第43卷 第3期
2011年3月
哈尔滨工业大学学报
JOURNAL OF HARBIN INSTITUTE OF TECHNOLOGY
Vol. 43 No. 3
Mar. 2011

一种类内方差与相关度结合的特征选择算法

张晓光[1,2], 孙 正[3], 徐桂云[1], 阮殿旭[1]

(1. 中国矿业大学 机电工程学院, 221116 徐州, doctorzxg@163.com; 2. 哈尔滨工业大学 现代焊接生产技术国家重点实验室, 150001 哈尔滨; 3. 枣庄学院 物理与电子工程系, 277160 山东 枣庄)

摘 要: 为了解决现有特征选择算法没有同时考虑特征之间以及特征与类别之间的相关性,且存在计算量大、适用范围窄等问题,从均方误差最小的分类训练准则出发,并借鉴线性鉴别分析的思想,提出了一种类内方差与相关度结合的特征选择算法,并使用核方法将其推广到可以解决非线性分类的特征选择问题. 该算法不仅同时考虑了样本特征之间以及特征与分类标号之间的相关性,而且使得类内方差最小,有效地提高了分类器的性能. 仿真实验表明:该算法适用于对特征数量多、特征相关性强的数据集进行特征选择,其选择的特征子集能够显著提高分类精度,具有较大的优越性.

关键词: 类内方差; 相关度; 特征选择; LDA; 分类

中图分类号: TP391.41 **文献标志码:** A **文章编号:** 0367-6234(2011)03-0132-05

A feature selection algorithm combining within-class variance with correlation measure

ZHANG Xiao-guang[1,2], SUN Zheng[3], XU Gui-yun[1], RUAN Dian-xu[1]

(1. College of Mechanical and Electrical Engineering, China University of Mining and Technology, 221116 Xuzhou, China, doctorzxg@163.com; 2. State Key Lab of Advanced Welding Production Technology, Harbin Institute of Technology, 150001 Harbin, China; 3. Physics & Electronics Engineering Department, Zaozhang University, 277160 Zaozhuang, Shandong, China)

Abstract: To solve the problems of large calculation and narrow application scope of present feature selection algorithms which do not take the correlation among sample features and that between sample features and classification labels into account, a feature selection algorithm combining within-class variance with correlation measure was proposed based on the principle of the minimum mean square error for classifier and the idea of linear discrimination analysis, then the algorithm was expanded to solve nonlinear feature selection problems by using kernel technology. The proposed algorithm can not only consider both the correlation among the features and that between the features and classification labels, but also minimize the within-class variance, effectively improving the generalization performance of classifier. The experimental results show that the proposed algorithm can be used to select important features from these data sets with numbers of correlation characteristics, and effectively improves the classification accuracy.

Key words: within-class variance; correlation measure; feature selection; LDA; classification

通常,维数约简方法主要有特征提取和特征选择两类. 其中,特征提取是指通过线性或非线性方式把原有的高维数据特征变换为低维数据特征,如 PCA、FDA、KPCA 等[1-4];特征选择是指从大量的原始特征中选择出少量有用的特征. 与特征提取方法相比,特征选择方法能够较好地保留原有特征的物理意义,便于对数据的理解,具有一定的优越性. 但是,目前大部分特征选择算法没有同时考虑特征之间以及特征与类别之间的相关

收稿日期: 2009-01-26.

基金项目: 现代焊接生产技术国家重点实验室开放课题研究基金资助;江苏省高技术研究资助项目(BG2007 013).

作者简介: 张晓光(1963—),男,教授,博士生导师.

性，且存在计算量大、适用范围窄等问题[4-6].

本文从均方误差最小训练准则出发，推导出基于相关度的特征选择准则. 该准则能够使所选特征之间的相关性最小，同时使特征与样本类标号之间的相关性最大，适用于线性可分数据集的特征选择问题. 考虑到该准则不能反映样本分布，进而无法保证所选特征是最优特征的问题，本文把线性鉴别分析（linear discriminant analysis；LDA）可以获取与样本分布一致的分类超平面的特性引入上述特征选择准则，提出了一种类内方差与相关度结合的特征选择算法. 分析了算法的物理意义，并从理论上对其优越性进行了说明.

1　LDA 的算法分析

LDA 是重要的特征提取方法，广泛用于聚类分析和模式分类中[6]. 该方法基于类内散度最小、类间散度最大准则，找出一个能够反映数据分布的最优鉴别向量，有利于数据模式的分类和分类器泛化能力的提高[1,6].

