遗传算法在暂态稳定评估输入特征选择中的应用

于之虹, 郭志忠

(哈尔滨工业大学电气工程及自动化学院, 黑龙江 哈尔滨 150001)

摘要: 针对主成分分析中利用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。选择反映电力系统运行状态的特征变量,建立暂态稳定评估模型;为了提高数据处理的效率,首先对原始数据进行了动态聚类分析;对数据进行主成分分析后,以类内类间距离判据作为适应度函数,采用二进制编码形式的遗传算法进行特征选择。通过对3机9节点和10机39节点新英格兰系统的计算,验证了所选方法的有效性。

关键词: 特征选择; 遗传算法; 暂态稳定评估; 电力系统

中图分类号: TM71 **文献标识码**: A **文章编号**: 1003-4897(2004)01-0016-05

0 引言

基于人工智能理论的暂态稳定评估(TSA)是一个典型的模式分类问题。影响电力系统暂态稳定的诸多因素,以及由SCADA收集到的现场海量的运行数据,在很大程度上都会导致分类器运行速度和识别能力的下降。因此,如何有效地提取和选择输入特征变量,压缩输入空间的大小以改善分类器设计,提高稳定判断的准确性已成为一个亟待解决的问题。

特征提取和选择的基本任务就是从许多特征中找出那些最有效的特征,去除与分类目标无关的或与其他特征量有较高相关性的冗余特征[1]。其中,特征提取是将原始数据构成的高维空间映射(或变换)为一个低维的样本空间;特征选择则指从一组特征中挑选出一些最有效的特征以降低特征空间维数。围绕这两项任务,本文首先采用主成分分析法对原始数据进行特征提取,将原有的高维样本空间映射为一个低维空间,然后阐述了以类内类间距离作为类别可分离判据,利用遗传算法进行特征选择的基本原理和应用实例。

1 基于主成分分析的特征提取

主成分分析法(Principal Component Analysis, PCA)是模式识别中一种有效的特征提取方法,其目的是用较少数量的特征对样本进行描述,降低特征空间维数,同时保留原始数据的主要信息。对于大样本、多变量的情况,该方法尤为有效。通常,对数据集 X,主成分的求解常转化为求 X 的协方差矩阵的特征根和其标准正交向量的问题,过程如下:

(1) 对原始数据样本集 $X = (x_{ij})_{n\times p}$ (n 为样本数, p 为输入特征数)进行标准化处理,得到 X′,即

$$x_i' = (x_i - \mu_i)/\sigma_i, \quad i \in [1, \cdots, p] \qquad (1)$$

式中 μ_i、σ_i 分别为特征变量 x_i 的均值和标准差;

(2) 建立标准化数据 X 的协方差矩阵 V,求 V 的 k ($k \le p$) 个不为0的特征值,并按特征值大小排序,得到 $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k > 0$ ($\lambda_{k+1} = \lambda_{k+2} = \cdots = \lambda_p = 0$) 和对应的标准正交特征向量。设对应于特征值 λ_i 的特征向量为 $u_i = (u_{i1}, u_{i2}, \cdots, u_{ip})'$, $i = 1, 2, \cdots, k$,即 $U = (u_1, u_2, \cdots, u_k)$;

(3) 计算主成分 $y_1, y_2, \cdots, y_k$

$$y_i = X'u_i, \quad i \in [1, 2, \cdots, k],\text{并记 } Y = (y_1, y_2, \cdots, y_k) = (X'u_1, X'u_2, \cdots, X'u_k) = X'u \qquad (2)$$

(4) 确定主成分的个数

根据给定阈值 L (在85%～95%之间取值),取累计贡献率 $\sum_{i=1}^{m}\lambda_i \Big/ \sum_{j=1}^{k}\lambda_j > L$ 对应的前 m ($m \le k$) 个主成分。

从上述过程可见,主成分分析法通过对现有样本空间进行某种正交变换组合,产生了一个新的样本空间。和原样本空间相比,新样本空间维数降低,特征变量间的相关性减小。

2 利用遗传算法进行特征选择

在第1条中,主成分的选择是依照传统方法,在 k 维主成分空间中选择前 m 个较大特征值所对应的特征向量(主分量特征)进行计算得到的;也有人提出应选择较小特征值对应的特征向量(次分量特征)计算得到主成分;还有意见认为应按 $\lambda_1 + \frac{1}{\lambda_1} \ge \lambda_2 + \frac{1}{\lambda_2} \ge \cdots$ 的顺序选取对应特征向量,计算主成

