基于属性约简与参数优化的SVM故障诊断研究

李 莎,陶 红,高 尚
(江苏科技大学 计算机科学与工程学院,江苏 镇江 212003)

摘 要:应用数据挖掘的方法从实时数据库中提取相应的故障诊断知识是一种有效途径,也是很有现实意义和研究价值的问题。为提高汽轮机组故障诊断的效率,并考虑其计算成本和复杂性,把关联分析作为数据的前处理器,通过计算属性间的相关系数,结合最大最小聚类方法,删除冗余属性。然后采用支持向量机进行故障诊断,构造SVM多分类器,采用粒子群优化算法对参数寻优并训练样本。并与BP神经网络和线性判别分析做比较,实验表明此故障诊断方法诊断速度快、准确率高,可以很好地应用于设备故障诊断。

关键词:支持向量机;关联分析;最大最小距离;粒子群;故障诊断
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2012)04-0175-04

SVM Fault Diagnosis Research Based on Attribute Reduction and Parameters Optimization

LI Sha, TAO Hong, GAO Shang
(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: Applying data mining methods to extract the appropriate fault diagnosis knowledge from the real-time database is an effective way, also is an practical significance and research value problem. In order to raise the efficiency of fault diagnosis of steam turbine units and consider its costs and complexity, use the correlation analysis as data pre-processor. Calculate the correlation coefficients between attributes, and combine with max-min distance, then keep only one of the attributes which most highly correlates. Then construct support vector machine classifier, applying particle swarm optimization to find optimal parameter. Experimental results show that SVM outperforms linear discriminant analysis (LDA) and back-propagation neural networks (BPN) in classification performance and can be well applied in fault diagnosis.

Key words: support vector machine; correlation analysis; max-min distance; particle swarm optimization; fault diagnosis

0 引 言

对于电力企业而言,越来越多的数据被DAS和DCS系统存储到实时数据库中,日积月累的历史数据占据着庞大的存储空间,这些数据背后往往蕴涵着丰富的知识,仅靠经验很难理解这些数据之间的关系,应用数据挖掘的方法从系统的历史数据库中提取相应的故障诊断知识应该是一种有效途径,也是很有现实意义和研究价值的问题[1,2]。支持向量机(Support Vector Machine,SVM)是数据挖掘的新方法,是一种小样本多元数据分析方法,具有很强的容错能力和泛化能力,克服了传统机器学习中过学习、欠学习、局部极小值、维数灾难等问题。SVM用于故障诊断的最大优势在于它适用于小样本决策,能进行实时在线监控。随着应用的广泛,支持向量机的不足也逐渐被关注,在模型参数选择方面,目前还缺乏有效的方法和理论依据,对其分类精度有着很大影响。其次,在对大量数据进行模式分类和时间序列预测时,如何缩短样本的训练时间,仍是需考虑的问题。文中提出关联分析结合最大最小聚类方法对样本属性约简,然后采用粒子群优化算法对SVM参数寻优,并用于汽轮机组故障诊断,实验表明前述方法诊断速度快,准确率高。

1 属性约简

一些软计算方法,如神经网络、遗传算法、决策树、粗糙集[3]和关联分析被频繁用于删除不相关的、冗余属性。当应用于工业中,需要考虑其计算成本和复杂性,故采用关联分析方法[4,5]。相关分析法是在分析某

收稿日期:2011-08-31;**修回日期**:2011-12-02
基金项目:人工智能四川重点实验室开放课题(2009RY001)
作者简介:李 莎(1986-),女,江苏泰州人,硕士研究生,研究方向为数据挖掘、模式识别与智能系统;高 尚,副教授,硕士生导师,主要从事系统理论及智能计算方面的研究。

