文章编号: 1001-0920 (2011) 08-1229-04

基于可信间隔的特征选择方法研究

姜慧研[a], 柴天佑[b,c]

(东北大学 a. 软件学院，b. 流程工业综合自动化教育部重点实验室，c. 自动化研究中心，沈阳 110819)

摘 要: 传统的特征选择方法没有很好地考虑数据的模式特性而导致性能下降. ReliefF 是较为有效的特征选择方法, 但存在特征权值随样本波动和不能去除冗余特征的问题. 对此, 从数据本身的模式特性出发, 提出了可信间隔的概念和基于可信间隔进行特征选择的方法. 以氧化铝回转窑烧结过程数据为实验数据进行特征选择和烧结工况识别实验, 结果表明, 所提出的方法能去除冗余特征, 有效地提高了识别率.

关键词: 可信间隔；特征选择；支持向量机；模式识别

中图分类号: TP391.4 **文献标识码:** A

Method for feature selection based on authentic distance

JIANG Hui-yan[a], *CHAI Tian-you*[b,c]

(a. School of Software，b. Key Laboratory of Integrated Automation of Process Industry of Ministry of Education, c. Automation Research Center，Northeastern University，Shenyang 110819，China. Correspondent：JIANG Hui-yan，E-mail：hyjiang@mail.neu.edu.cn)

Abstract: Traditional feature selection methods don't consider the pattern feature of data, which leads to the performance degradation. ReliefF is a more effective method for feature selection, but the weight of feature can fluctuate with samples and the method cannot remove redundant feature. Therefore, the concept of authentic distance and a new method used to feature selection are proposed based on the pattern feature of data, and the feature selection and experiments of sintering working conditions recognition are operated with the data of alumina rotary kiln sintering process as experimental data. The experimental results show that the proposed method can effectively remove unrelated features and greatly improve recognition rate.

Key words: authentic distance；feature selection；support vector machine；pattern recognition

1 引 言

特征提取和选择是模式识别领域的重要研究课题之一. 为了提高准确率, 人们往往最大限度地提取特征信息, 使特征向量能够充分包含类别信息. 但是特征所包含的信息是否充分很难确定, 结果不仅增大了特征向量的维数, 而且由于每个特征对样本的不确定性度量的贡献不同, 其中一些与问题本身无关的干扰特征或冗余特征将导致计算成本的增加和对未知数据泛化能力的降低. 因此, 选择合适的特征来描述模式, 对提高识别性能和降低计算成本具有重要意义.

特征选择可以看作是一个求解组合优化问题的过程. 即对给定的特征集合, 通过建立一种评价标准获得有助于分类的最优特征子集[1], 使得由它产生的类分布可以无限地接近训练类分布[2]. 因此, 在降低计算成本和保证识别率的前提下, 如何能够找到这样的特征子集便成为一个重要课题.

根据评价标准与分类器的关系, 特征选择基本包括嵌入式 (Embedded), 过滤式 (Filter) 和封装式 (Wrapper) 3 种模型[3-4]. 其中: Embedded 模型利用分类器的某些特性而非直接使用分类器进行特征子集的评判; Wrapper 模型利用分类器的反馈进行特征选择, 选择的过程中需要不断从后端分类器获得反馈, 即将分类器的识别率作为评价标准来构造识别性能最优的特征子集; Filter 模型利用类间可分性进行特征选择, 在特征选择过程中不考虑后端分类器的特征, 即将距离等可分性指标作为评价标准对特征进行排序, 然后选择位于前列的特征构成特征子集. 由于 Embedded 模型将分类器的训练和特征选择融合到一

收稿日期: 2010-05-24；**修回日期:** 2011-02-16.

基金项目: 国家863计划重点项目(2007AA041404)；国家自然科学基金重点项目(50834009).

作者简介: 姜慧研(1963–), 女, 教授, 博士, 从事图像处理与分析、计算机辅助诊断等研究；柴天佑(1947–), 男, 教授, 博士生导师, 中国工程院院士, 从事自适应控制、多变量智能解耦控制等研究.

