基于支持向量机的暂态稳定分类中的特征选择

向丽萍[1]，王晓红[1]，王建[2]，项杨[1]，谢桦[3]，王晓茹[1]

(1.西南交通大学电气工程学院，四川 成都 610031；2.清华大学电机工程与应用电子技术系，北京 100084；
3.北京交通大学电气工程学院，北京 100044)

摘要：特征选择是支持向量机(SVM)分类实现中非常重要的环节。针对传统方法进行特征选择的缺陷，提出了基于遗传算法的特征选择方法。综述和提出了支持向量机暂态稳定分类的初始特征；建立了IEEE16机86节点系统的暂态稳定分类初始特征样本集；利用主成分分析和遗传算法对维数较大的初始特征进行了有效降维；并通过因子负荷，完成了暂态稳定输入特征的选择；经过支持向量机分类器测试，显示选出的特征有很好的分类效果。

关键词：特征选择；主成分分析；遗传算法；支持向量机

Feature selection for SVM based transient stability classification

XIANG Li-ping[1], WANG Xiao-hong[1], WANG Jian[2], XIANG Yang[1], XIE Hua[3], WANG Xiao-ru[1]

(1.Southwest Jiaotong University, Chengdu 610031,China; 2. Tsinghua University, Beijing 100084, China;
3.Beijing Jiaotong University, Beijing 100044, China)

Abstract: Feature selection plays a very important role in realizing support vector machine(SVM) classifier. Aimed at the disadvantages existing in feature selection by traditional method, a new method based on genetic algorithm to select the input features. In this paper, a set of features which fit for transient stability assessment is summarized.The primary feature pattern of IEEE16-machine 86-bus system system is established. Using principal component analysis(PCA) and genetic algorithm (GA) to efficiently reduce the dimension of the primary feature. By using the idea of factor loading, it reconstructs the input space to accomplish feature selection. SVM classifier test demonstrates the validity of the proposed approach.

Key words: feature selection; principal component analysis; genetic algorithm; SVM

中图分类号：TM711 文献标识码：A 文章编号：1003-4897(2007)09-0017-05

0 引言

用于数据挖掘的海量数据可能包含成千上万的特征，其中大部分特征与挖掘任务是不相关的或是冗余的，这些特征增加了数据量，减慢了挖掘进程，并有可能使发现的知识质量很差。特征选择就是一个从原有的特征集合中选择一个(相对某种评价准则)最优特征子集的过程。

电力系统暂态稳定评估的特征选择，需解决以下几个关键问题：数据源、原始输入特征的选取、特征选择算法的选取、产生的特征子集性能优劣的评价标准。目前的研究多采用IEEE 9机39节点系统，文献[1]中提出了基于遗传算法的特征选择，取得了较好的结果，但对于主成分分析变换后的特征含义不清。文献[2]中提出了34个原始特征，并在此基础上采用了Tabu搜索技术进行特征选择，所选特征具有代表性，但采用等频离散化的方法对输入特征进行离散化，减弱了特征信息，可能对分类影响较大。

本文在综合现有研究的基础上，提出了暂态稳定评估的初始特征，建立了IEEE16机系统的初始样本集；利用主成分分析法提取出一组有较好分类效果的综合特征；然后利用遗传算法进行特征选择，从提取的综合特征中选择出使类内、类间距离判据最大的一个综合特征，并通过因子负荷，完成了暂态稳定输入特征的选择。

1 原始特征的选择

电力系统暂态稳定性与系统的运行工况、故障地点、故障类型、故障切除时间、故障后的网络结

基金项目：国家973重点基础研究发展规划资助项目(G1998010301)

