Leakage current pulse time of different section of peak value	污秽等级 Contam in- ation condition
1	0.015	70.0	4	0.15	< 1	0/0/0/0/0	NL
2	0.150	100	14	5.22	< 50	0/0/0/0/0	CM
3	0.300	100	27	12.8	262	0/0/0/0/0	MS
4	0.380	90.0	17	15.9	378	687/112/12/0/0	SR
5	0.450	75.0	32	13.4	354	896/202/67/9/0	SR
6	0.450	100	19	58.0	478	734/176/43/7/2	SR
7	0.013	70.0	-4	0.12	< 1	0/0/0/0/0	NL
8	0.034	84.0	6	0.62	< 4	0/0/0/0/0	NL
9	0.287	100	14	12.6	253	647/102/8/0/0	SR
10	0.026	80.0	17	5.2	< 2	0/0/0/0/0	CM
11	0.142	85.0	31	3.9	< 50	0/0/0/0/0	MS
12	清洁 Clean	70.0	22	0.08	< 0.5	0/0/0/0/0	NL

2.2 评定模型的建立

2.2.1 数据的归一化处理 样本数据的归一化处理。为加快训练速度,利用公式(5)将输入量 F_k 、 H 、 T 归一化到 $[0, 1]$, 即

$$\bar{X}_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \quad (5)$$

式中, $X_{\max}$ 和 $X_{\min}$ 分别为 F_k 、 H 、 T 的最大值和最小值。

泄漏电流峰值的归一化处理。由于脉冲电流峰值大小具有一定随机性,精确地按照其值的大小划分一条曲线没有太大的实际意义。因此,根据试验

及文献[11]和[12]报道的情况,采取分段划分的方式较好。依试验测试结果将其分为 6 个区间,即泄漏电流峰值的归一化处理函数为离散型函数,用公式(6)表示。

$$F_p = \begin{cases} 0.0 & I_p < 20 \text{ mA} \\ 0.2 & 20 \text{ mA} \leq I_p < 100 \text{ mA} \\ 0.4 & 100 \text{ mA} \leq I_p < 250 \text{ mA} \\ 0.6 & 250 \text{ mA} \leq I_p < 350 \text{ mA} \\ 0.8 & 350 \text{ mA} \leq I_p < 450 \text{ mA} \\ 1.0 & I_p \geq 450 \text{ mA} \end{cases} \quad (6)$$

泄漏电流脉冲频度的归一化处理。由于一般情况下绝缘子上的脉冲电流峰值大多数在 20 mA 以下, 只是在临近闪络前的几天内才会产生越来越多的高幅值电流脉冲, 且各种幅值的脉冲电流峰值出现的几率, 及所反映的由污秽导致的外绝缘性能变化情况的能力不同, 如当有 450 mA 以上的电流脉冲出现时, 则意味着闪络即将发生。因此, 对泄漏电流脉冲频度需作如下的归一化处理: 对于各种不同幅值的泄漏电流峰值, 首先计算其在一定时间内 (取 5 min) 的出现次数 n_i , 然后计算各幅值区段泄漏电流脉冲个数与该时间段泄漏电流总脉冲个数的比值, 并根据其重要性分别赋予不同的权值, 通过运算决定最终脉冲频度输入量。

定义的归一化函数为:

$$n = \sum_{i=1}^5 n_i \quad (7)$$

$$F_{f1} = \frac{3n_1}{10n} \quad 20 \text{ mA} \leq I_p < 100 \text{ mA} \quad (8)$$

$$F_{f2} = \frac{n_2}{2n} \quad 100 \text{ mA} \leq I_p < 250 \text{ mA} \quad (9)$$

$$F_{f3} = \frac{4n_3}{5n} \quad 250 \text{ mA} \leq I_p < 350 \text{ mA} \quad (10)$$

$$F_{f4} = \frac{n_4}{n} \quad 350 \text{ mA} \leq I_p < 450 \text{ mA} \quad (11)$$

$$F_{f5} = \begin{cases} 0 & (n_5 = 0) \\ 1 & (n_5 = 1) \end{cases} \quad I_p \geq 450 \text{ mA} \quad (12)$$

$$F_f = F_{f1} \quad F_{f2} \quad F_{f3} \quad F_{f4} \quad F_{f5} = \max(F_{f1}, F_{f2}, F_{f3}, F_{f4}, F_{f5}) \quad (13)$$

将样本数据进行归一化处理后, 选取 200 个样本, 其中 150 个样本作为训练样本集, 其余 50 个样本作为测试样本集。

2.2.2 评定模型的训练 将 150 个训练样本输入到支持向量机中进行样本训练。分别采用两种核函

数作为内积的回旋:

多项式核函数。 $K(x, x_i) = [(x \cdot x_i) + 1]^d$ ($d = 1, 2, \dots$), 其中 d 为多项式的阶数, 这里取 $d = 3$ 。多项式核函数的 SVM 分类器相当于带有超平面参数 γ 的线性判别分析, γ 可以控制对训练数据的拟合度。 γ 的值越大, 分离的超平面离给定的训练样本越近。采用交叉验证的方法选择合适的 γ 值以避免数据的“过拟合”和“欠拟合”, 本研究选择适合各个分类器的最优 γ 值为 1。

$$\text{径向基核函数。} K(x, x_i) = \left\{ -\frac{x - x_i}{2\sigma^2} \right\}^2,$$

其中 $x - x_i = \sqrt{\sum_{k=1}^n (x^k - x_i^k)^2}$, σ 为核宽度。径向基核函数为输入数据的非线性映射提供了很大的灵活性, 这对于复杂的、非线性的、不可分的分类问题是很有用的。使用径向基核函数的 SVM 分类器需要确定 2 个超平面参数, 即 γ 和 σ 。对于 γ 而言, γ 值太小会导致数据的“欠拟合”, γ 值太大会导致数据的“过拟合”, 从而使泛化性能恶化; 对于 σ 而言, σ 值太大会造成数据的“欠拟合”, σ 值太小会造成数据的“过拟合”, 因此同样必须通过交叉验证的方式来对这两个参数加以选取。对于每个分类器, 训练准确率最高时的 γ 和 σ 值是不同的。本研究选择适合各个分类器的最优超平面参数为 $\gamma = 0.7598$, $\sigma = 4.738$ 。