对 m 个类的训练数据样本 $\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_m\}$，其中 $\boldsymbol{X}_i=\{(\boldsymbol{x}_{ij},y_i)\}_{j=1}^{n_i}$ 表示类 c_i 的样本集合，$\boldsymbol{x}_{ij}\in\mathbf{R}^{d\times 1}$ 是 d 维的列向量，y_i 是类标号，$i=1,2,\cdots,m$，及 $j=1,2,\cdots,n_i$. 设 n_i 为类 c_i 的样本数量，样本总体数量为 n，那么 $n=\sum_{i=1}^{m}n_i$. 令 $\bar{\boldsymbol{u}}_i$ 为类 c_i 的样本均值，$\bar{\boldsymbol{u}}$ 为总体样本的均值，即

$$\bar{\boldsymbol{u}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij},\quad i=1,2,\cdots,m,$$

$$\bar{\boldsymbol{u}}=\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij},$$

则类内散度和类间散度矩阵可以定义为

$$\boldsymbol{S}_{\mathrm{w}}=\frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{u}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{u}}_i)^{\mathrm{T}},$$

$$\boldsymbol{S}_{\mathrm{b}}=\frac{1}{m}\sum_{i}^{m}p_r(c_i)(\bar{\boldsymbol{u}}-\bar{\boldsymbol{u}}_i)(\bar{\boldsymbol{u}}-\bar{\boldsymbol{u}}_i)^{\mathrm{T}}.$$

其中，$p_r(c_i)$ 是类 c_i 的先验概率，既可以把它们设定为 $p_r(c_i)=n_i/n$，也可以按照统计的方法确定类 c_i 在样本集合中出现的概率作为先验概率 $p_r(c_i)$. 根据 LDA 思想，其特征提取准则为

$$J_S(\boldsymbol{w})=\max_{\boldsymbol{w}}\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}}.$$

其中：$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{b}}\boldsymbol{w}$ 为类间散度，表示类间的距离；$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{w}$ 为类内散度，表示同类样本的方差，其值越小，表示同类样本聚集程度越大；$\boldsymbol{w}$ 为鉴别向量，也称为特征变换向量.

LDA 求取的鉴别向量能够有效地与样本分布相一致，具有较好的泛化性能，但是鉴别向量存在不能划分线性可分样本的问题，影响了分类器的性能[6-8]. 基于经验风险的均方误差最小化准则（minimum square error；MSE）训练得到的分类器能够有效区分线性可分的分类样本，但是不能保证分类超平面方向与样本数据分布一致，因而其存在训练精度很高而泛化性能不高的问题.

从上述分析可以看出，如果把 LDA 与 MSE 准则相结合，那么选择的特征不仅能够反映样本的数据分布，同时还可能具有较高的分类精度. 为此，本文首先从 MSE 准则出发，给出了基于相关度的特征选择准则，然后把 LDA 思想与相关度准则相结合，提出了一种新的特征选择算法.

2　基于相关度的特征选择准则

设分类训练样本集合为

$$\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n},$$

其中 $\boldsymbol{x}_i\in\mathbf{R}^d$，$y_i\in\{-1,1\}$，$i=1,\cdots,n$.

令

$$\boldsymbol{X}=[x_1,x_2,\cdots,x_n],\boldsymbol{Y}=[y_1,y_2,\cdots,y_n].$$

这里 $\boldsymbol{X}$ 可以看作是一个 d 维随机向量，其各时刻的值为 $x_1,x_2,\cdots,x_n$；$\boldsymbol{Y}$ 可以看作是一个随机变量，其各时刻的值为 $y_1,y_2,\cdots,y_n$. 假设线性分类器模型为

$$y=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x},$$

其中 $\boldsymbol{w}$ 和 $\boldsymbol{x}$ 均是 d 维列向量，设该线性分类器的训练准则为

$$\min_{w}\|\boldsymbol{Y}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\|^2. \tag{1}$$

令

$$r=\|\boldsymbol{Y}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}\|^2,$$

则

$$r=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}-\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{Y}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}. \tag{2}$$