构以及发电机参数（包括励磁系统和调速系统的参数）有关。就电力系统暂态受扰模式的输入与输出间的匹配关系而言，对于某一电力系统，当系统的发电水平及发电分布、负荷水平及故障等条件都确定后，系统的暂态稳定性就确定了。文献[3]采用7机24节点系统，选用了稳态及故障时刻的状态信息为输入，共24个原始特征；文献[4]提出在故障发生时刻、中间时刻和故障切除时刻这3个时刻采集系统中各发电机轴间的最大相对摇摆角及它们的变化率共5个变量，反映了系统的故障信息和功角变化趋势，包含了丰富的系统稳定性信息；文献[2]在文献[3,5,6]的基础上采用34个系统特征作为ANN的初始输入特征集，并用于特征选择，取得较好结果。在分析和综合上述研究的基础上，提出采用74个系统初始特征来构成初始的候选特征集。如表1所示。

表1 初始输入特征

Tab.1 Primary imput feature

1：系统中发电机机械输入功率的平均值
2：故障切除时刻所有发电机转子动能的平均值
3：系统总的能量调整
4：故障切除时对系统的冲击
5：故障切除时coi坐标下转速的和
6：故障切除时与coi相差最大的转速
7：故障切除时与coi相差最大的功角
8：故障切除时领前机与殿后机的功角之差
9：故障切除时最大的加速度之差
10：故障切除时最大的加速度的变化率之差
11：故障切除时最大的转子动能之差
12：故障切除时最大的转子动能变化率之差
13：$TZ_{13}=TZ_{12}/TZ_{10}$
14：$TZ_{14}=TZ_{12}/TZ_{11}$
15：$TZ_{15}=TZ_{11}/TZ_{8}$
16：$TZ_{16}=TZ_{12}/TZ_{8}$
17：$TZ_{17}=TZ_{11}/TZ_{10}$
18：故障瞬间所有发电机转子初始加速度的最大值
19：具有最大初始加速度发电机的初始角度
20：所有转子初始加速度的最小值
21：故障切除时刻所有转子动能的最大值
22：故障切除时具有最大动能发电机的转子角度
23：所有发电机初始加速度的方差
24：故障切除时具有最大转子角度发电机的转子动能
25：所有发电机初始加速功率的均值
26：所有发电机初始加速功率的方差
27：所有发电机相对初始加速功率的均值
28：所有发电机相对初始加速功率的方差
29：所有发电机初始加速度的均值
30：故障瞬间发电机所受的最大有功冲击
31：故障瞬间发电机所受的最小有功冲击
32：所有发电机相对初始加速度的平均值
33：所有发电机相对初始加速度的方差
34～38：取t_0、0.5t_c和t_c（t_0为故障发生时刻，t_c为故障切除时刻）3个时刻系统中各发电机大轴间的最大相对摇摆角δ_i（i=0，1，2）及它们的变化率V_i（i=1，2），共5个变量
39～70：故障切除时各发电机有功功率和无功功率
71：故障瞬间发电机发出的有功功率之和
72：故障切除时发电机发出的有功功率之和
73：故障切除时所有发电机转子动能的平均值
74：所有发电机转子初始加速度的均方根误差

2 利用主成分分析方法实现特征降维

主成分分析法PCA (Principal Component Analysis)是由霍特林提出的，它将原来的信息重新组合成一组相互独立的少数几个综合指标来代替原来指标，并能反映原指标的主要信息。其原理为[7]：

在原始变量的p维空间中，找到新的p个坐标轴，新变量与原始变量的关系可以表示为：

$$\begin{cases} y_1=\mu_{11}x_1+\mu_{12}x_2+\mu_{13}x_3+\cdots+\mu_{1p}x_p \\ y_2=\mu_{21}x_1+\mu_{22}x_2+\mu_{23}x_3+\cdots+\mu_{2p}x_p \\ \cdots\cdots\cdots\cdots \\ y_p=\mu_{p1}x_1+\mu_{p2}x_2+\mu_{p3}x_3+\cdots+\mu_{pp}x_p \end{cases}$$

我们称这p个新变量为原始变量的主成分，每个主成分均为原始变量的线性组合。

本例中就是将74个输入特征运用主成分分析，在力保数据信息丢失最小原则下，对高维变量空间进行降维处理，其数学模型的求解步骤如下[8]：

（1）对样本数据进行标准化处理。

（2）计算变量的相关系数矩阵$\boldsymbol{R}$。

（3）求出$\boldsymbol{R}$的特征根，并按特征值大小排序，得到$\lambda_1\geq\lambda_2\geq\lambda_3\geq\cdots\cdots\geq\lambda_p\geq0$；即对应的单位特征向量$\mu_1,\mu_2,\mu_3,\cdots,\mu_p$。