个问题或指标时，将与该问题或指标相关的其他问题或指标进行对比，分析其相互关系或相关程度的一种分析方法，用少数几对综合变量来反映 2 组变量间的线性相关性质。给定两个属性，这种分析可以根据可用的数据度量一个属性能在多大程度上蕴涵另一个。对于数值属性，通过计算属性 A 和 B 之间的相关系数（又称皮尔逊积矩系数 Pearson product coefficient），可以估算这两个属性的相关度 $r_{A,B}$。即

$$r_{A,B}=\frac{\sum_{i=1}^{N}(a_i-\bar{A})(b_i-\bar{B})}{N\sigma_A\sigma_B}=\frac{\sum_{i=1}^{N}(a_i b_i)-N\bar{A}\bar{B}}{N\sigma_A\sigma_B}-1\leqslant r_{A,B}\leqslant 1$$

相关系数值为 -1（完全负相关关系）~ +1（完全正相关关系）之间，相关系数为 0 时，表示不存在相关关系。当属性维数较高时，通过关联分析无法直接准确找出最简属性。最大最小距离算法也称小中取大距离算法，首先根据确定的距离阈值寻找聚类中心，然后根据最近邻规则把样本划分到各聚类中心对应的类别中。由于聚类中心之间的距离通常总大于各类样本的类内平均距离，即聚类中心之间的关联强度总小于各类样本的类内平均强度，因此相关分析结合最大最小聚类可以找到较优的约减属性。算法描述如下：

步骤 1：任选一个模式样本作为第一聚类中心 Z_1；

步骤 2：选择与 Z_1 相关强度最小的样本作为第二聚类中心 Z_2；

步骤 3：逐个计算每个模式样本与已确定的所有聚类中心之间的相关系数，并选出其中的最大值相关系数；

步骤 4：在所有的最大值相关系数中选出一个最小值相关系数，如果该值小于设定的阈值 T，则将产生最小相关系数的那个模式样本定义为新增聚类中心，并返回上一步骤。否则，聚类中心的计算步骤结束；

步骤 5：重复步骤 3 和步骤 4，直到没有新的聚类中心出现为止。得到 K 组聚类中心（即属性集），根据最大聚类比选出最优属性集。

2 参数优化

支持向量机的核心思想是建立一个超平面作为决策曲面，使正样本与负样本之间的分离边界最大化。同时，通过引入内积核使支持向量机成为一种通用逼近器。对于采用径向基核的支持向量机的主要参数是惩罚系数 C 和核函数宽度 σ，求解最佳参数有多种方法。完成一个完全的网格搜索非常费时。混沌优化算法对于搜索空间小时效果显著，但当搜索空间大时却不能令人满意。遗传算法[6~8]的编程实现比较复杂，首先需要对问题进行编码，找到最优解之后还需要对问题进行解码，另外三个算子的实现也有许多参数，如交叉率和变异率，并且这些参数的选择严重影响解的品质，而目前这些参数的选择大部分是依靠经验。其他一些方法如深度优化搜索[9]、交叉验证等都无法达到预期效果。而粒子群算法（PSO）是一种有效的全局寻优算法，采用速度——位移模型，没有遗传算法的“交叉”（Crossover）和“变异”（Mutation）操作，它通过追随当前搜索到的最优值来寻找全局最优。这种算法以其实现容易、精度高、收敛快等优点引起了学术界的重视，并且在解决实际问题中展示了其优越性。