起, 其性能受到分类器的制约; Wrapper 模型的性能较好, 但当特征数目较多时, 学习算法中的迭代次数很高, 直接应用存在一定的困难; Filter 模型的优点是简单、速度快, 在去除无关特征方面的能力较强, 但去除冗余特征的能力较弱. 当这 3 种模型的假设与数据不吻合时, 性能会下降.

由于上述特征选择方法缺乏对数据本身模式特性的分析而导致性能下降, 本文提出一种基于数据本身特性的可信间隔的概念, 并研究了将可信间隔作为类间可分性测度进行特征选择的方法.

2 Relief 和 ReliefF 特征选择算法

Filter 模型去除无关特征的能力较强, Relief 算法[5]是 Filter 模型的代表性算法. 该算法借鉴了最近邻算法的思想, 利用 "假设间隔 (RF)" 计算特征权值, 并基于权值进行特征选择. 即通过调整每个特征的权值凸显特征与类别的相关程度, 在保持样本分类不变的情况下最大化决策面能够移动的距离[6].

假设间隔可由下式计算:

$$r=\frac{1}{2}(\|x-x_m\|-\|x-x_h\|). \tag{1}$$

其中: x 为训练样本, x_h 和 x_m 分别为与 x 同类或不同类的最近邻样本, $\|x-x_h\|$ 为同类样本间的差异, $\|x-x_m\|$ 为非同类样本间的差异.

Relief 算法适于两类训练样本的情况, ReliefF 算法[7]利用 k 近邻思想解决多类识别问题. 假设特征量均为数值型的, ReliefF 算法如下:

1) 权值 $w_j:=0$.

2) for $i:=1$ to n (n 为样本个数).

2.1) 随机选取样本 x_i;

2.2) for $l:=1$ to k

$h_l:=$ 第 l 个与 x_i 同类的最近邻样本,

$m_l:=$ 第 l 个与 x_i 异类的最近邻样本;

endfor

2.3) for $j:=1$ to N

$\mathrm{d}h_j:=h_l$ 与 x_i 的第 j 个特征的差异,

$\mathrm{d}m_j:=m_l$ 与 x_i 的第 j 个特征的差异,

$$w_j:=w_j-\frac{\mathrm{d}h_j}{nk}+\frac{\mathrm{d}m_j}{nk}. \tag{2}$$

endfor

endfor

3) 根据权值大小对特征进行排序和选择.

3 基于可信间隔的特征选择算法

Relief 和 ReliefF 算法基于假设间隔更新特征权值, ReliefF 算法对于多类识别较为有效. 但当特征干扰严重和特征值较小时, 特征权值的差异也较小, 权重效应被弱化. 为了凸显特征间的差异性, 本文提出了 "相似间隔 (SD)" 的概念, 并利用相似间隔凸显特征的差异.

相似间隔的计算公式为

$$r_{ij}^k=\exp(-|g_i^k-g_j^k|). \tag{3}$$

其中: r_{ij}^k 为第 i 个样本和第 j 个样本在第 k 个特征上的相似间隔, g_i^k 为第 i 个样本的第 k 个特征.

假设间隔和相似间隔均没有考虑分类或识别系统性能, 为了兼顾对错误分类的惩罚和凸显特征间的差异, 本文进一步提出了包含模式特性的 "可信间隔 (AD)" 的概念, 以及将可信间隔作为类间可分性测度的特征选择新方法, 将 "分类准确率最优" 作为评估准则来决定特征子集的规模以获取最优特征子集.

可信间隔的计算公式为

$$r_{ij}^k=A\cdot\rho^k\exp(-|g_i^k-g_j^k|). \tag{4}$$

其中: r_{ij}^k 为第 i 个样本和第 j 个样本在第 k 个特征上的可信间隔; g_i^k 和 g_j^k 分别为样本集中第 i 个样本和第 j 个样本的第 k 个特征; ρ^k 为对所有 m 个样本仅用第 k 个特征识别时的识别率, 反映模式的特性; m 为样本数; A 为惩罚因子, A 越大, 对错误分类的样本惩罚越大.