（4）计算主成分 $Y_i=\mu_i'\ X^*$，Y_i为第i主成分。

（5）计算累积方差贡献率，确定主成分的个数，根据给定阈值$L(0.85\leq L\leq0.95)$，取累积贡献率$\sum_{i=1}^{k}\lambda_i/\sum_{i=1}^{p}\lambda_i>L$对应的前$k$（$k\leq p$）个主成分。

（6）计算因子(成分)负荷量[9,10]

利用主成分分析中p个特征值和对应的特征向量计算因子负荷矩阵：

$\rho_{Ki}=\rho_{Y_KX^*}=\mu_{iK}\sqrt{\lambda_K}$，其中$\rho_{Ki}$即为第$K$个主成分，与第$i$个变量之间的相关系数，得因子负荷矩阵为

$$\boldsymbol{A}=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pk} \end{pmatrix}=\begin{pmatrix} \mu_{11}\sqrt{\lambda_1} & \mu_{21}\sqrt{\lambda_2} & \cdots & \mu_{k1}\sqrt{\lambda_k} \\ \mu_{12}\sqrt{\lambda_1} & \mu_{21}\sqrt{\lambda_2} & \cdots & \mu_{k2}\sqrt{\lambda_k} \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{1p}\sqrt{\lambda_1} & \mu_{2p}\sqrt{\lambda_2} & \cdots & \mu_{kp}\sqrt{\lambda_k} \end{pmatrix}$$

将上述因子负荷矩阵改变坐标轴，进行旋转，能够重新分配各因子(成分)解释原始变量方差的比例，使因子更易于理解。

3　利用遗传算法进行特征选择

遗传算法是一种全局寻优算法，像撒网一样，在变量空间中进行寻优，由 N 个数字串组成的群体在遗传因子的作用下，同时对空间中不同的区域进行充分搜索，从而构成一个不断优化的群体序列。遗传算法是通过保持在解空间不同区域中各个点的搜索，而不是盲目地穷举或瞎碰，故相对其他优化方法而言，遗传算法能以很大的概率找到优化问题的全局最优解。运用遗传算法进行特征选择步骤如下：

1)初始化群体。染色体编码，生成初始种群。特征选择问题是从最初的 D 个特征变量中选择出其中的 d 个特征。在用遗传算法进行特征选择时，采用二进制染色体编码，用一个 D 位的由0或1构成的字符串表示一种特征组合，数字1表示对应的特征被选中，数字0表示对应的特征未被选中。特征变量能否被选中是随机的，即初始种群中的每一个染色体都是随机产生的。

2）定义适应度函数。为了得到一组对分类最有效的特征，本文采用基于类内、类间距离的可分判据 J 作为适应度函数，定义如下：

$$J = tr(S_b) \Big/ tr(S_w) \tag{1}$$

其中：

$$S_b = \sum_{i=1}^{c} P_i (m_i - m)(m_i - m)^T \tag{2}$$

$$S_w = \sum_{i=1}^{c} P_i \frac{1}{n} \sum_{k=1}^{n_i} (x_k^{(i)} - m_i)(x_k^{(i)} - m_i)^T \tag{3}$$

$$m_i = \frac{1}{n} \sum_{k=1}^{n_i} x_k^{(i)} \tag{4}$$

$$m = \sum_{i=1}^{c} P_i m_i \tag{5}$$

式中：m_i 表示第 i 类样本集的均值向量；m 表示所有各类的样本集总均值向量；称 S_b 为类间离散度矩阵；S_w 为类内离散度矩阵。直观上，希望被提取特征的类间离散度的值 S_b 尽量大，类内离散度的值 S_w 尽量小，则 J 越大，这样有利于分类。