文中预采用 PSO 算法对 SVM 两个参数进行寻优[10]。PSO—SVM 具体流程如下：

步骤 1：PSO 参数设置以及初始化粒子群；

步骤 2：计算每个粒子的适应度值；

步骤 3：更新粒子位置和速度；

步骤 4：看是否满足最大迭代次数，如果没有，返回步骤 2，重新计算每个粒子的适应度值；

步骤 5：获得优化的 SVM 分类器参数；

步骤 6：获得 PSO-SVM 分类器。

3 实例分析

为了准确诊断汽轮机组故障，必须深入挖掘故障信息。具体步骤如下：

1）从实时监测系统采集相关数据进行预处理，去除噪声和缺失数据；

2）对数据进行归一化处理，形成初始连续属性决策表；

3）对样本数据特征选择，计算属性间的相关系数，结合最大最小聚类编程求得最佳属性集，根据计算结果删除冗余属性；

4）构建支持向量机，PSO 寻找最佳参数，选取部分作为训练数据，剩余作为测试数据，训练 SVM 分类器；

5）绩效评估。

文中通过实时监测系统，共采集 30 组数据，以汽轮机组振动信号的频谱特征中 0 ~ 0.4f、0.4f ~ 0.6f、0.6f ~ 1f、1f、2f、3f、4f、> 4f 共 8 个不同频段 $\{x_1,x_2,\cdots,x_8\}$ 上的幅值分量能量作为故障征兆属性，对汽轮机发电机组的 4 种常见故障：油膜涡动、不平衡、不对中和动静碰摩进行诊断[11]。对数据进行归一化预处理，采用的归一化映射为 $f:x\rightarrow y=\frac{x-x_{\min}}{x_{\max}-x_{\min}}$，式中，$x,y\in R^n$，$x_{\min}=\min(x)$，$x_{\max}=\max(x)$，归一化的效果是原始数据被规整到 [0,1] 区间归一化，得到如表 1 所示的初始属性决策表。

表1　初始连续属性决策表

样本	0~0.4f	0.4f~0.6f	0.6f~1f	1f	2f	3f	4f	>4f	故障类型
1	0.0667	0.0103	0.0286	0.8498	0.2045	0.3333	0.2917	0.2778	1
2	0.4667	0.0412	0.1714	0.364	0.3636	0.0513	0.25	0.3333	1
3	0	0	0	0.9735	1	0.6667	0.2917	0.6111	1
4	0.0333	0	0	0.7933	0.1818	0.2051	0.125	0.1667	1
5	0.0333	0.0103	0.0286	0.189	0.1591	0.0256	0.0417	0.1667	1
6	0.1667	0.0103	0.0857	0.2014	0.1477	0.0513	0	0.1667	1
7	0	0	0	0	0.5341	0.0769	0.0833	0.3333	2
8	0	0	0	0.0353	0.5795	0.0513	0.0417	0.2778	2
9	0	0	0	0.3322	0.8977	0.0769	0.1667	0.2222	2
10	0.0333	0.0103	0.0286	0.0283	0.6705	0.1026	0.0833	0.1111	2
11	0	0	0	0.2049	0.6705	0.0769	0.0417	0.3333	2
12	0.0333	0.0103	0.0286	0.03	0.6705	0.1026	0.125	0.1667	2
13	0.0333	0.4536	0.0286	0.0371	0.0114	0.0513	0.0833	0.0556	3
14	0.1333	1	0.2	0.0671	0.0114	0.0513	0.0833	0.2222	3
15	0.0667	0.6186	0.0571	0.0601	0.0568	0.0256	0.0417	0.1667	3
16	0.3333	0.732	0.1714	0.0636	0.0341	0.1282	0.375	0.0556	3
17	0.5333	0.4021	0.1714	0.0406	0.0909	0	0	0	3
18	0.9667	0.4227	0.9429	0.1943	0.8295	0.8718	0.9167	0.8889	4
19	0.8667	0.2784	1	0.2032	0.8295	0.7179	0.5	0.7778	4
20	1	0.4124	0.8571	0.2067	0.8409	0.8718	0.8333	0.9444	4
21	0.6667	0.4124	0.8286	0.2615	0.7727	0.2564	0.875	0.8333	4
22	0.8667	0.2784	1	0.2049	0.8409	0.7436	0.5	0.8333	4
23	0.6667	0.3093	0.6571	0.2933	0.6818	0.8462	1	0.8889	4
24	0.0333	0	0	0.8074	0.1591	0.1795	0.1667	0.1111	1
25	0	0	0	1	0.8409	1	0.5833	0.7222	1
26	0	0	0.0286	0.1555	0.8977	0.0769	0.1667	0.2778	2
27	0	0	0	0.0424	0.5341	0.0513	0.0417	0.2778	2
28	0.1667	0.299	0.2	0.0389	0	0.0256	0.0417	0.1111	3
29	0	0.4227	0.0857	0.0442	0	0.0513	0.2083	0.0556	3
30	0.7	0.299	0.6857	0.2951	0.6591	0.8205	0.8333	1	4