分。这些选取方法都缺乏一般的理论支持。为此,本文提出采用遗传算法进行特征选择,利用遗传算法的全局寻优能力,搜索最优的特征组合。

遗传算法是通过模拟生物进化过程中的繁殖、变异和自然选择来求解最优化问题。利用遗传算法进行特征选择的过程为:

(1) 令进化代数 $t=0$,生成初始群体 $P(t)$。各种遗传算法的实施过程基本类似,所不同的是针对问题的具体编码方式和适应度函数的实现过程。特征选择问题是从数据样本最初的 D 个特征变量中选择出其中的 d 个特征。在用遗传算法解决这个问题时,可采用二进制染色体编码,即用一个 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。为了加快收敛,在产生初始群体时,假设绝大多数特征变量都将被选择,对字符串中的每一位以 0.9 的概率取值为 1。

(2) 定义适应度函数。为了得到对稳定判断最有效的特征,本文采用类内类间距离判据 J[1]作为适应度函数。具体定义为:

$$J = tr(S_b)/tr(S_w) \tag{3}$$

其中,$S_b=\sum_{i=1}^{c}P_i(m_i-m)^{T}(m_i-m)$,为类间离散度矩阵;$S_w=\sum_{i=1}^{c}P_i\frac{1}{n_i}\sum_{k=1}^{n_i}(x_k^{(i)}-m_i)^{T}(x_k^{(i)}-m_i)$,为类间离散度矩阵;$m_i=\frac{1}{n_i}\sum_{k=1}^{n_i}x_k^{(i)}$,$m=\sum_{i=1}^{c}P_i m_i$,$c$ 为类别数,n_i 为第 i 类中的样本数,P_i 是相应第 i 类的先验概率,$x_k^{(i)}$ 是第 i 类中的 D 维特征向量。

在上式中,如果表示同类样本之间距离的 S_b 的值越小,表示异类样本间距离的 S_w 的值越大,则 J 值越大,此时的分类效果也越好。

(3) 计算 $P(t)$ 中每个个体的适应度函数值。将第一次迭代具有最高适应度值的个体作为第一次迭代的最优解,记录其适应度值。对于第二次迭代以上的个体,将这一代计算所得的最大适应度值与记录适应度值相比较。若小于记录值,则记录值保持不变;若大于记录值,则将这一代具有最大适应度值的个体作为群体中的最优解,修改记录值。

(4) 进行选择、交叉和变异操作,产生下一代。

(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。

在初始群体中,因为个体随机选取,所以其分布散度最大,随着进化的进行,群体的分布散度逐步减小,为此本文利用群体熵来刻划群体个体在进化过程中分布散度的衰减过程。

$$E=\sum_{i=1}^{D}[p_i\log p_i+(1-p_i)\log(1-p_i)] \tag{4}$$

p_i 为群体中第 i 位为 1 的频率,D 为特征维数。

经过上述特征提取和特征选择,暂态稳定评估的输入空间由最初的 p 维下降到了 d 维($d\leq p$),至此,暂态稳定评估问题就转化为 d 维输入空间中的分类问题。

3　输入特征变量的确定

在暂态稳定分析中,所选特征变量应符合:①适用于不同规模的电力系统,避免出现"维数灾"问题;②适用于在线计算。就暂态稳定问题而言,它同电力系统的运行方式有着密切联系。因此,这里选用反映电力系统运行状态的数据及其统计值作为特征变量,考虑到系统的运行工况、故障形式、故障点位置及故障切除后系统的结构等因素均有可能对暂态稳定产生影响,特征变量定义如下(变量中的出力、负荷均包括有功、无功两部分):

(1) 系统总的发电出力;

(2) 根据电网络中的联络线,将系统划分为若干区域后每个区域的总发电出力;

(3) 系统的总负荷;

(4) 各区域的总负荷;

(5) 系统的电压最大、最小值及其对应的母线号;

(6) 联络线(或断面)上的传输功率;

(7) 有功传输功率最大的线路(或断面)标号;

(8) 最大的线路(或断面)有功传输功率及与此对应的无功传输功率;

(9) 系统有功、无功网损;

(10) 故障发生位置(定义为故障线路 + 出口位置);

表示结构的特征变量,详见文献[2]:

(11) $I_Z=(\sum_{j=1}^{NG}\Delta Z_{ij})-(\sum_{j=1}^{NL}\Delta Z_{ik})$ (5)