3）计算群体中每个个体的适应度函数值。将个体解码，选中个体中基因位为“1”的特征并对应到原始数据中，组成一个新的空间，用类内、类间距离判据计算其适应度值。

4)进行选择、交叉和变异操作，产生新一代群体。

5)重复3)、4)，直至进化代数超过给定的最大进化代数为止。特征选择的结果就为最后一次迭代后群体中的最优解。

4　暂态稳定分类的特征选择

4.1　数据来源及初始特征样本集组织

初始特征样本集采用 IEEE16 机仿真系统，共有16台发电机，68条母线，86条线路。仿真时间为3 s。所有的故障都设定为三相短路，近端节点断开，然后远端节点断开，形成永久故障。故障时间分别设定为0.15～0.2 s，0.25～0.3 s，0.3～0.35 s。并分别设置80%，100%，110%，120%四个负荷水平。获得了390个样本（表2）。选择样本集4和样本集5共85个样本作为测试样本，其他305个作为训练样本。所有的样本根据第一摇摆失稳被分为两类：稳定和不稳定。

表2 样本构成表

Tab.2 Structure of feature pattern

样本集	负荷水平/(%)	故障时间/s	样本个数
1	80	0.15～0.2	77
2	80	0.25～0.3	72
3	100	0.15～0.2	72
4	100	0.25～0.3	56
5	110	0.3～0.35	29
6	120	0.15～0.2	55
7	120	0.25～0.3	29

4.2　特征选择

本文进行特征选择的流程图如图1所示.

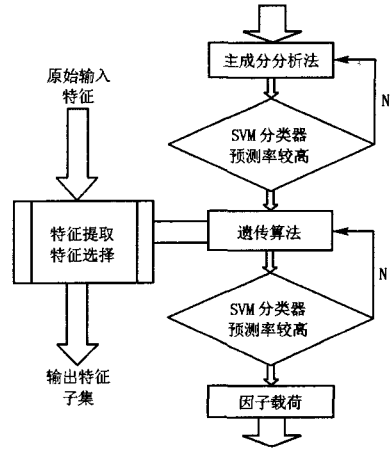

图1 特征选择的流程

Fig.1 Flow chart of feature selection

其中 SVM[11]采用 C-SVM 二分类器，用于给出暂态稳定评估的结果，SVM 分类器的分类效果也作为特征选取的评判标准。

4.2.1 主成分分析

在仿真中，假设发电机机械功率和故障发生时刻系统中各发电机轴间的最大相对摇摆角保持不变，消去 74 个原始特征中的特征 1 和特征 34，对剩下的 72 个特征进行主成分分析。

运用 SPSS 统计分析软件[8]进行主成分分析。设定累计贡献率为 0.9。由实验可得前 22 个主成分的贡献率达到 91%，其它的主成分贡献率都很低，可以视为冗余信息，舍去。通过主成分分析，72 个特征量压缩到 22 个，用 SVM 进行测试，结果显示压缩后的测试样本预测率达到 97.6%，与全部 72 特征的预测率 98.8%基本相同，但特征数却不到原来的 1/3，达到分类性能的要求，可以进行下一步基于遗传算法的特征选择。经过 PCA 选出的 22 个特征（主成分），每一个都是原始输入特征的线性组合。

4.2.2 遗传算法进一步降维

用遗传算法对主成分分析选取的22个综合特征进一步压缩。遗传算法中控制参数选取的不同对算法的性能产生较大的影响，尤其是对算法收敛性的影响。这些参数包括群体规模N、进化代数M、编码长度、交叉概率P_c、变异概率P_m等。本文通过大量的仿真实验，最后选定种群规模N=20，个体长度（编码长度）为22，进化代数M=150，交叉概率P_c=0.7，变异概率P_m=0.1。

图 2 为设定值经过 150 次迭代后最优解的变化和种群均值的变化。随着进化的进行，群体中的个体趋于相同，最终形成最优解，证明所选择的进化代数是足够的，交叉概率和选择概率是恰当的。通过 150 次迭代后，得到个体：