基于可信间隔的特征选择算法如下:

1) for $i:=1$ to N (N 为特征向量维数).

2) for $j:=1$ to n (n 为样本个数).

2.1) 选取样本 x_j;

2.2) 建立与样本 x_j 同类的样本集合 C_1 和异类的样本集合 C_2;

2.3) 基于式 (3) 计算 C_1 内所有样本与 x_j 的可信间隔, 并进行累加, 累加后的可信间隔设为 $r_{ij}[C_1]$; 同理, 计算 C_2 内所有样本与 x_j 的可信间隔的累加和 $r_{ij}[C_2]$.

3) 基于下式计算第 j 个样本的第 i 个特征 g_i^j 权值:

$$w_i^j:=|r_{ij}[C_1]-r_{ij}[C_2]|. \tag{5}$$

4) 基于下式计算第 i 个特征的权值:

$$w_i:=\frac{1}{n}\sum_{j=1}^{n}w_i^j. \tag{6}$$

5) 按权值从大到小的顺序对特征进行排序, 组成一个相关性从强到弱的特征序列.

6) 基于支持向量机[8]进行模式识别, 将 "分类准确率最优" 作为特征子集的评估准则, 据此决定删除多少排序靠后的特征数. 即在满足一定识别率的基础上使计算成本最小.

4　实验结果与分析

4.1　实验对象与数据描述

回转窑内部烧结工况是决定氧化铝质量的重要因素, 其准确识别对提高回转窑控制系统的安全性、可靠性和产品质量具有重要意义. 由于回转窑长达百米且处于不断旋转和高温煅烧中, 其结构的特殊性和烧结法工艺的复杂性, 使得烧结过程存在多变量强耦合、强非线性、大惯性和不确定性干扰等综合复杂特性. 迄今为止, 烧结工况的识别基本依靠人工看火方法, 烧结工况被分为“正常”、“过烧”和“欠烧”3 种模式. 人工看火的结果受看火工素质和经验的制约, 缺乏客观性和科学性.

本文以氧化铝回转窑为实验对象研究窑内烧结工况的识别问题. 实验数据是通过回转窑图像采集系统获取烧结视频, 然后利用图像采集卡量化成彩色数字图像. 本文从中随机选取 318 个图像样本进行实验, 其中 138 个样本用于训练, 180 个样本用于测试. 在训练样本中, 欠烧、正常和过烧工况样本数分别为 46, 47 和 45. 在测试样本中, 欠烧、正常和过烧工况样本数分别为 60, 59 和 61.

4.2　特征提取实验

根据现场看火工的经验, 物料区、火焰区、黑把子区和充分燃烧区包含重要的烧结信息, 因此本文将这些区域定义为关心区域(ROI), 如图 1 所示. 其中: 充分燃烧区是煤粉从喷煤管喷出后与回转窑内高温空气混合在一起, 瞬间爆炸燃烧形成的高亮区域; 火焰区是由煤粉燃烧照亮区域; 物料区是由生料浆烧结形成的区域; 黑把子区是由未燃煤粉形成的黑色区域.

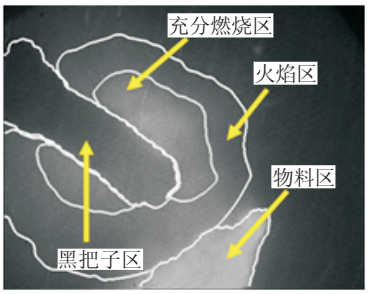


图 1　关心区域

本文利用文献 [9] 的方法分割 ROI 区域, 具体步骤如下:

1) 在图像 $I(x,y)$ 中选择物料区的敏感区域(物料区下部的矩形区域), 在物料区内、外分别设置初始轮廓曲线作为初始零水平集, 基于快速行进法演化内、外初始零水平集. 当内、外零水平集的能量均达到最小时, 迭代结束. 将迭代结束时的内、外两条零水平集曲线基于“逻辑与”方式融合, 提取物料区域 M; 同理, 提取黑把子区域 B.

2) 选择物料区边缘靠近火焰区部分的像素作为能量源集合 Γ.

