基于最大树聚类的多超球体一类分类算法及其应用研究

刘丽娟　陈　果
南京航空航天大学，南京，210016

摘要：提出了一种基于最大树聚类的多超球体一类分类算法。首先应用最大树聚类算法将训练样本聚为多个子类，再对各子类分别进行一类支持向量机（one－class SVM，OC－SVM）分类器训练，得到由各子类对应的超球体形成的多超球体一类分类模型。分别将该方法应用于仿真数据、UCI 标准数据集以及转子故障诊断三个实例中，结果表明了该方法的有效性。

关键词：一类分类；最大树聚类；多超球体；转子；故障诊断
中图分类号：V263.6；TP277　　DOI：10.3969/j.issn.1004－132X.2012.03.003

Study on One－class Classification with Multi Hyper－spheres Based on Maximal Tree Clustering and Its Applications

Liu Lijuan　Chen Guo
Nanjing University of Aeronautics and Astronautics，Nanjing，210016

Abstract：An one－class Classification with multi hyper－spheres based on maximal tree clustering algorithm was presented herein. The training samples were firstly clustered into several sub－classes by the maximal tree clustering algorithm，and then，the sub－classes data were trained separately using one－class SVM（OC－SVM）and the multi hyper－spheres classifying models were established. The new method was applied to the instances of the simulation data set，UCI data sets and the rotor faults diagnosis，and the results show the effectiveness of the new method.

Key words：one－class classification；maximal tree clustering；multi hyper－sphere；rotor；fault diagnosis

0　引言

相对于多类分类算法对样本数量的要求较高，一类分类[1-4]方法仅仅需要一类样本对象。如状态监测与故障诊断运行状态中，相对于大量正常状态的样本，异常状态的样本往往很少，而且表现出各种各样的异常模式[5]，而其主要任务是识别状态正常与否，采用一类分类法就能有效解决该问题。

一类分类器仅需一类样本通过机器学习生成一个闭合的超球体作为该类样本的决策边界。如果测试样本点在超球体的外面，则认为这些样本点是异常样本（野点），反之则判断为正常样本。但是在实际应用中发现，即使是正常状态的训练样本，在数据分布或者结构信息上还是会存在差异（特别是当训练样本的数据是成簇分布时），如果只按照单超球体一类分类建模，那么构造的单个超球体不仅包围了训练数据，而且还包围了簇间的空白区域[6]，这样一来很可能将非正常的样本也错误地判为正常样本。虽然通过引入核函数，调节核参数（如高斯核参数）可以使上述情况有所改善，但是这无法从根本上解决问题。因此本文采用多个超球体来覆盖训练样本，研究了基于最大树聚类的多超球体[7-8]一类分类器，分别将该方法应用于仿真数据、UCI 标准数据集以及转子故障诊断三个实例中，并且与常用的基于单超球体的一类分类方法进行了比较，结果表明了该方法的有效性。

1　单超球体一类分类器

一类分类器针对一类对象（如故障诊断中的正常运行状态，为正类），而相对于该类对象的其他对象（如故障诊断中的非正常运行状态，为负类）统称异常对象（野点）。单超球体一类分类器本质上是寻找一个能够包含全部正类样本的最小超球体，在球体外的点视为野点。设有一个正类样本集 $\{x_1, x_2, \cdots, x_N\}$，将该正类样本集全部样本包围的最小球体的半径设为 R，球心设为 a，为了实现错误划分和区域范围之间的折中，在优化过程中引入松弛变量，此时样本集满足：

收稿日期：2011—03—15
基金项目：国家自然科学基金资助项目（50705042，61179057）；
航空科学基金资助项目（2007ZB52022）

$$\left.\begin{aligned}&\min\quad L(R)=R^2+C\sum_{i=1}^{N}\xi_i\\&\text{s.t.}\ (x_i-a)(x_i-a)^{\mathrm{T}}\leqslant R^2+\xi_i\\&\qquad\xi_i\geqslant 0\end{aligned}\right\}\tag{1}$$

定义 Lagrange 函数：

$$L(R,\boldsymbol{a},\Lambda,\boldsymbol{\xi})=R^2+C\sum_{i=1}^{N}\xi_i-\sum_{i=1}^{N}\alpha_i[R^2+\xi_i-(\boldsymbol{x}_i^2-2\boldsymbol{a}\boldsymbol{x}_i+\boldsymbol{a}^2)]-\sum_{i=1}^{N}\gamma_i\xi_i\tag{2}$$