Individual=0010000000000000000000

即 22 个综合特征中的第 3 个综合特征被选中。该个体在群体中具有最高的适应度函数值，是算法的全局最优解。

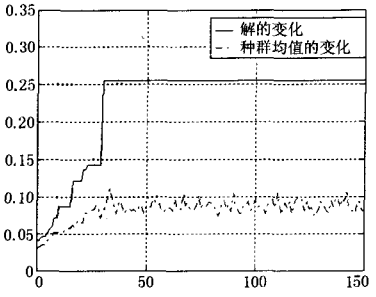

图2 经过150次迭代后最优解的变化和种群均值的变化

Fig.2 Evolution of optimum solution and population equalizing by iteration

4.2.3 因子负荷量

通过遗传算法，把22个综合特征压缩到1个，即选择出一个主成分。由主成分分析法可知，每个综合特征（主成分）是原始输入特征的线性组合。

在本文中，遗传算法选出一个主成分，列出其因子负荷表，即变量与主成分的相关系数，经过旋转后分析第三主成分与变量的关系，得出起主要作用的变量。旋转后部分因子负荷矩阵见表3。

表3 旋转后部分因子负荷表

Tab.3 Part of factor loading by revolution

特征	负荷量	特征	负荷量	特征	负荷量	特征	负荷量
3	-0.386	13	0.263	30	0.402	54	-0.407
4	0.852	14	0.931	31	-0.691	64	0.229
5	-0.306	16	0.949	51	0.278	65	0.292
9	0.313	21	0.793	52	-0.498	69	0.787
11	0.415	24	0.29	53	0.588	72	-0.261
12	0.932	26	0.285				

选出2组特征子集：

1为TZ_4, TZ_{12}, TZ_{14}, TZ_{16}, TZ_{21}, TZ_{31};

2为TZ_4, TZ_{12}, TZ_{14}, TZ_{16}, TZ_{21}, TZ_{31}, TZ_{53}, TZ_{69} 。

4.3 仿真测试

用SVM分类器分别测试这两个特征子集和全部特征的分类能力，比较它们的分类效果，如表4所示。

表4 不同组合特征子集的分类预测率

Tab.4 Forecast quotiety of different feature subset

特征组合	训练样本预测率/(%)	测试样本预测率/(%)
1	99	95.3
2	99.3	96.5
全部特征	100	98.8

表4可以看出，通过主成分分析和遗传算法选出的特征子集1和特征子集2有较好的预测率。与全部特征相比，特征空间不到原来的1/10，预测率几乎没有发生变化。与文献[1]相比，将数据集压缩了92%。

5 结论

本文根据IEEE16机仿真数据，建立了74个特征的高维暂态稳定评估的初始特征样本集；通过主成分分析提取出了有较好分类效果的一组综合特征；然后以类内、类间距离为适应度函数，利用遗传算法从主成分分析法提取的综合特征中选择了一个综合特征；最后利用因子负荷，确定6个输入特征；在特征选择过程中采用分类泛化性能比较好的SVM分类器进行测试，以此作为反馈，对算法和参数做出修正；最终所选特征用SVM分类器测试，新的特征达到与原始特征接近的分类效果，满足分类性能的要

求。因此，本文成功地提取了用于暂态稳定分类的有效特征。

参考文献

[1] 于之虹，郭志忠．遗传算法在暂态稳定评估输入特征选取中的应用[J]．继电器，2004，32(1)：16-20.
YU Zhi-hong, GUO Zhi-zhong.Feature Selection Based on Genetic Algorithm for Transient Stability Assessment[J]. Relay,2004，32(1):16-20.

[2] 顾雪平，张文朝．基于Tabu搜索技术的暂态稳定分类神经网络的输入特征选择[J]．中国电机工程学报，2002.
GU Xue-ping, ZHANG Wen-chao. Feature Selection by Tabu Search for Neural-network Based Transient Stability Classification[J]. Proceeding of the CSEE, 2002.

