陈 果

遗传算法特征选取中的几种适应度函数构造新方法及其应用

陈 果,邓 堰

(南京航空航天大学 民航学院,南京 210016)

摘 要: 针对遗传算法特征选取技术,提出 4 种适应度函数构造方法,即,基于改进的距离判据、基于平均值方差比、基于 Fisher 判别准则以及基于最近邻分类法的适应度函数,并通过仿真实例对方法进行了验证。最后,将新方法应用于转子故障诊断,结果表明: 笔者提出的遗传算法特征选择的 4 种适应度函数的正确有效性。

关 键 词: 特征选取; 遗传算法; 适应度函数; 转子; 故障诊断

中图分类号: O322; TH113.1 **文献标识码:** A **文章编号:** 1003-8728(2011)01-0124-05

Several New Methods for Features Extraction Based on Genetic Algorithm and Their Application

Chen Guo, Deng Yan

(College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

Abstract: In this paper, aiming at feature selection based on genetic algorithm, four fitness functions are constructed, that is the function of improved distance criterion, of mean-variance ratio, of Fisher criterion and the most near neighbor classifying function. These methods are verified by a simulation example. Finally, The new methods are applied to diagnose the rotor faults, and the results show that the new methods are correct and effective.

Key words: feature selection; genetic algorithm; fitness function; rotor; fault diagnosis

在旋转机械故障诊断实践中,由于诊断对象的复杂性,故障特征和故障类别的对应关系不甚明了,人们提出了大量的原始特征以进行故障识别。但受分类器规模、训练过程的复杂性以及计算机容量等诸多因素的制约,往往不能取得良好的效果。因此,如何有效地选择特征量以改善分类器设计,提高诊断精度已成为一个亟待解决的课题[1]。

国内外大量专家学者对特征选择方法作了深入地研究,出现了各种各样的方法: 如,主成分分析法(PCA)[1]、神经网络法[2,3]、无监督聚类法[4,5]、粗糙集理论法[6]、遗传算法[7,8]、基于特征相关性和冗余性分析的特征选择方法[9]、基于动态规划方法的特征选择法[10]等。其中,遗传算法因其简单通用,鲁棒性强,适用于并行处理,已广泛应用于计算机科学、优化调度、运输问题、组合优化等领域,也被广泛应用于特征选择,并取得了较好的结果[7,8]。

但是,遗传算法的适应度函数对特征选取具有很大的影响,因此将模式识别原理应用于遗传算法的适应度函数构造,探讨遗传算法的特征选取效果,对于完善遗传算法特征选取方法具有重要意义。

1 遗传算法特征选取基本原理

遗传算法特征选择的基本原理是用遗传算法寻找一个最优的二进制编码,码中的每一位对应一个特征,若第 i 位为“1”,则表明对应特征被选取,该特征将出现在分类器中,为“0”,则表明对应特征未被选取,该特征将不出现在分类器中。其基本步骤为:

收稿日期: 2009-10-09

基金项目: 国家自然科学基金项目(50705042)和航空科学基金项目(2007ZB52022)资助

作者简介: 陈 果(1972-),教授,博士生导师,研究方向为航空发动机状态监测与故障诊断、转子动力学、智能诊断与专家系统、机器学习与知识获取、图像处理及模式识别等,cgzyx@263.net

(1) 编码。采用二进制编码方法,二进制码的每一位的值,"0"表示特征未被选中,"1"表示特征被选中。

(2) 初始群体的生成。随机产生 N 个初始串构成初始种群,通常种群数确定为 50 ~ 100。

(3) 适应度函数。适应度函数表明个体或解的优劣性。针对特征选取问题,适应度函数的构造非常重要,它主要依据类别的可分性判据以及特征的分类能力。适应度函数的有效性将直接决定遗传算法的搜索方向和进化结果。笔者将重点讨论 4 种遗传算法特征选取的适应度函数的构造方法。