其中，C 为惩罚因子，ξ_i 为对应第 i 个样本的松弛变量，$\Lambda=\{\alpha_i\}$，对应的 Lagrange 系数 $\alpha_i\geqslant 0$，$\gamma_i\geqslant 0$。将式(2) 分别对 R 和 α 求偏微分，并令其等于 0，得到相关的优化方程如下：

$$\left.\begin{aligned}&\max\ L=\sum_{i=1}^{N}\alpha_i(x_i,x_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j(x_i,x_j)\\&\text{s.t.}\ \sum_{i=1}^{N}\alpha_i=1\qquad 0\leqslant\alpha_i\leqslant C\end{aligned}\right\}\tag{3}$$

引入高斯径向基核函数 $K(\boldsymbol{x},\boldsymbol{y})$，即

$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{\sigma^2}\right)\tag{4}$$

用核函数 $K(\boldsymbol{x},\boldsymbol{y})$ 替代(x,y)，得到对应的优化方程：

$$\left.\begin{aligned}&\max\ L=\sum_{i=1}^{N}\alpha_iK(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\\&\text{s.t.}\ \sum_{i=1}^{N}\alpha_i=1\qquad 0\leqslant\alpha_i\leqslant C\end{aligned}\right\}\tag{5}$$

实际上，根据 KKT(Karush — Kuhn — Tucker) 条件，大部分 α_i 为 0，只有一小部分 $\alpha_i>0$，而与这些不为零的 α_i 所对应的样本点决定了超球体边界的构成，为此，将这些样本点称为支持对象(support objection)。

对于待定状态数据 z，其到球心的距离的平方为

$$f(z)=K(z\cdot z)-2\sum_{i=1}^{N}\alpha_iK(z,\boldsymbol{x}_i)+\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\tag{6}$$

取任一支持对象 $\boldsymbol{x}_s$，则球体半径的平方为

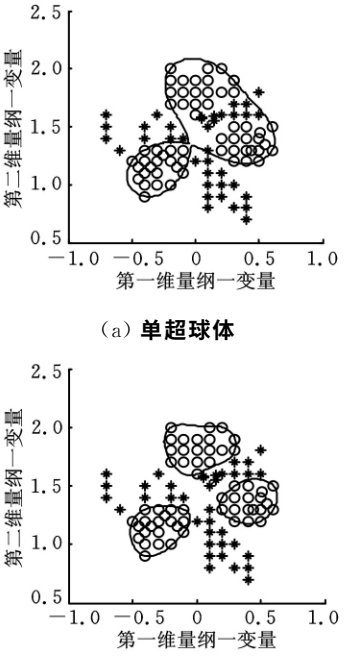
$$R^2=K(\boldsymbol{x}_s\cdot\boldsymbol{x}_s)-2\sum_{i=1}^{N}\alpha_iK(\boldsymbol{x}_s,\boldsymbol{x}_i)+\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_jK(\boldsymbol{x}_i,\boldsymbol{x}_j)\tag{7}$$

依据下式可判断 z 是否为正类样本：

$$\left.\begin{aligned}&f(z)\leqslant R^2\qquad z\text{ 为正类}\\&f(z)>R^2\qquad z\text{ 为负类}\end{aligned}\right\}\tag{8}$$

2　基于最大树聚类的多超球体一类分类器

单超球体一类分类器在进行建模时，没有考虑到样本间的分布结构以及同类样本之间存在的差异，因此本文采用多超球体来代替单超球体覆盖训练样本。图 1 中，“o”表示的是正常样本，“*”表示的是异常样本。图 1a 所示是采用单超球体覆盖训练样本的示意图，图 1b 所示是采用多超球体覆盖训练样本的示意图。通过比较发现图 1b 的方法较之于图 1a 的方法具有更高的识别率。



(a) 单超球体

(b) 多超球体

图 1　单超球体与多超球体比较图

相对于单超球体的一类分类器，多超球体一类分类器首先要对训练样本进行聚类，然后对聚类后的各子类分别进行一类支持向量机分类器学习，最后得到对应的多个超球体一类分类模型。