式中 i 为薄弱节点号,j 为与薄弱节点联系紧密的发电机节点号,k 为主要负荷节点号,NG 为发电机个数,NL 为主要负荷数。

$\Delta Z_{ij}=|Z_{ij}'-Z_{ij}|$,　$\Delta Z_{ik}=|Z_{ik}'-Z_{ik}|$

Z_{ij}、Z_{ik} 分别是网络操作前 i、j 与 i、k 节点间的互阻抗;Z_{ij}'、Z_{ik} 分别是网络操作后 i、j 与 i、k 节点

间的互阻抗。

$$I_L = \Delta Z_{km}' - \Delta Z_{km} \quad (6)$$

$$\Delta Z_{km}' = |Z_{ik}' - Z_{im}'|, \quad \Delta Z_{km} = |Z_{ik} - Z_{im}|$$

式中 i 是被切除线路的首端节点号，k、m 是薄弱线路的首、末节点号；Z_{ik}、Z_{im} 分别是网络操作前 i、k 与 i、m 节点间的互阻抗；Z_{ik}'、Z_{im}' 分别是网络操作后 i、k 与 i、m 节点间的互阻抗。

4 算例分析

4.1 IEEE-9 仿真结果

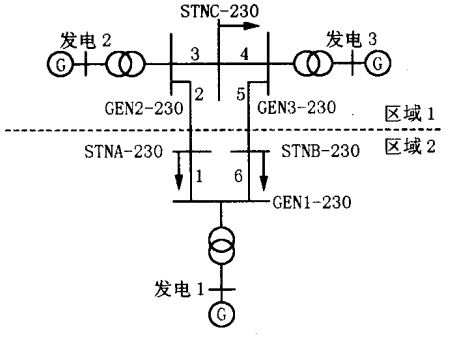

图 1 IEEE-9 电力系统接线图

Fig. 1 Connection of IEEE-9 power system

系统如图 1 所示，联络线 2、5 将该系统分为区域 1 和区域 2；本系统选择了 25 个初始特征变量，见表 1；在 75 %～120 %(以 5 %为变化步长)基准负荷下，对每一负荷条件随机设置 5 种不同的发电出力；设定系统每条线路两端发生三相短路故障，进行 $N-1$ 校验，0.15 s 跳开故障线路切除故障，共得到 600 个观测样本。随机抽取 480 个样本作为训练样本，其余 120 个样本用来测试。

考虑到电力系统的运行有明显的规律性和重现性，根据最大化类内相似性、最小化类间相似性的原则，首先对上述观测数据进行动态聚类分析。基于系统所具有的基本运行方式和 4 种典型运行方式(夏小、夏大、冬小、冬大)，将训练样本聚成 5 类，根据每一测试样本与每个聚类中心的距离，将测试样本分成对应的 5 个样本集。接着对每一训练-测试子集先后进行主成分分析和 GA 特征选择。在遗传算法实验中，群体大小设为 50，终止进化代数为 25，采用轮盘赌方法选择个体，交叉概率为 0.6，变异概率为 0.1。

以第一组训练-测试子集为例，对除 Tz17 和 Tz23 之外的 23 个连续特征变量进行主成分分析，连续特征变量数下降为 12 个，之后执行 GA 运算，又从这 12 个连续特征变量中选择出了其中的 7 个。图 2 为进化过程中最优个体的适应度变化趋势，图 3 为群体熵的变化趋势。可见随着进化的进行，群体中的个体迅速趋同，最终趋于最优解，证明所选择的交叉概率和变异概率是恰当的。