(4) 将适应度最大的个体,即种群中最好的个体无条件地复制到下一代新种群中,然后对父代种群进行选择、交叉和变异等遗传算子运算,从而繁殖出下一代新种群其它 $n-1$ 个基因串。通常采用转轮法作为选取方法,适应度大的基因串选择的机会大,从而被遗传到下一代的机会大,相反,适应度小的基因串选择的机会小,从而被淘汰的机率大。交叉和变异是产生新个体的遗传算子,交叉率太大,将使高适应度的基因串结构很快被破坏掉,太小则使搜索停止不前,一般取为 0.5 ~ 0.9。变异率太大,将使遗传算法变为随机搜索,太小则不会产生新个体,一般取为 0.01 ~ 0.1。

(5) 如果达到设定的繁衍代数,返回最好的基因串,并将其作为特征选取的依据,算法结束。否则,回到(4)继续下一代的繁衍。

2　4 种新的遗传算法特征选取适应度函数

2.1　基于改进的距离判据适应度函数

基于距离的可分性判据直接依靠样本计算,直观简洁,物理概念清晰,因此目前应用较为广泛。基于距离的可分性判据的出发点是: 各类样本之间的距离越大、类内散度越小,则类别可分性越好。

给定一组表示联合分布的模式集(训练集),假定每一类的模式向量在观察空间中占据不同的区域是合理的,类别模式间的距离或平均距离则是模式空间中类别可分离性的度量。定义:

(1) 总体类内散布矩阵

$$S_w = \sum_{i=1}^{C} P(\omega_i)\, E\left[(X^{(i)} - M_i)(X^{(i)} - M_i)^{\mathrm{T}}\right] \quad (1)$$

(2) 总体类间散布矩阵

$$S_b = \sum_{i=1}^{C} P(\omega_i)(M_i - M)(M_i - M)^{\mathrm{T}} \quad (2)$$

式中: C 为类别数; $M_i = \frac{1}{N_i}\sum_{k=1}^{N_i} X_k^{(i)}$ 为第 i 类均值向量; $M = \frac{1}{N}\sum_{i=1}^{N} X_i = \sum_{i=1}^{C} P(\omega_i) M_i$ 为样本集总的均值向量; $P(\omega_i)$ 为第 i 类的先验概率。

显然,如果同类样本之间的距离越小,而异类样本之间的距离越大,则分类效果越好。于是分别以类内散度矩阵 S_w 和类间散度矩阵 S_b 的迹来度量以上二距离,进而给出类内-类间距离判据 J。

$$J = \frac{tr(S_b)^n}{tr(S_w)} \quad (3)$$

式中: 参数 n 用于调节类内散度矩阵 S_w 的迹和类间散度矩阵 S_b 的迹对特征分类性能的贡献率。根据实际情况,n 可以取大于 0 的任意值,当 $n=1$ 即为文献[3]的适应度函数。

2.2　基于平均值方差比的适应度函数

特征选择的一种可行技术是基于平均值和方差的比较[11]。通常,数据特征值分布的均值和方差是关于特征的重要信息。通常,如果一个特征描述了不同种类的实例,则可以检查不同种类的样本。用特征的方差对特征的均值进行标准化,然后在不同类之间进行比较。如果均值偏离很远,此特征的重要性增加。如果均值是不可区分的,这个特征的重要性减弱。

设 A 和 B 是测量两个不同类特征的值的集合,n_1 和 n_2 是相应的样本数。特征 A 和 B 的方差和定义为

$$S_{\text{特征}} = \sqrt{\frac{\mathrm{var}(A)}{n_1} + \frac{\mathrm{var}(B)}{n_2}} \quad (4)$$

特征重要性定义

$$F_{\text{重要性}} = \frac{|\,\mathrm{mean}(A) - \mathrm{mean}(B)\,|}{S_{\text{特征}}} \quad (5)$$

将平均值方差法应用于构造遗传算法的适应度函数,其基本步骤为:

设样本类别数为 C,对某染色体个体,所选取的特征为 M 个,则按式(5)可以得到 M 个特征对每类的重要性指标,从而构成特征重要性指标矩阵 F,即

$$F = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1M} \\ F_{21} & F_{22} & \cdots & F_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ F_{C1} & F_{C2} & \cdots & F_{CM} \end{bmatrix}$$

所以,从特征重要性矩阵 F 可以计算特征重要性指标平均值 F_{AVG},即

$$F_{\mathrm{AVG}} = \frac{1}{C \times M}\sum_{i=1}^{C}\sum_{j=1}^{M} F_{ij} \quad (6)$$