2.1　聚类

本文采用最大树[9-10] 聚类算法进行聚类。用绝对值减数法：

$$r_{jk}=1-\alpha\sum_{l=1}^{n}|x_i^{(j)}-x_i^{(k)}|\tag{9}$$

建立模糊相似矩阵 $\boldsymbol{R}=(r_{jk})_{N\times N}$($r_{jk}$ 为任意两个样本 $x^{(j)}$ 与 $x^{(k)}$ 之间的相似系数)。在求得最大树后为了选择合适的参数 λ 对训练样本进行聚类划分，根据文献[11] 中提到的关于误差平方和 J_e 与聚类后子类数 c 的关系(给定的 n 个样本若能聚类成 $\hat{c}$ 个子类，那么在聚类过程中，J_e 会随着 c 的增大而迅速减小，直到 $c=\hat{c}$($\hat{c}$ 为拐点处的子类数)，然后 J_e 减小速度变缓，直到 $c=n$ 为止，J_e 不再减小)。依据这一关系，本文取聚为 $\hat{c}$ 类时所对应的参数 λ，根据该参数得到样本集所分成的 $\hat{c}$ 个子类。

2.2　基于最大树聚类的多超球体一类分类法流程

基于最大树聚类的多超球体一类分类法具体

的过程如下：

(1)对所得样本数据进行特征提取，得到对应的训练样本集、测试样本集。

(2)将训练样本集按最大树聚类算法聚为多个子类。根据所聚成的子类个数将训练样本集的各个子类分别进行一类支持向量机分类器学习，得到各个子类所对应的单超球体，各子类对应的单超球体相组合就构成对应于训练样本集的多超球体一类分类模型。

(3)采用得到的多超球体一类分类模型对测试样本集进行决策。只要存在一个超球体能包含测试样本，就将该测试样本视为正常类；若没有一个超球体能包含该测试样本，就将该测试样本视为异常类。

图 2 是其对应的流程图，可以看出当训练样本集聚类为一个子类(即 $m=1$)时，所得的多超球体分类模型就是单超球体分类模型。即单超球体一类分类器可以看作是多超球体一类分类器将其对应的训练样本集聚为一个子类的特例。

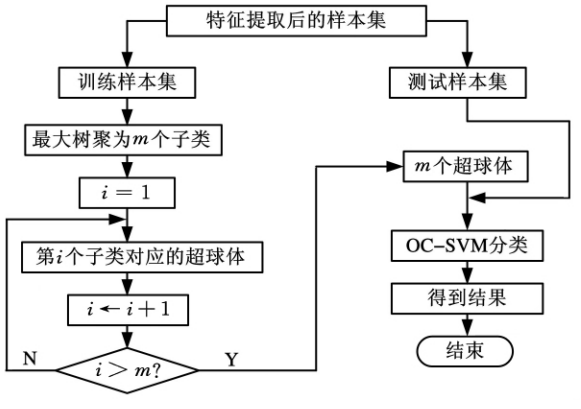


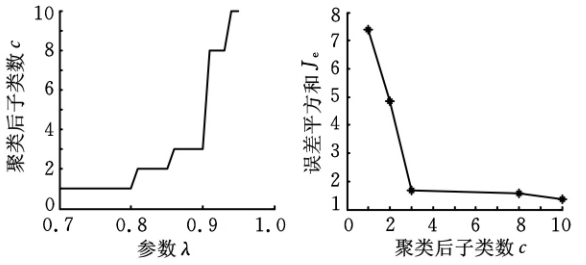
图 2 算法流程

3 实验与应用

3.1 仿真数据实验

为了验证基于最大树聚类的多超球体一类分类法的可行性，本文首先选用图 1 所示的具有聚类特性的仿真数据进行验证。从图 1 可以看出，正常样本聚类特征明显，倾向于聚为 3 个子类。

随机选择正常样本的 2/3 作为训练样本，剩余的 1/3 样本作为正类测试样本，所有的异常样本作为负类测试样本。采用最大树聚类法(α 取 0.5)，选取聚类子类数 10 以内对应的结果，如图 3 所示。根据图 3a 所示的参数 λ 与聚类后子类数 c 的关系，图 3b 所示的误差平方和 J_e 与聚类后子类数 c 的关系，选取参数 $\lambda=0.88$，聚类后聚为 3 个子类，这一点与图 1 中样本簇分布的趋势一致。