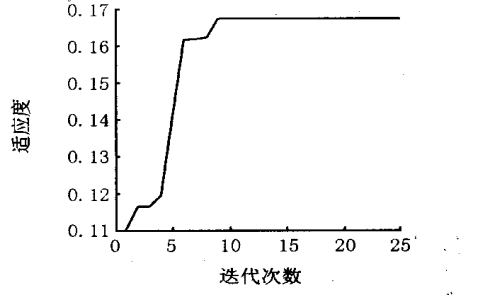

图 2 IEEE-9 系统进化过程中适应度的变化趋势

Fig. 2 Evolution of the optimal fitness of IEEE-9 power system

利用关联分类法[3]对压缩前后的数据进行训练测试。仿真结果表明结合 PCA 特征提取和 GA 特征选择最终得到的 9 个综合特征变量保持了原有

表 1 IEEE-9 系统输入特征变量

Tab. 1 Input characteristic variables of IEEE-9 power system

序号	特征	序号	特征	序号	特征
Tz1	整个系统的有功出力	Tz 9	区域 1 的有功负荷	Tz 17	有功传输功率最大的线路标号 *
Tz2	整个系统的无功出力	Tz10	区域 1 的无功负荷	Tz18	*对应的线路的有功传输功率
Tz 3	区域 1 的有功出力	Tz 11	区域 2 的有功负荷	Tz 19	*对应的线路的无功传输功率
Tz 4	区域 1 的无功出力	Tz 12	区域 2 的无功负荷	Tz 20	系统有功网损
Tz 5	区域 2 的有功出力	Tz 13	联络线 2 上的有功传输功率	Tz 21	系统无功网损
Tz 6	区域 2 的无功出力	Tz 14	联络线 2 上的无功传输功率	Tz 22	各节点的电压最小值
Tz 7	整个系统的有功负荷	Tz 15	联络线 5 上的有功传输功率	Tz 23	故障类型
Tz 8	整个系统的无功负荷	Tz 16	联络线 5 上的无功传输功率	Tz 24	I_L
				Tz 25	I_z

25 个特征变量的分类能力,同时将训练数据集压缩了 64 %;在压缩前后,二者的稳定判断正确率完全一致,均为 98 %。

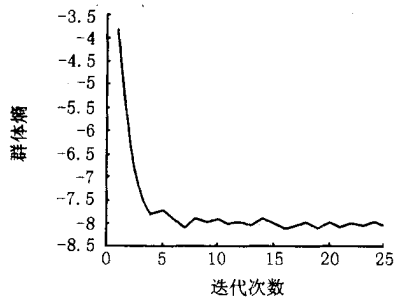

图 3　IEEE - 9 系统进化过程中群体熵的变化趋势

Fig. 3　Evolution of the population entropy of IEEE - 9 power system

4.2　IEEE - 39 仿真结果

为了验证本文提出的方法对不同系统具有有效性,按照 4.1 中的步骤,对 IEEE - 39 系统进行了仿真试算。系统如图 4 所示,其中节点 2 - 3、8 - 9、26 - 27、4 - 14、6 - 11、16 - 17 之间的联络线 3、14、31、8、12、21 将该系统分为区域 1、2 和 3;本系统选择了 40 个初始特征变量,见表 2;设线路在靠近母线 2、4、6、8、10、11、14、15、16、18、19、21、22、23、24、25、26、27、29 侧发生三相短路故障,通过 0.15 s 跳开故障线路切除故障,共得到 1000 个观测样本。随机抽取 600 个样本作为训练样本,200 个样本作为测试样本。

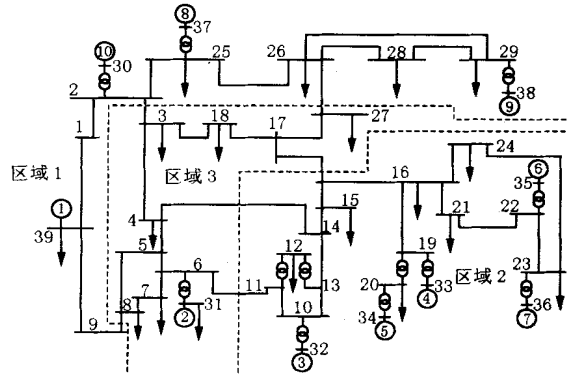

图 4　IEEE - 39 电力系统接线图

Fig. 4 Connection of IEEE - 39 power system

以第一组训练 - 测试子集为例,对连续特征变量 Tz1～Tz36 进行主成分分析,连续特征变量数下降为 11 个;在执行 GA 运算后,又从这 11 个连续特征变量中选择出了其中的 3 个。图 5、6 分别为进化过程中最优个体的适应度和群体熵的变化趋势。

表 2　IEEE - 39 系统输入特征变量

Tab. 2　Input characteristic variables of IEEE - 39 power system

序号	特征	序号	特征	序号	特征
1	整个系统的有功出力	14	区域 2 的无功负荷	27	联络线 5 上的有功传输功率
2	整个系统的无功出力	15	区域 3 的有功负荷	28	联络线 5 上的无功传输功率
3	区域 1 的有功出力	16	区域 3 的无功负荷	29	联络线 5 上的有功传输功率
4	区域 1 的无功出力	17	母线电压最大值	30	联络线 5 上的无功传输功率
5	区域 2 的有功出力	18	母线电压最小值	31	有功传输功率最大的线路标号 *
6	区域 2 的无功出力	19	联络线 3 上的有功传输功率	32	* 对应线路上的有功传输功率
7	区域 3 的有功出力	20	联络线 3 上的无功传输功率	33	* 对应线路上的无功传输功率
8	区域 3 的无功出力	21	联络线 5 上的有功传输功率	34	系统有功网损
9	整个系统的有功负荷	22	联络线 5 上的无功传输功率	35	系统无功网损
10	整个系统的无功负荷	23	联络线 5 上的有功传输功率	36	I_L
11	区域 1 的有功负荷	24	联络线 5 上的无功传输功率	37	I_Z
12	区域 1 的无功负荷	25	联络线 5 上的有功传输功率	38	对应最大电压的母线标号
13	区域 2 的有功负荷	26	联络线 5 上的无功传输功率	39	对应最小电压的母线标号
				40	故障类型 Ftype