同时,从特征重要性矩阵 F 统计出大于 F_{AVG} 的特征重要性指标之和,即

$$S_F = \sum_{i=1}^{C}\sum_{j=1}^{M} F_{ij},\ \text{if } F_{ij} > F_{\mathrm{AVG}} \quad (7)$$

构造适应度函数为

$$J = S_F \times F_{\mathrm{AVG}}^n, \quad n > 0 \tag{8}$$

显然,特征重要性指标平均值越大表明特征所选择的特征分类能力越强,同时平均值以上的特征重要性指标之和也反映了所选择特征的重要性。因此所构建的适应度函数 J 反映了所选择特征的分类能力。其中指数 n 的作用是平衡特征重要性指标平均值 F_{AVG}和大于 F_{AVG}的特征重要性指标和 S_F,通常取为 2 ~5。笔者选取 $n=2$。

2.3 基于 Fisher 准则的适应度函数

在模式识别理论中,Fisher 准则是寻找一个投影方向,使样本在该方向上的分类效果最好。下面讨论在遗传算法中利用该准则来构造适应度函数。

设样本类别数为 C,对某染色体个体,所选取的特征为 M 个。

Step 1 对第 i 类,将样本分为两类,即,属于第 i 类的样本为一类,其他样本为另一类,则最佳投影方向为 $Y = S_w^{-1}(m_1^* - m_2^*)$,在最大投影方向得到的 $J_i(Y)_{\max} = \dfrac{Y^{\mathrm{T}} S_B Y}{Y^{\mathrm{T}} S_w Y}$。其中,$m_1^* = \dfrac{1}{n_1}\sum_{X\in\omega_1} X$,$m_2^* = \dfrac{1}{n_2}\sum_{X\in\omega_2} X$ 分别为第 1 类和第 2 类均值向量。

S_w 为总体类内散布矩阵

$$S_w = \sum_{i=1}^{2} P(\omega_1)\, E\left[(X^{(i)} - m_i^*)(X^{(i)} - m_i^*)^{\mathrm{T}}\right] \tag{9}$$

S_b 为总体类间散布矩阵

$$S_b = \sum_{i=1}^{2} P(\omega_i)(m_i^* - m)(m_i^* - m)^{\mathrm{T}} \tag{10}$$

式中: $m = \dfrac{1}{N}\sum_{i=1}^{N} X_i = \sum_{i=1}^{2} P(\omega_i)\, m_i^*$ 为样本集总的均值向量; $P(\omega_i)$ 为第 i 类的先验概率。

Step 2 i 从 1 到 C,重复 Step 1,可以得到 $J_i(Y)_{\max}$,$i=1,2,\cdots,C$。

Step 3 构造适应度函数

$$J = \frac{\sum_{i=1}^{C} J_i(Y)_{\max}}{M^n} \tag{11}$$

显然,该适应度函数反映了所选择特征对分类能力的平均贡献率,其中,n 为平衡分类能力和特征数目的参数,通常取 $0 < n < 1$,笔者选取 $n=0.5$。

2.4 基于最近邻分类法的适应度函数

近邻法是一种非参数模式识别方法,属于有监督学习,我们可以利用它的分类识别率作为特征评价函数。最近邻法的基本思想很简单,设有 n 个样本 $\chi = \{X_1, X_2, \cdots, X_n\}$ 每个样本都已标以类别标志。如果在这 n 个样本中与待分类样本 X 相距最近的一个样本为 $X' \in \chi$,则把 X 分到 X' 所在的类别中去。利用最近邻法构造适应度函数的步骤为:

Step 1 将样本随机分为训练样本和测试样本集。

Step 2 对每一特征组合,利用所选取的特征重新构造训练样本和测试样本,运用最近邻法对测试样本进行识别,得到识别率 R。

Step 3 考虑所选择的特征数目 M,则构造适应度函数为

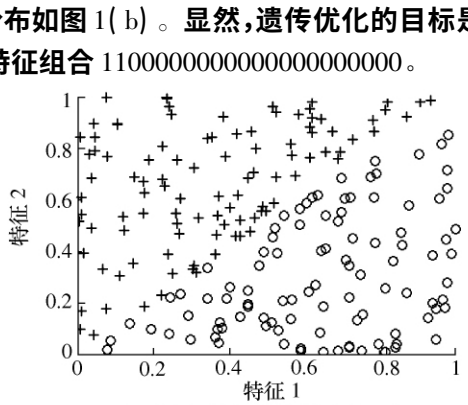
$$J = R^{(1+M^n)} \tag{12}$$

由于 $0 \leqslant R \leqslant 1$,因此,特征数 M 越小、识别率越大,则适应度函数 J 值越大,这正好符合特征选取的原则,即用最少的特征获取最高的识别率。其中,n 为平衡特征数目和识别率权重的参数,通常,$0 \leqslant n \leqslant 1$,笔者选取 $n=0.5$。

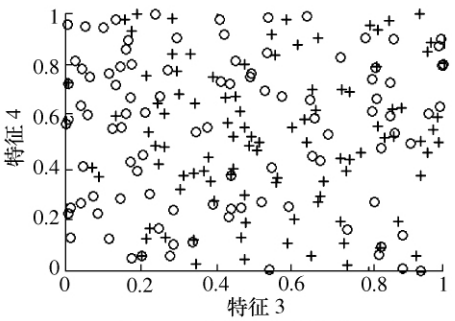
3 仿真实例分析

3.1 仿真样本

为了测试遗传算法的寻优能力,作了如下仿真试验。样本集由两类样本组成,各有 100 组样本,原始特征集的维数为 22,其中前两个特征是有效的分类特征,后 20 个特征量的取值为 0 到 1 之间的随机数,不含有任何分类信息。各样本在特征 1 和特征 2 张成的空间的分布如图 1(a) 所示,特征 3 和特征 4 的分布如图 1(b)。显然,遗传优化的目标是得到最优特征组合 11000000000000000000。



(a) 样本在前两维空间中的分布



(b)样本在第 3、4 维空间的分布

图 1 仿真样本分布图

3.2　不同适应度函数下的特征选取实验

本文遗传算法的参数为: 种群数 100,进化代数 100,交叉率为 0.5,变异率为 0.05。图 2 为以式(3)的距离判据作为适应度函数时的最优特征组合进化图,其中图 2(a)和图 2(b)分别为 $n=1$ 和 $n=2$ 的进化过程。

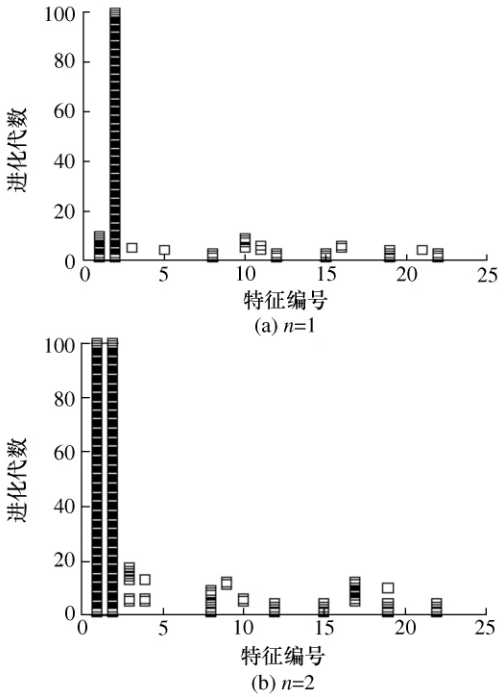


图 2　以距离判据作为适应度函数时的最优特征组合进化图

表 1 为以距离判据作为适应度函数得到的前 3 个最优特征组合。从计算结果可以看出,当 $n=2$ 时,遗传算法得到的最优特征正好是我们要寻找的目标,而当 $n=1$ 时,遗传算法得到的最优特征并不是真正的最优特征组合。而文献[8]的适应度函数正是 $n=1$ 时的情形,显然其特征选择结果并不可靠。显然,笔者的距离判据中的参数 n 对获取最优特征组合具有重要作用。它可以根据实际情况选取大于 0 的值,通过比较研究,笔者选取 $n=2$。