(a)参数 λ 与子类数关系　(b)子类数与误差平方和关系

图 3 仿真数据聚类结果

根据聚类的结果，分别采用单超球体一类分类器和多超球体一类分类器进行学习，两种算法中涉及的惩罚因子 C 与高斯核参数 σ，均采用文献[12-14]中提到的粒子群优化算法对其进行参数自适应优化。两种算法在最优参数下得到的识别率如表 1 所示。其中，T 为正类训练样本数；T_1 为正类测试样本数；T_2 为负类测试样本数；N 为支持向量个数；R_1 为正类识别率；R_2 为负类识别率；R 为平均识别率，$R=(R_1+R_2)/2$。

表 1 多超球体一类分类器与单超球体一类分类器对仿真数据的实验结果

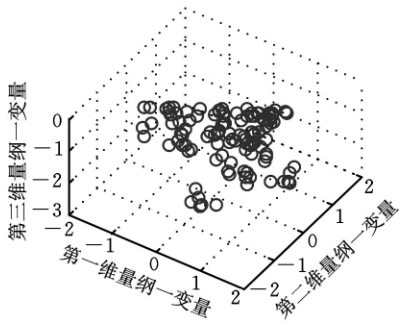
算法	T	T_1	T_2	N	R_1(%)	R_2(%)	R(%)
单超球体一类分类	39	20	37	6	70.00	54.05	62.03
多超球体一类分类				10	85.00	100	92.50

表 1 所示结果表明，当训练样本呈聚类特征分布时，多超球体一类分类算法相对于单超球体一类分类算法具有优越性。

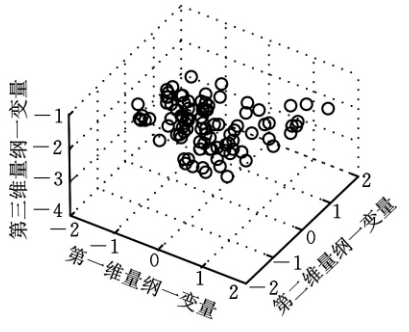
3.2 UCI 标准数据集实验

为了进一步验证该算法的可行性。本文选取 UCI 数据库中的 Sonar 这个两类数据集产生两个单类数据来验证。获取的 Sonar 数据集包含两类，分别记为 Sonar1、Sonar2。首先对获取的数据在信息量保持 0.95 的情况下，得到主成分分析(principle component analysis，PCA)特征压缩后的两类样本数据。图 4a 与图 4b 分别是部分 Sonar1 和 Sonar2 数据取最大 3 维主分量的可视化分布图，从一定程度上反映了高维数据簇分布的趋势。

和仿真实验中一样，分别针对每一类样本集，随机选取其中的 2/3 样本作为正类训练样本，剩余的 1/3 同类样本作为正类测试样本，另一类的 1/3 样本作为负类测试样本。图 5、图 6 分别是对 Sonar1 及 Sonar2 采用最大树聚类法(α 取 0.2)，对应聚类子类数 10 以内的结果。因此 Sonar1、Sonar2 分别取：参数 $\lambda=0.87$、聚类后聚为 3 个子类以及参数 $\lambda=0.88$、聚类后聚为 3 个子类。

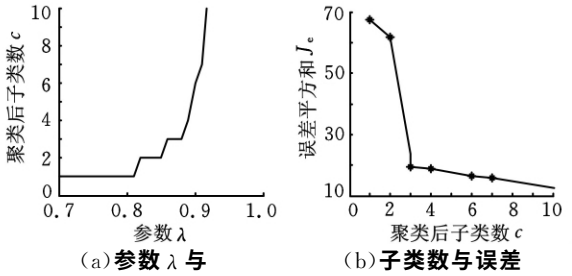


(a)Sonar1 样本三维分布图



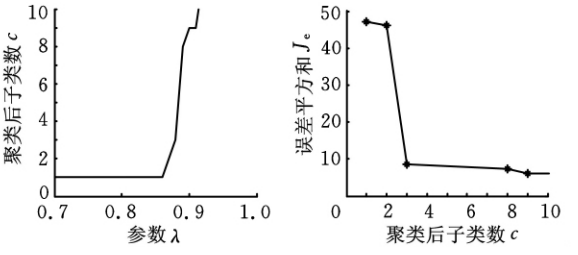
(b)Sonar2 样本三维分布图

图 4　Sonar 数据集的可视化分布图



(a)参数 λ 与子类数关系　(b)子类数与误差平方和关系

图 5　Sonar1 聚类的结果



(a)参数 λ 与子类数关系　(b)子类数与误差平方和关系

图 6　Sonar2 聚类的结果

同时采用粒子群优化算法对多超球体一类分类器与单超球体一类分类器两种算法中涉及的惩罚因子 C 与高斯核参数 σ 进行参数自适应优化，两种算法在最优参数下得到的识别率如表 2 所示。