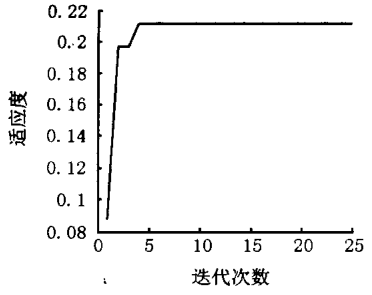

图 5 IEEE-39 系统进化过程中适应度的变化趋势

Fig. 5 Evolution of the optimal fitness of IEEE-39 power system

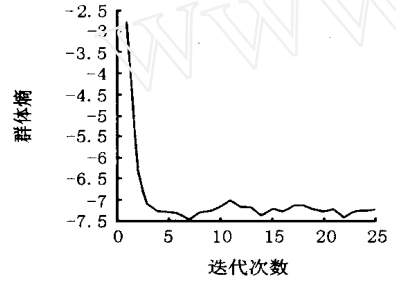

图 6 IEEE-39 系统进化过程中群体熵的变化趋势

Fig. 6 Evolution of the population entropy of IEEE-39 power system

关联分类法的仿真结果表明,结合 PCA 特征提取和 GA 特征选择最终得到的 7 个综合特征变量同原有 40 个特征变量具有相同的分类能力,同时将训练数据集压缩了 82.5 %;在压缩前后,二者的稳定判断正确率没有发生变化。

另外,从特征选择的结果可以看出,使类内类间距离判据达到最优值的特征组合不一定全由主分量特征构成,次分量特征的作用不应被轻易忽视。

5 结论

本文针对主成分分析法在主成分特征选择中存在的不确定性,提出了以类内类间距离为适应度函数,利用遗传算法对主成分分析形成的综合特征变量进行选择。该方法充分利用遗传算法全局寻优、计算速度快的特点,能高效去除原始数据中的冗余特征。仿真结果表明该方法适于不同规模的电力系统,能够处理大数据量样本,具有实用化价值。

参考文献:

[1] 边肇祺, 张学工 (BIAN Zhao-qi, ZhANG Xue-gong). 模式识别 (Pattern Recognizition) [M]. 北京: 清华大学出版社 (Beijing: Tsinghua University Press), 2000.

[2] 李卫东,唐艳丽 (LI Wei-dong, TANG Yan-li). 电力系统运行模式结构特征提取方法的研究 (A Methodology for Extracting Power System Operational Configuration Characteristics) [J]. 中国电力 (Electric Power), 1998, 31(4):29-31.

[3] LIU Bing, Hsu Wynne, MA Yi-ming. Integrating Classification and Association Rule Mining [A]. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining[C]. New York: AAAI Press, 1998. 80-86.

收稿日期: 2003-03-17; **修回日期**: 2003-05-30

作者简介:

于之虹 (1975-),女,博士研究生,研究方向为电力系统稳定分析;

郭志忠 (1961-),男,教授,博士生导师,研究领域为电力系统稳定分析,电力市场,光电互感器。

Feature selection based on genetic algorithm for transient stability assessment

YU Zhi-hong, GUO Zhi-zhong

(Dept. of Electrical Engineering, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Aimed at the disadvantages existing in feature selection by traditional combination optimization method in PCA (Principal Component Analysis), a new method based on genetic algorithm to select the input features is put forward. In this approach, the feature set to describe the system status and post-fault network configuration change are selected for transient stability assessment and the initial data is preprocessed by dynamic clustering analysis firstly. With the within-class/between-class distance criterion used as fitness function, a binary genetic algorithm is employed to select an effective subset of features forming the feature set after PCA, and the input dimension is reduced remarkably. As an example, the 3-machine 9-bus WSCC system and the 10-machine 39-bus New England system are used for simulation. The results reveals the validity of the proposed approach.

Key words: feature selection; genetic algorithm; transient stability assessment; power system