注:1~4分别表示油膜涡动、不平衡、不对中和动静碰摩

1)对数据进行相关分析,通常情况下通过以下取值范围判断变量的相关强度:0.8~1.0极强相关;0.6~0.8强相关;0.4~0.6中等程度相关;0.2~0.4弱相关;0.0~0.2极弱相关或无相关。通过分析,设定阈值 $T=0.45$,计算得到五组属性集:$\{x_1,x_2,x_4,x_5\}$,$\{x_2,x_3,x_4,x_5\}$,$\{x_2,x_4,x_6\}$,$\{x_2,x_4,x_7\}$,$\{x_2,x_4,x_8\}$,根据最大聚类比选取$\{x_1,x_2,x_4,x_5\}$为最优属性集。

2)构建支持向量机。SVM是2类分类器,对于多个故障类型进行诊断时,必须构造多类分类器[12]。采用一对一多类分类法,由6个2类SVM分类器组成多分类模型。SVM类型设置为C-SVC,采用RBF核函数 $k(x_i,x)=\exp(-\|x_i-x\|^2/\sigma^2)$ 来训练分类器,将30组数据前20组用来训练,后10组用来测试。粒子群算法对RBF核函数宽度 σ 和惩罚系数c参数寻优,设PSO速度调节参数 $C_1=1.5$,$C_2=1.7$,借助Matlab仿真平台测得当 bestc = 16.51,bestσ = 2 时可得到最佳分类精度,终止代数为200,种群数量POP=20,如图1所示,并与BP神经网络和线性判别分析(LDA)相比结果如表2。

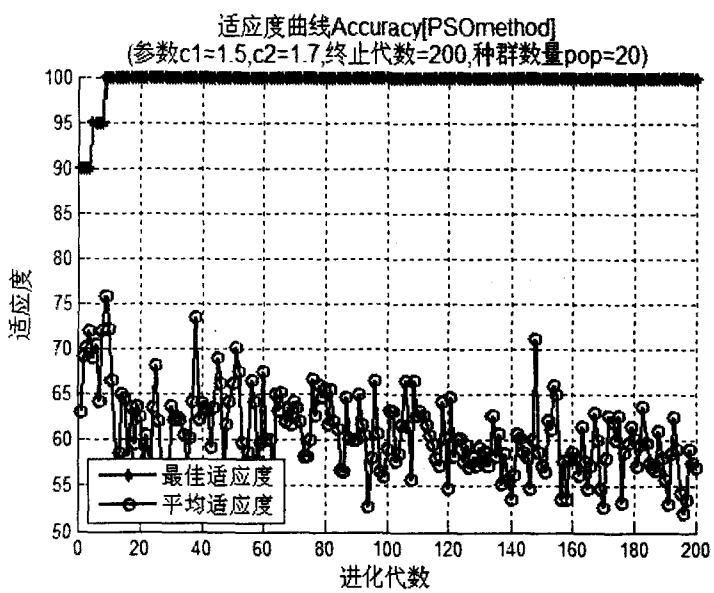

图1　Matlab仿真结果

表2 实验结果

Method performance	BPN		LDA		PSO-SVM		
	Training (%)	Testing (%)	Training (%)	Testing (%)	Training (%)	Testing (%)	T(s)
约简前	85	100	100	60	100	100	0.21
约简后	70	80	85	70	100	100	0.10

3)在相同的参数下,使用没有经过关联分析处理的数据对SVM多类分类器进行训练和诊断时间为0.21s,是经过处理后的两倍多,当数据样本和故障类型增加时,效果会更显著。经过PSO参数寻优,准确性达到100%,明显优于BP神经网络和线性判别分析,可以很好地应用于设备故障诊断。

4 结束语

在汽轮机组故障诊断中,采集到的数据中存在一定的冗余属性和重复样本,利用关联分析,计算属性间相关系数,可以对其进行约简处理以达到降低数据维数的目的,使得样本更具有代表性,实现简单,效率更高。当数据样本和故障类型增加时,效果会更显著。此外,应用粒子群方法选择参数能避免支持向量机的过学习,一定意义上寻得最优参数,提高了分类准确率。文中将进一步研究,并把关联分析直接应用于电厂实时数据库,以达到预期效果。

参考文献:

[1] 梁 娜.基于数据挖掘的火电厂故障诊断研究[D].保定:华北电力大学,2007.