对已知的训练样本集合，$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 是常数，$\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\boldsymbol{Y}$，在两边分别乘以 $\boldsymbol{X}$ 和 $\boldsymbol{X}^{\mathrm{T}}$，并将 $\boldsymbol{X}$ 的自相关矩阵和 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的相关矩阵分别记为 $\boldsymbol{C}_{XX}$、$\boldsymbol{C}_{XY}$，则式（2）的等价形式为

$$r\boldsymbol{C}_{XX}=-2\boldsymbol{C}_{XY}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}+\boldsymbol{C}_{XX}\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}. \tag{3}$$

对已知训练样本集合 $\boldsymbol{X}$，自相关矩阵 $\boldsymbol{C}_{XX}$ 为已知的对称矩阵，即 $\boldsymbol{C}_{XX}=\boldsymbol{C}_{XX}^{\mathrm{T}}$. 在式（3）两边分别乘以 $\boldsymbol{w}^{\mathrm{T}}$ 和 $\boldsymbol{w}$，则变为

$$r\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}=(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}-2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY})\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}. \tag{4}$$

把式（4）看作为对 $\boldsymbol{w}$ 的最小优化求解问题，由于 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}$ 为标量，则式（1）表示的分类器训练准则等价于如下形式：

$$\min_{w}r=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}-2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY}. \tag{5}$$

根据经验风险最小化训练准则,如果要使式(1)所示的均方误差最小,那么就应该使得式(5)中 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}$ 尽可能最小,同时使得 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY}$ 值尽可能最大,这里系数向量 $\boldsymbol{w}$ 可以看作特征选择因子. 由相关性的物理意义可知,$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}$ 可以看作为数据样本特征之间相关性的度量;$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY}$ 可以看作为数据样本特征与类标号之间相关性的度量,本文把上述 2 种相关性度量分别称为特征相关度以及特征与类标号相关度. 很明显,要使分类器训练经验风险最小,特征选择因子 $\boldsymbol{w}$ 必须使特征相关度最小的同时使特征与类标号相关度最大. 基于上述考虑,本文给出一种基于相关度的特征选择准则(CMFS),即

$$\text{CMFS}: J_{LS1}(\boldsymbol{w}) = \max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}}.$$

上述准则较好地克服了目前特征选择算法中因没有考虑特征之间的相关性问题而导致分类器性能不佳的问题. 对 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w}$ 规格化,即 $\boldsymbol{w}^{\mathrm{T}}C_{XX}\boldsymbol{w} = 1$,那么 上面优化问题可写为如下形式:

$$\text{CMFS}: \begin{aligned} & J_{LS1}(\boldsymbol{w}) = \max_{w} \boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY}, \\ & \text{s.t. } \boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w} = 1. \end{aligned}$$

上述优化问题的目标函数与约束条件的 Lagrange 函数为

$$L(\boldsymbol{w}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY} - \lambda(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w} - 1). \tag{6}$$

由式(6)对向量 $\boldsymbol{w}$ 进行求导并等于 0 得

$$\boldsymbol{C}_{XY}\boldsymbol{C}_{YX}\boldsymbol{w} = 4\lambda^2\boldsymbol{C}_{XX}\boldsymbol{w},$$

即

$$(0.25\boldsymbol{C}_{XX}^{-1}\boldsymbol{C}_{XY}\boldsymbol{C}_{YX} - \lambda^2)\boldsymbol{w} = 0.$$

显然,优化问题的解是矩阵 $\boldsymbol{C}_{XX}^{-1}\boldsymbol{C}_{XY}\boldsymbol{C}_{YX}$ 的最大特征值对应的特征向量. 需要注意的是,上述特征选择准则是根据 MSE 准则推导出来的. 由第 1 节可知,根据 CMFS 求取的特征向量虽然具有较高的分类精度,但难以保证分类器的泛化性能. 而基于 LDA 求取的鉴别向量用于分类精度不高,但是能够反映样本的分布. 因此,下面本文把 LDA 思想引入到 CMFS,充分利用二者的优点同时克服它们的缺点.

3 类内方差与相关度结合的特征选择算法

结合 LDA 特征选择算法和 CMFS 的优点,提出了一种类内方差与相关度结合的特征选择准则(WVCMFS),即

$$\text{WVCMFS}: J_{\mathrm{s}}(\boldsymbol{w}) = \max_{w} \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XY}}{(1-a)\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w} + a\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{w}\boldsymbol{w}}. \tag{7}$$

其中,$a \in [0,1]$ 是一个正数因子,用来控制类内方差度在特征选择准则中的作用. 从式(7)可以看出,特征选择准则不仅考虑了样本特征与类标号之间的相关性以及特征之间的相关性,而且能够使得类内方差最小,因而选择得到的特征能够有效地提高分类器性能.