2.2.3 评定模型的测试 对评定模型进行测试时, 每个分类器的训练和测试重复 100 次, 模型参数的选择及对应的训练集和测试集的准确率详见表 2。针对绝缘子污秽程度评定问题, 从训练和测试结果来看, 径向基核函数比多项式核函数具有更高的准确率。

表 2 SVM s 分类器的训练和测试结果

Table 2 Training and testing results of SVM s classifier

模型核函数 Model kernel function	核参数 Kernel parameter			准确率/% Right ratio	
	γ	d	σ^2	训练集 Training set	测试集 Testing set
多项式 Polynomial	1	3		91.02	88.46
径向基 RBF	0.7598		4.738	93.37	91.02

3 结 语

目前, 绝缘子污秽程度的评定主要是基于传统的统计学和神经网络的方法, 传统的统计学方法适用于样本数目逼近无穷大的大样本, 但当样本数目有限或者在处理小样本时, 便难以取得理想效果, 很

难适应泄漏电流的电气特征量、环境参数和绝缘子污秽程度之间复杂多变的关系。而神经网络又存在收敛速度慢、结构选择难和局部极小等问题。本研究尝试将绝缘子表面污秽程度的评定问题转化为一个多类分类问题, 用泄漏电流有效值、泄漏电流峰值、泄漏电流脉冲频度、环境相对湿度和环境温度等参

数作为输入变量,以污秽程度作为输出,利用支持向量机方法来解决此问题。在样本内容和数量相同的情况下,将基于两种核函数的SVM分类器对绝缘子污秽程度的评定结果与其他方法的评定结果^[1]相

比较,前者的准确率更高。这说明支持向量机能较好地解决小样本、非线性等实际问题,具有很强的泛化能力。将其应用于绝缘子污秽在线监测系统中来评定绝缘子表面污秽程度,具有一定的工程实用价值。

[参考文献]

- [1] 焦尚彬,刘丁,郑岗,等.基于模糊逻辑方法的高压绝缘子污秽程度评定[J].电力系统自动化,2005,29(7):84-87.
- [2] 解建军,焦尚彬,张青.输电线路绝缘子污秽在线监测系统的设计与实现[J].西北农林科技大学学报:自然科学版,2005,33(10):109-114.
- [3] 张寒,文习山,丁辉.用神经网络预测基于气象因素的绝缘子等值附盐密度[J].高压电器,2003,39(6):31-32,35.
- [4] Vapnik V. The nature of statistical learning theory[M]. New York: Springer Verlag, 1999.
- [5] 许涛,贺仁睦,王鹏,等.基于输入空间压缩的短期负荷预测[J].电力系统自动化,2004,28(6):51-54.
- [6] 刘蔚,韩祯祥.基于支持向量机的配电网重构[J].电力系统自动化,2005,29(7):48-52.
- [7] 张恒喜,郭基联,朱家元,等.小样本多元数据分析方法及应用[M].西安:西北工业大学出版社,2002.
- [8] 刘志刚,李德仁,秦前清,等.支持向量机在多类分类问题中的推广[J].计算机工程与应用,2004,40(7):10-13.
- [9] 国家技术监督局 GB/T 5582-93 高压电力设备外绝缘污秽等级[S].北京:中国标准出版社,1994.
- [10] 国家技术监督局 GB/T 4585 2-91 交流系统用高压绝缘子人工污秽试验方法:固体层法[S].北京:中国标准出版社,1993.
- [11] Fierro-Chavez J L, Ramirez-Vazquez I, Montoya-Tena G. On-line leakage current monitoring of 400 kV insulator strings in polluted areas[J]. IEE Proc Generation, Transmission and Distribution, 1996, 143(6): 560-564.
- [12] Richards C N, Renowden J D. Development of a remote insulator contamination monitoring system[J]. IEEE Transactions on Power Delivery, 1997, 12(1): 389-397.

Application of support vector machine in assessing contamination condition of high voltage insulators

JIAO Shang-bin¹, LIU Ding¹, ZHENG Gang¹, ZHANG Qing^{1,2}, WANG Hong-jiang¹

(1 Xian University of Technology, Xian, Shaanxi 710048, China;

2 Yangling Institute of Vocation and Technology, Yangling, Shaanxi 712100, China)

Abstract: Assessing the insulator surface contamination condition can be described as multi-pattern recognition. The traditional methods for assessing this can not achieve the required accuracy for some engineering application due to the limited sample data sets and the complex factors that affect the surface contamination condition of insulator. In the research, based on the full advantage of SVM's ability to solve the problem with relatively few samples and nonlinear and high dimensions, the DAG (Directed Acyclic Graph) SVM model for multi-pattern recognition is used to assess the contamination condition of the high voltage insulator. The model is easily realized and can find out the super-plane between patterns without local minima, and has strong universal ability. The nonlinear relationship between the electrical characteristics of different contamination condition insulators, the environment factors and the contamination condition is considered synthetically by the SVM method, and the surface contamination condition of insulator is assessed. The results show that the model is suitable to the contamination assessment.

Key words: Support Vector Machine (SVM); high voltage insulators; insulator contamination condition assessment; pattern recognition