[2] Chen Kaiying, Chen Longsheng. Using SVM based method for equipment fault detection in a thermal power plant[J]. Computers in Industry, 2011, 62(1): 42-50.

[3] 郭小芸,马小平.基于粗糙集的故障诊断特征提取[J].计算机工程与应用,2007,43(1):221-224.

[4] Han J, Kamber M. 数据挖掘概念与技术[M].范 明,孟小峰,译.北京:机械工业出版社,2001.

[5] 任江涛,黄焕宇,孙婧昊,等.基于相关性分析及遗传算法的高维数据特征选择[J].计算机应用,2006,26(6):1403-1405.

[6] 连 可,陈世杰,周建明,等.基于遗传算法的SVM多分类决策树优化算法研究[J].控制与决策,2009,24(1):7-12.

[7] Wu Chih-Hung, Ken Yun, Huang Tao. Patent classification system using a new hybrid genetic algorithm support vector machine[J]. Applied Soft Computing, 2010, 10(4): 1164-1177.

[8] 徐庆伶,江西莉.一种基于支持向量机的半监督分类方法[J].计算机技术与发展,2010,20(10):115-117.

[9] 向昌盛,周子英.支持向量分类机的参数选择方法研究[J].计算机技术与发展,2010,20(9):95-97.

[10] 姚全珠,蔡 婕.基于PSO的LS-SVM特征选择与参数优化算法[J].计算机工程与应用,2010,46(1):134-136.

[11] 汪 江.汽轮机组振动故障诊断SVM方法与远程监测技术研究[D].南京:东南大学,2005.

[12] 江 伟,罗 毅,涂光瑜.基于多类支持向量机的变压器故障诊断模型[J].水电能源科学,2007,25(1):52-55.

(上接第174页)

有无标度特性,即具有"马太效应"现象,新上市的股票更容易与那些度很大的节点相联。同时,在该股票关联网络中存在着对整个网络具有重要作用的节点,这些节点所对应的股票价格的大幅度波动会对整个股市网络产生较大的影响,这些节点的移除也会影响到股市网络的稳定性。因此在证券交易市场,应该加强对这些股票的监管,确保证券交易市场的稳定性。

参考文献:

[1] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(3): 167-256.

[2] 孙博文,孙百瑜,刘天立,等.中国股市的拓扑结构及其复杂性质研究[J].黑龙江大学自然科学学报,2006,23(3):366-369.

[3] 熊胜君,杨朝军.沪深股票市场行业效应与投资风格效应的实证研究[J].系统工程理论与实践,2006(4):44-49.

[4] Kim H J, Lee Y, Kahng B, et al. Weighted scale-free network in financial correlations[J]. J Phys Soc Jpn, 2002, 71(9): 2133-2136.

[5] 汪小帆,李 翔,陈关荣.复杂网络理论及应用[M].北京:清华大学出版社,2006.

[6] 李耀华,姚洪兴.金融危机下股市网络的结构特性研究[J].成都信息工程学院学报,2010,25(1):107-111.

[7] Watts D J, Strogatz S H. Collective dynamics of small world networks[J]. Nature, 1998, 393(6): 440-442.

[8] 黄玮强,庄新田.中国股票关联网络拓扑性质与聚类结构分析[J].管理科学,2008,21(3):94-102.

[9] Kim H J, Kim I M. Scale-free network in stock market[J]. J Kor Phys Soc, 2002, 40(6): 1105-1108.

[10] 兰旺森,张所地.基于复杂网络的中国股市房地产版块股票强相关性研究[J].数学的实践与认识,2009,39(4):62-66.

[11] 鲁巍巍,林正春.基于复杂网络理论的沪深A股分析[J].科学技术与工程,2009,9(11):2859-2862.

[12] Albert R, Barabasi A L. Statistical mechanics of complex network[J]. Rev Mod Phys, 2002, 74(1): 47-97.