3) 按下式计算图像受到能量源作用后的能量衰减值 E 和除去物料对火焰耦合作用后的图像 I':

$$E(x,y,x_i,y_i)=\frac{I(x_i,y_i)+b}{1+\left(\dfrac{a}{100}\right)^2[(x-x_i)^2+(y-y_i)^2]},\tag{7}$$

$$I'(x,y)=I(x,y)-\frac{1}{n}\sum_{i=1}^{n}E(x,y,x_i,y_i).\tag{8}$$

其中: $I(x,y)$ 为原图像, Γ 为能量源, n 为 Γ 内像素的个数, $a=0.28$, $b=15$, $(x_i,y_i)\in\Gamma$, $i=1,\cdots,n$.

4) 利用改进的大津方法从图像 $I'(x,y)$ 中分割出火焰区候补区域. 即首先利用下式计算阈值 t^*:

$$t^*=\arg\max\{\omega_0(t)(\mu_0(t)-\mu)^2+\omega_1(t)(\mu_1(tt)-\mu)^2\};\tag{9}$$

然后基于 t^* 对图像 $I'(x,y)$ 进行二值化处理; 最后利用火焰区候补区域与黑把子区的差分运算分割火焰区.

5) 设像素总数为火焰区的像素数, 利用大津方法分割充分燃烧区.

对烧结工况的“正常”、“过烧”和“欠烧”3 种模式, 均从 ROI 中提取特征向量 $\kappa=[g_1,g_2,\ldots,g_{44}]$, 组成烧结工况特征集合 S. 其中: $g_1,\cdots,g_{30}$ 是从黑把子区、充分燃烧区和火焰区提取的平均灰度、方差、面积、高度、纵横比、重心坐标以及轴向长等特征; $g_{31},\cdots,g_{43}$ 是从物料区提取的能量、熵、惯性矩以及局部平稳等特征; g_{44} 是图像整体方差特征.

4.3　实验结果与性能比较

在式 (4) 中, A 是对错分样本惩罚性能的惩罚因子. 本文基于试凑方法进行 A 值选择实验, 所得结果如图 2 所示. 其中: set1 是包含所有 44 个特征的集合, set2 是包含基于权值排序的前 30 个特征集合. 从图 2 中可以看出, $A=78$ 时 set1 和 set2 的识别率均达到最大, 因此取 A 值为 78.

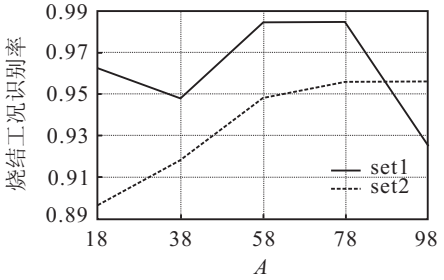


图 2　基于不同 A 值的烧结工况分类结果

分别基于 ReliefF 算法、相似间隔 (SD) 和可信间

隔(AD)进行特征权值的计算, 所得结果如图3所示.

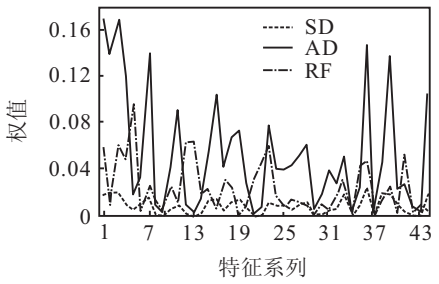
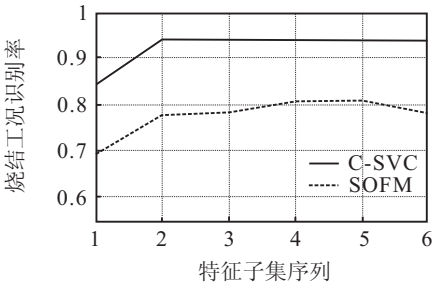


图 3 基于3种方法的特征权值计算结果

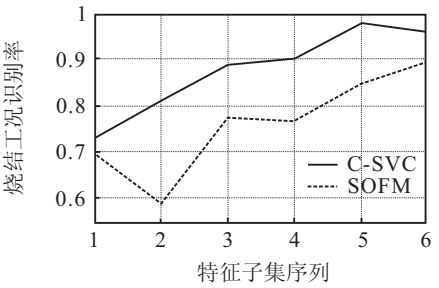
从图3可以看出, 基于ReliefF算法计算的特征权值差异较小; 与SD方法相比, 基于AD方法计算的特征权值更好地凸显了特征间的差异性.