表 1　以距离判据作为适应度函数得到的前 3 个最优特征组合

参数 n	适应度	前 3 个最优特征组合
1	0.6276	0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	0.5040	1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	0.3864	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2	0.0273	1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	0.0171	1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
	0.0165	1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0

图 3(a) ~ 图 3(c) 分别为以平均值方差比、Fisher 准则和最近邻分类构造适应度函数时,遗传算法最大适应度的进化过程; 图 4(a) ~ 图 4(c) 分别为以平均值方差比、Fisher 准则法和最近邻分类法构造适应度函数时,所得到的最优特征组合随进化代数变化的趋势图。从图 2、图 3、和图 4 中不难看出,3 种适应度函数下的遗传算法特征选取,均得到了期望的特征组合,由此验证了笔者方法的正确有效性。

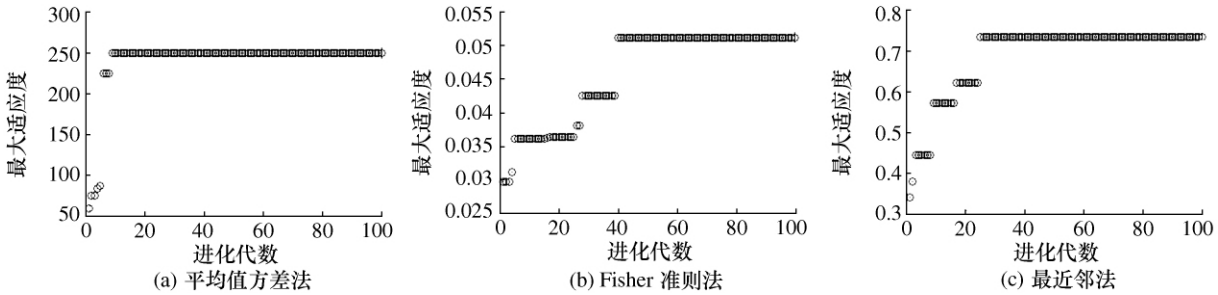


图 3　最大适应度进化图

4　转子故障诊断实验分析

笔者利用 ZT-3 多功能转子故障模拟实验台和 DHDAS 信号测试分析系统获取了不同转速下的不平衡故障样本 26 个、碰摩样本 29 个及油膜涡动样本 40 个。

由于多分类问题可以转化为两分类问题来考虑,且两类问题具有的网络结构简单,训练样本要求少,训练时间大大缩短等优点,笔者将 3 分类问题转化为 3 个 2 分类问题,由各子网络分别负责诊断一种故障,最后根据 3 个神经网络的输出的最大值来判断最终故障类型。

由于神经网络训练的不确定性,以及泛化能力受结构参数影响很大的,为了获得具有最大泛化能力的神经网络结构和参数,笔者利用文献[14]提出的结构自适应神经网络来自动获取最佳的子网络结构参数,其基本思想是: 将故障样本随机选择一半作为训练样本,剩余的作为测试样本,用训练样本训练网络,再用测试样本对训练好的网络测试,用所产生

的测试误差构造遗传算法的适应度函数，利用遗传算法的全局搜索能力自动获取测试误差最小，且具有最佳泛化能力的网络结构参数。其中结构自适应神经网络参数统一为：种群数 10，进化代数 10，交叉率为 0.5，变异率为 0.05。