从表 2 可以看出，Sonar1 中多超球体的平均识别率要比单超球体的平均识别率提高了近 15%，而 Sonar2 中两种方法的平均识别率比较接近，这是由于 Sonar2 的数据聚类特征不是很明显，这与图 4b 的三维可视图的分布是相符的。对标准数据多超球体一类分类器首先考虑了数据内部的簇分布情况，其对正负类样本的平均识别率总体上与单超球体一类分类器相比均有所提高，可见该算法的有效性。

3.3　转子故障诊断

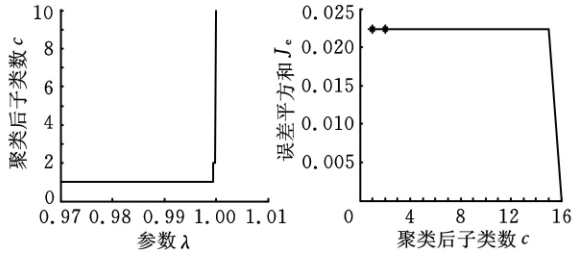
借助 ZT－3 多功能转子实验台以及 DH5922 动态信号测试分析系统，在不同的转速下采集了不平衡、不对中、碰摩以及油膜涡动 4 类转子故障样本：不平衡 25 个，不对中 22 个，碰摩 29 个，油膜涡动 31 个。

表 2　多超球体一类分类器与单超球体一类分类器对标准数据集的实验结果

数据集	T	T_1	T_2	单超球体一类分类				多超球体一类分类			
				N	R_1(%)	R_2(%)	R(%)	N	R_1(%)	R_2(%)	R(%)
Sonar1	74	37	33	9	72.97	48.48	60.73	13	83.78	66.67	75.23
Sonar2	64	33	37	5	87.88	72.97	80.43	9	90.91	72.97	81.94

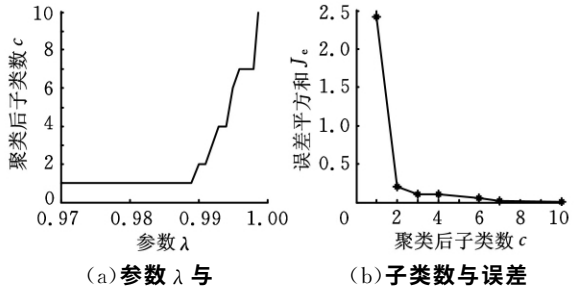
本文将实验提取的 4 类转子故障的样本数据进行频谱分析，得到信号频谱后，对频谱进行归一化处理，然后直接对频谱数据在信息量保持率为 95%的情况下进行 PCA 特征压缩。分别对压缩后的 4 类特征样本集建立其对应的多超球体一类分类器：不平衡对应所有类别、不对中对应所有类别、碰摩对应所有类别、油膜涡动对应所有类别。建立每个模型时，分别对每一类故障数据随机选取其中 2/3 的样本数据作为正类训练样本集，将剩余的 1/3 样本作为正类测试样本集，将其他各故障的 1/3 样本组合成负类测试样本集，依照本文提出的基于最大树聚类的多超球体一类分类器进行学习。

图 7～图 10 所示为采用最大树聚类法（α 均取 0.6），分别对四种转子故障的训练样本集进行聚类的结果。图 7b 中 10 个子类以内聚为 n 个子类与 $n+1$ 个子类间的误差平方和的差距很小（小于 0.001），因此对于不平衡样本，聚类后的子类个数仍为 1。因此根据图 7～图 10 所示聚类后参数 λ 与子类数 c 的关系、子类数 c 与误差平方和 J_e 的关系，分别对每一类故障选择的参数 λ 以及所得的子类数是：不平衡——0.98，1；不对中——0.99，2；碰摩——0.99，2；油膜涡动——0.98，2。