[3] 管霖，曹绍杰．基于人工智能的大系统分层在线暂态稳定评估[J]．电力系统自动化，2000，24（1）：22-26.
GUAN Lin, Tso S K. Combination of Heuristic Reasoning and ANN to Realize On-line Transienr Stability Assessment in Large Scale Power Systems[J]. Automation of Electric Power Systems,2000,24(1):22-26.

[4] 刘玉田，林飞．基于相量测量技术和模糊径向基网络的暂态稳定性预测[J]．中国电机工程学报，2000，24(2)：19-23.
LIU Yu-tian, LIN Fei. Application of PMU and Fuzzy Radial Basis Function Network to Power System Transient Stability Prediction[J].Proceeding of the CSEE, 2000,24(2):19-23.

[5] 顾雪平，曹绍杰．人工神经网络和短时仿真结合的暂态安全评估事故筛选方法[J]．电力系统自动化，1999，23(4)：16-20.
GU Xue-ping,Tso S K,Integration of ANNS and Short-duration Numerical Simulation for Contingency Screening of Transient Security Assessment[J]. Automation of Electric Power Systems,1999,23(4):16-26.

[6] 张琦，韩祯祥，等．用于暂态稳定评估的人工神经网络输入空间压缩方法[J]．电力系统自动化，2001，25(5)：32-36.
ZHANG Qi, HAN Zhen-xiang,et al. Input Dimension Reduction in Neural Network Training for Transient Stability Assessment[J]. Automation of Electric Power Systems, 2001,25(2):32-35.

[7] 卢纹岱，等．SPSS for Windows统计分析[M]．北京：电子工业出版社，2002. 311-316.
LU Wen-dai,et al. SPSS for Windows Statistical Analyse[M]. Beijing:Publishing House of Electronics Industry,2002.311- 316.

[8] 刘海燕，等．概率论与数理统计(下)[M]．北京：国防工业出版社，2001. 53-161.
LIU Hai-yan,et al. The Theory of Probability and Statistics[M].Beijing:National Defence Industry Press, 2001.153-161.

[9] 陈正昌,等.多变量分析方法统计软件应用[M]．北京：中国税务出版社,2005.
CHEN Zheng-chang, et al. Multivariable Analysis Method, Statistical Software Application[M].Beijing: Chinese Tax Affairs Publishing House,2005.

[10] 薛薇. SPSS统计分析方法及应用[M]．北京：电子工业出版社，2004.
XUE Wei.SPSS Statistical Analyse Method and Application[M]. Beijing:Publishing House of Electronics Industry,2004.

[11] 王晓红，王晓茹，李群湛．二分类数据的分类结果可视化算法[J]．西南交通大学学报，2006，41(3)：329-334.
WANG Xiao-hong, WANG Xiao-ru, LI Qun-zhan. Algorithm for Visualization of Classification Results of Two-category Data[J]. Journal of Southwest Jiaotong University, 2006,41(3):329-334.

收稿日期：2006-11-27；　修回日期：2007-01-15

作者简介：

向丽萍(1981-)，女，硕士研究生，研究方向为数据挖掘与电力系统紧急控制；E-mail:lpxiang@mars.swjtu.edu.cn

王晓红(1977-)，女，讲师，博士研究生，研究方向为数据挖掘与电力系统紧急控制；

王　建(1982-)，男，博士研究生生，研究方向为分布式电力系统计算。

（上接第 4 页　continued from page 4）

[4] 许正亚．变压器及中低压网络数字式保护[M].北京：中国水利水电出版社，2004.
XU Zheng-ya. Digital Protection for Power Transformer and Medial-low Voltage Electric Power Net[M].Beijing: China Water Power Press，2004.

收稿日期：2006-09-14　　修回日期：2007-03-07

作者简介：

韩　笑（1969-），男，硕士，副教授，主要从事电力系统继电保护，电力系统优化，电力系统仿真的研究与教学工作；E-mail:hxslqc@sina.com

戈祥麟(1964-)，男，大专，技师，从事电力系统继电保护施工、维护、调试工作；

汪经华（1967-），男，学士，工程师，主要从事电力系统继电保护工程施工及管理工作。