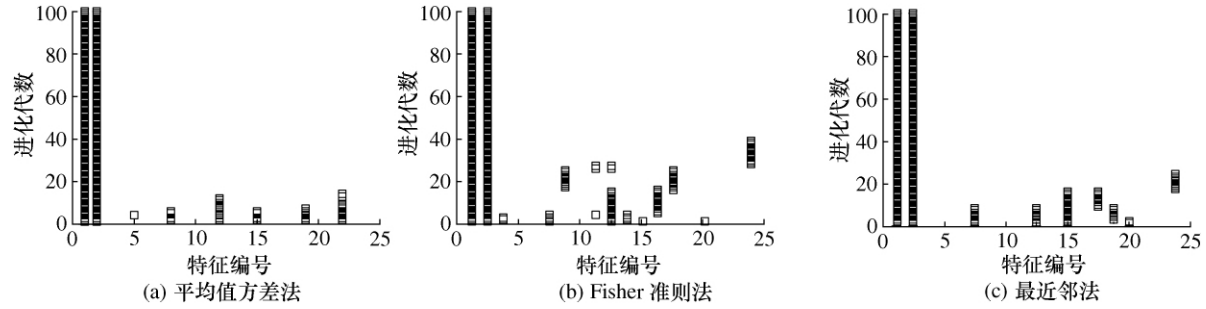


图 4 最优特征组合进化图

笔者将故障样本随机分为训练样本集(32 个样本)、测试样本集(31 个样本)及未知样本集(32 个)，每个样本集中均包含了不平衡、碰摩以及油膜涡动 3 种故障。其中训练样本直接参与神经网络训练；测试样本参与对每次训练好的神经网络的测试，其识别率作为遗传算法的适应度函数；未知样本参与对优化好的神经网络的识别率的测试，作为最佳神经网络的分类性能评价依据。

用频谱分析法提取 17 个频谱特征，即 0.2×，0.25×，0.33×，0.43×，0.5×，0.67×，0.75×，1×，2×，3×，4×，5×，6×，7×，8×，9×，10×。分别用不同的适应度函数进行遗传算法特征选取，特征选取结果及所选取的特征识别结果如表 2 所示。

表 2 结构自适应神经网络对频谱特征的识别结果

选取方法		最优特征组合	特征数	子网络的识别率(%)			对测试样本识别率(%)	对未知样本识别率(%)
				NN1	NN2	NN3		
不选取(原始特征)		1111111111111111	17	100%	96.8%	100%	96.8%	90.6%
适应度函数	改进距离	00111001111100100	9	100%	96.8%	100%	96.8%	96.9%
	平均值方差比	00011001101001100	7	100%	96.8%	96.8%	96.8%	93.8%
	Fisher 准则	00011000011010000	5	100%	100%	96.8%	96.8%	93.8%
	最近邻分类	10001001001000000	4	100%	93.5%	96.8%	93.5%	96.9%

注：NN1——不平衡子网络；NN2——碰摩子网络；NN3——油膜涡动子网络。

从表 2 可以看出，在 17 个频谱特征下，对未知样本的识别率为 90.6%；基于改进距离判据构造的适应度函数，用遗传算法优化后特征数缩减为 9 个，对未知样本的识别率为 96.9%；基于平均值方差比判据构造的适应度函数，用遗传算法优化后特征数缩减为 7 个，对未知样本的识别率为 93.8%；基于 Fisher 准则构造的适应度函数，用遗传算法优化后特征数缩减为 5 个，对未知样本的识别率为 93.8%；基于最近邻分类法构造的适应度函数，用遗传算法优化后特征数缩减为 4 个，对未知样本的识别率为 96.9%。由此可见，笔者 4 种遗传算法适应度函数均能对转子故障特征实现有效的缩减，并且不会降低对故障的识别率，在转子故障诊断特征提取中具有很好地应用前景。

5 结论

(1) 针对遗传算法特征选取方法，根据模式识别理论，分别基于改进的类内-类间距离判据、平均值方差比、Fisher 判别准则及最近邻分类法构造了 4 种适应度函数，并运用仿真算例进行了验证，表明了笔者方法的正确有效性。

(2) 笔者将遗传算法特征提取方法应用于转子故障诊断，对 17 个频谱特征进行了特征选取，并依据结构自适应神经网络进行了识别。结果表明，本文 4 种遗传算法特征选取方法能够在保证不低于原始特征分类能力的基础上大幅度地缩减原始特征数。本文方法简洁实用，可以应用于基于大量特征的转子故障诊断中。

(下转第 132 页)

$$N_p = \frac{p_p \times Q_P}{\eta} = \frac{20.15 \times 10^3 \times 32.6}{0.8 \times 10^3 \times 60} = 13 \text{ kW} \quad (11)$$