分别基于上述3种特征选择方法和两种不同识别方法, 对6个特征子集进行烧结工况识别实验. 其中, 特征子集1～特征子集6包含的特征数分别为5, 10, 20, 30, 40, 44. 识别方法包括支持向量机中的C-SVC方法和自组织神经网络(SOFM)方法.

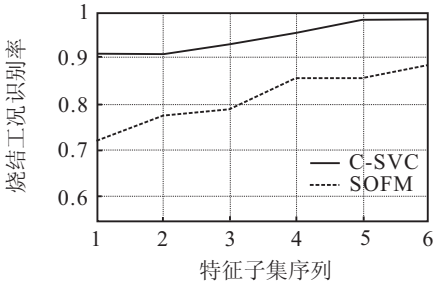
在ReliefF算法, SD方法和AD方法等不同特征选择方法下, 基于C-SVC和SOFM分类器得到的烧结工况识别结果如图4所示.



(a) 基于ReliefF算法



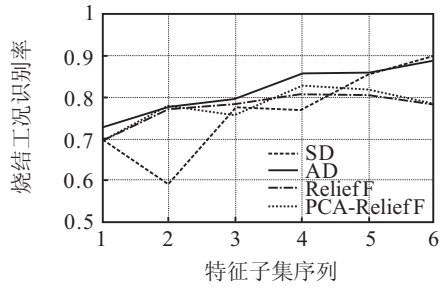
(b) 基于SD方法



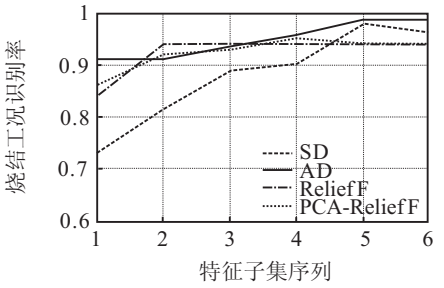
(c) 基于AD方法

图 4 基于不同特征选择方法的烧结工况分类结果

在不同识别方法下, 将SD, AD, ReliefF, PCA-ReliefF方法[10]选择的特征作为输入, 所得到的烧结工况识别结果如图5所示. 其中, PCA-ReliefF方法是指先基于主成分分析(PCA)方法对实验数据进行降维, 再基于ReliefF方法进行特征选择.



(a) 基于SOFM方法



(b) 基于C-SVC方法

图 5 基于不同识别方法的烧结工况识别结果

从以上实验可以看出:

1) 将基于可信间隔(AD)的特征选择方法获得的特征作为输入, 在不同分类器上所得到的识别结果均为最佳;

2) 特征子集5是最有效表达图像特征的集合;

3) 如果确定了特征选择算法, 则支持向量机的分类结果较好.

5 结 论

本文提出了一种基于可信间隔的特征选择方法. 该方法充分考虑了研究对象的模式特性和特征之间的差异, 先利用可信间隔计算特征权值, 然后通过支持向量机评估和选择最优特征子集. 将本文方法应用于氧化铝回转窑烧结工况识别实验, 所得结果表明, 与ReliefF和PCA-ReliefF算法相比, 基于可信间隔的特征选择方法能更有效地选择对工况识别有重要贡献的特征, 取得了较好的识别效果.

参考文献(References)

[1] Almuallim H, Dietterich T G. Learning boolean concepts in the presence of many irrelevant features[J]. Artificial Intelligence, 1994, 69(1/2): 279-305.

[2] Koller D, Sahami M. Toward optimal feature selection[C]. Proc of Int Conf on Machine Learning. Bari: IEEE, 1996: 284-292. (下转第1238页)