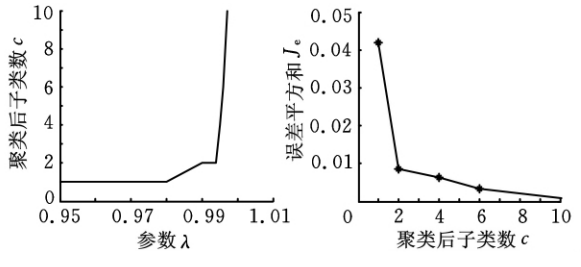
(a)参数 λ 与子类数关系　(b)子类数与误差平方和关系

图 7　不平衡样本聚类结果



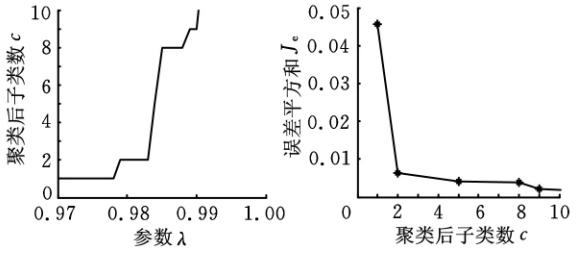
(a)参数 λ 与子类数关系　(b)子类数与误差平方和关系

图 8　不对中样本聚类结果



(a)参数 λ 与子类数关系　(b)子类数与误差平方和关系

图 9　碰摩样本聚类结果



(a)参数 λ 与子类数关系　(b)子类数与误差平方和关系

图 10　油膜涡动样本聚类结果

根据聚类后的结果采用本文提到的多超球体一类分类器建立模型，同时与常用的单超球体一类分类器比较了实验结果。同样对两种算法均以粒子群优化算法优化各算法中所涉及的惩罚因子 C 与高斯核参数 σ。在最优参数下所得到的识别率如表 3 所示。从表 3 的实验结果可知，由于不平衡样本经最大树聚类后仍聚为一个子类，故对于不平衡样本的单超球体一类分类算法即可看成是其多超球体一类分类算法的特例，两者结果一样。其他三类故障样本经聚类后均聚为两个子类：不对中样本采用多超球体一类分类算法不仅支持向量个数比单超球体一类分类法少了，且其对应的识别率也提高了；碰摩样本使用多超球体一类分类法后在支持向量个数增加的情况下，识别率有了提高；油膜涡动样本对应的多超球体一类分类法虽然支持向量个数增加了，但是最后的识别率同样达到了 100%。由此可见，该算法相对于常用的单超球体一类分类法在识别率上表现了其有效性。

表 3　多超球体一类分类器与单超球体一类分类器对转子故障的识别率

故障类型	T	T_1	T_2	单超球体一类分类				多超球体一类分类			
				N	R_1(%)	R_2(%)	R(%)	N	R_1(%)	R_2(%)	R(%)
不平衡对应所有	16	9	28	2	88.90	92.86	90.88	2	88.90	92.86	90.88
不对中对应所有	14	8	29	6	87.50	72.41	79.96	4	87.50	75.86	81.68
碰摩对应所有	19	10	27	3	90.00	88.89	89.45	5	90.00	92.59	91.30
油膜涡动对应所有	21	10	27	3	100	100	100	5	100	100	100

4　结语

本文从考虑数据内在分布的角度出发研究了一种基于最大树聚类的多超球体一类分类算法。首先对经 PCA 特征降维后的训练样本集采用最大树聚类算法实现聚类，得到对应的内在分布簇形成的各子类；然后对各簇子类分别进行一类支持向量机分类器训练，并且利用粒子群优化算法获取最优参数，得到各子类对应的超球体；最后建立由各子类对应的超球体而形成的多超球体一类分类模型。分别将该方法应用于仿真数据、UCI 标准数据集以及转子故障数据这三个实例中，实验结果表明，当样本数据呈簇类分布时，尤其是聚类特征比较明显时，该方法相对于常用的单超球体一类分类方法具有可行性及有效性。

参考文献：

[1] Juszczak P. Learning to Recognise: a Study on One－class Classification and Active Learning [D]. Delft: Delft University of Technology, 2006.

[2] Camci F, Chinnam R B. General Support Vector Representation Machine for One－class Classification of Non－stationary Classes[J]. Pattern Recognition, 2008, 41: 3021-3034.

[3] Tsang I W, James T K, Li S. Learning the Kernel in Mahalanobis One－class Support Vector Machines [C]//Proceeding of the International Joint Conference on Neural Networks. Vancouver, Canada, 2006: 1169-1175.