因调平时所需的总流量较小，选用电机 Y160M-4 (11 kW) B5-V1。

(3) 控制阀的选择

一般选择控制阀的额定流量应比系统管路实际通过的流量大一些，必要时，允许通过阀的最大流量超过其额定流量的20%。集成块与料斗升降、卸料开门选用10通径的阀。

选择电液比例阀 PVG32-4 联阀组，它的流量为10 L/min，压力为25 MPa。

(4) 管道尺寸的确定

管道尺寸由选定的标准元件连接口尺寸确定。

(5) 油箱容量的确定

本系统属于中压系统，按经验公式[6]计算 $V = (3 \sim 5) Q_p = 4 \times 32.6 = 130.4$ L，考虑到集装箱内的空间较小，集装箱内有空调，室内温度较低，确定油箱实际容量 $V_s = 180$ L。

3 结论

(1) 平台下落时，由于平台过重，平台抖动厉害，响声异常，采用靠平台的自重下落，通过 PLC 给电磁比例阀信号，控制液控锁的开口，通过给信号量的大小控制了平台下降速度，还避免了平台下落时的抖动和油缸下降时发出的响声。

(2) 成品料斗升降过程中，料斗出现明显的抖动，用平衡阀代替液控锁，完全解决了该问题。

(3) 用手打泵给防摆油缸自动加油，淘汰了传统的人工给防摆油缸灌油过程，有效地提高了的工作效率和系统的可靠性。

参考文献

[1] 宋钢. 沥青水泥砂浆车在无咋轨道施工中的应用[J]. 现代城市轨道交通，2008，(4)：42～45

[2] 江创华，韩国梁，李忠元等. 高速铁路无咋轨道沥青水泥砂浆车[J]. 工程机械：2009，40：7～9

[3] 侯友山，石博强，于淼等. TL345J 铰接式自卸车液压系统设计[J]. 液压与机床：2009，37(3)：89～92

[4] 陈俊，刘剑雄，陈雪菊. 钢卷翻转输送机的液压系统设计[J]. 机电产品开发与创新，2009，22(1)：36～37

[5] 雷天觉. 液压工程手册[M]. 北京：机械工业出版社，1998

[6] 于连科，王伟. 平板硫化机附机液压泵站的设计[J]. 辽宁工学院学报床，2001，21(1)：39～42

(上接第128页)

参考文献

[1] 肖建华. 智能模式识别方法[M]. 广州：华南理工大学出版社，2006

[2] Michael Egmont-Petersen, Talmon J L. Assessing the importance of feature for multi-layer perceptrons[J]. **Neural Network**, 1998, 11(4): 623～635

[3] 高仁祥，张世英，刘豹. 基于神经网络的变量选择方法[J]. 系统工程学报，1998，13(2)：32～37

[4] Law M H C, *et al*. Simultaneous feature selection and clustering using mixture model[J]. **IEEE Transaction on Pattern Analysis and Machine Intelligence**, 2004, 26(9): 1154～1166

[5] 张莉等. 基于K-均值聚类的无监督的特征选择方法[J]. 计算机应用研究，2005，(3)：23～24

[6] Kuncheva L I. Fuzzy rough sets application to feature selection[J]. **Fuzzy Sets and Systems**, 1992, 51(2): 147～153

[7] Jack L B, Nandi A K. Feature selection for ANNs using genetic algorithms in condition monitoring[A]. **ESANN'1999 Proceedings-European Symposium on Artificial Neural Networks**[C], Bruges(Belgium), 1999: 313～318

[8] 史东锋，屈梁生. 遗传算法在故障特征选择中的应用研究[J]. 振动、测试与诊断，2000，20(3)：171～176

[9] 王新峰，邱静，刘冠军. 基于特征相关性和冗余性分析的机械故障特征选择研究[J]. 中国机械工程，2006，17(4)：379～382

[10] 章新华. 一种特征选择的动态规划方法[J]. 自动化学报，1998，24(5)：675～680

[11] Kantardzic M. **Data Mining Concepts, Models, Methods, and Algorithms**[M]. New York: IEEE Press, 2002

[12] Kohavi R, John G H. Wrappers for feature subset selection[J]. **Artificial Intelligence**, 1997, 97(1～2): 273～324

[13] 周志红，周新聪，袁成清. 基于过滤器-封装器组合模型的故障特征选择算法[J]. 中国机械工程，2007，18(16)：1988～1991

[14] 陈果. 一种实现结构风险最小化思想的结构自适应神经网络模型[J]. 仪器仪表学报，2007，28(10)：1874～1879