切入磨削与纵向磨削的磨削力分析与比较

李　厦　李郝林
上海理工大学，上海，200093

摘要：研究了同时包含切入磨削和纵向磨削的复杂外圆磨削过程。根据纵向磨削过程的特点，将砂轮等效成若干个小砂轮，在传统阶梯模型的基础上构建了砂轮磨损的抛物线模型。推导了基于两种模型的纵向磨削切向分力和切入磨削切向分力的比较公式，两切向分力的比值反映了切入磨削和纵向磨削转换时切向分力的变化情况，它主要与磨削系数、砂轮宽度和纵向进给速度有关。采用砂轮主轴功率信号分析磨削切向分力，通过实验验证了抛物线模型更符合实际情况的结论。研究结果为采用磨削力信号和功率信号研究复杂磨削过程的监控提供了参考依据。

关键词：切入磨削；纵向磨削；磨削力；功率信号
中图分类号：TG580.6　　DOI：10.3969/j.issn.1004－132X.2012.03.004

Analysis and Comparison of Grinding Forces between Plunge Grinding and Traverse Grinding

Li Sha　Li Haolin
University of Shanghai for Science and Technology, Shanghai, 200093

Abstract: Complex cylindrical grinding processes including both plunge grinding and traverse grinding were investigated. According to the traverse grinding characteristics, the grinding wheel was equivalent to a number of small wheels, a parabolic model for wheel wear was built based on the traditional steps model. The comparision formula of tangential grinding force was derived based on two tangential grinding forces of plunge grinding and traverse grinding the ratio of the tangential grinding force reflected tangential grinding varieties from the plunge grinding to the traverse grinding and was concerned with grinding coefficient, the wheel width and the traverse feed rate. The tangential grinding force was analyzed by the grinding wheel spindle power signals in experiments and the parabolic model is more realistic. Using force signals and power signals to monitor the complex grinding processes provides a reference method.

Key words: plunge grinding; traverse grinding; grinding force; power signal

0　引言

磨削加工技术是先进制造技术中的重要内容，磨削加工的品质往往决定着工件的最终加工精度。外圆磨削是一种主要的磨削方式，它包括切入磨削(plunge grinding)和纵向磨削(traverse grinding)。切入磨削时，砂轮与工件之间只有径向运动；纵向磨削时，砂轮与工件之间既有径向运

收稿日期：2011—06—14
基金项目：上海市科学技术委员会科研基金资助项目(10DZ2292100)；上海市教委重点学科建设资助项目(J50503)

[4] Tax D. One－class Classification: Concept－learning in the Absence of Counter－examples[D]. Delft: Delft University of Technology, 2001.
[5] 谭真臻，陈果，孙丽萍. 基于Hilbert谱图特征的航空发动机转子故障智能诊断[J]. 机械科学与技术，2010，29(9)：1177-1181.
[6] 冯爱民，陈松灿. 基于核的单类分类器研究[J]. 南京师范大学学报(工程技术版)，2008，8(4)：1-6.
[7] 戴蒙，林家骏，刘云翔. 基于FCM聚类的多超球体一类分类数字图像隐藏信息[J]. 中国图像图形学报，2008，13(10)：1918-1921.
[8] Wang D, Yeung D S, Tsang E C C. Structured One－class Classification[J]. IEEE Trans. on Systems, Man, and Cybernetics－Part B: Cybernetics, 2006, 36(6): 1283-1294.
[9] 肖健华. 智能模式识别方法[M]. 广州：华南理工大学出版社，2006.
[10] 杨梦宁，杨丹，张强劲. 基于最大树法的模糊图像分割方法[J]. 计算机科学，2005，32(8)：190-191.
[11] Duda R O, Hart P E, Stork D G. 模式分类[M]. 李宏东，姚天翔，等，译. 2版. 北京：机械工业出版社，2003.
[12] Chapelle O, Vapnik V, Bousquet O, et al. Choosing Multiple Parameters for Support Vector Machines[J]. Machine Learning, 2002, 46(1): 131-159.
[13] 王东，吴湘滨. 利用粒子群算法优化SVM分类器的超参数[J]. 计算机应用，2008，28(1)：134-135.
[14] 邵信光，杨慧中，陈刚. 基于粒子群优化算法的支持向量机参数选择及其应用[J]. 控制理论与应用，2006，23(5)：740-743.

(编辑　王艳丽)

作者简介：刘丽娟，女，1986年生。南京航空航天大学民航学院硕士研究生。主要研究方向为图像处理与模式识别、智能故障诊断。陈　果，男，1972年生。南京航空航天大学民航学院教授、博士研究生导师。