与第 2 节优化问题方法求解类似,首先对式(7)的优化问题中目标函数的分母进行规格化,即

$$(1-a)\boldsymbol{w}^{\mathrm{T}}\boldsymbol{C}_{XX}\boldsymbol{w} + a\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{w}\boldsymbol{w} = 1.$$

那么优化问题的解可通过式(8)求出:

$$(0.25((1-a)\boldsymbol{C}_{XX} + a\boldsymbol{S}_{w})^{-1}\boldsymbol{C}_{XY}\boldsymbol{C}_{YX} - \gamma^2)\boldsymbol{w} = 0. \tag{8}$$

可以看出,式(7)所示的优化问题的解是矩阵 $((1-a)\boldsymbol{C}_{XX} + a\boldsymbol{S}_{w})^{-1}\boldsymbol{C}_{XY}\boldsymbol{C}_{YX}$ 的最大特征值对应的特征向量 $\boldsymbol{w}^*$.

本文采用搜索特征选择方法从输入数据特征中选择一定数量的特征.

假设待选的特征集合为 $S_0 = \{x_1, x_2, \cdots, x_d\}$,已选特征集合为 F_0,从 S_0 中去除一个特征 x_i,剩余特征集合记为 $\{S_0 - x_i\}$,然后根据式(8)计算特征集合 $\{S_0 - x_i\}$ 对应的最大特征值,记为 $\gamma_{\{S_0 - x_i\}}$,$i = 1, 2, \cdots, d$. 根据 $\gamma_{\{S_0 - x_j\}} = \min_i \gamma_{\{S_0 - x_i\}}$ 选择特征 x_j,并记 $S_1 = \{S_0 - x_j\}$ 为 S_0 去除特征 x_j 得到的新特征集合,已选特征集合变为 $F_1 = \{F_0, x_j\}$. 同样使用上述方法从待选特征集合 S_1 中选择下一个特征到 F_1 中,此时已选、候选特征集合分别变为 F_2 和 S_2. 依次类推,直至选取预先设定的 M 个特征.

可以看出,上述特征选择方法每一次只能选择一个特征,在选择过程中已选特征与待选特征之间的相关性并不能被很好地考虑,不能保证最终选择的特征子集是最优的.

4 推广的类内方差与相关度结合的特征选择算法

由于上述特征选择算法针对的是线性分类数据样本,而实际很多分类数据样本是非线性的,为了寻找一种针对非线性分类数据样本的特征选择算法,本文使用核方法把上述算法推广到解决非线性特征选择问题.

首先,通过非线性特征映射 $\psi(\cdot)$ 将分类训练样本集合 $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ 映射到高维特征空间中,特征空间分类样本变为 $\{(\psi(\boldsymbol{x}_i), y_i)\}_{i=1}^{n} \subset \mathbf{R}^{N_F}$,然后,在高维特征空间 $\mathbf{R}^{N_F}$ 上的样本集进行分类.

在特征空间上线性分类器的形式为

$$\boldsymbol{Y} = \boldsymbol{w}_{\psi}^{\mathrm{T}} \boldsymbol{Z},$$

其中：

$$\boldsymbol{Z} = [\psi(\boldsymbol{x}_1), \psi(\boldsymbol{x}_2), \cdots, \psi(\boldsymbol{x}_n)];$$

$\boldsymbol{w}_{\psi}$为高维特征空间中的线性分类器法向量，

$$\boldsymbol{w}_{\psi} = \sum_{i=1}^{n} \beta_i \psi(\boldsymbol{x}_i) = \boldsymbol{Z}\boldsymbol{\beta};$$

$$\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_n]^{\mathrm{T}} \in \mathbf{R}^n.$$

对一个非线性特征映射$\psi(\cdot)$，存在一个核函数$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y})$，使得

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{x})^{\mathrm{T}} \psi(\boldsymbol{y}).$$

式(8)可以写为如下形式：

$$\boldsymbol{Z}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{\beta} = 4\gamma^2((1-a)\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} + a\boldsymbol{S}_{u\psi})\boldsymbol{Z}\boldsymbol{\beta}. \quad (9)$$

定义核矩阵$\boldsymbol{K}$和$\boldsymbol{H}$为

$$\boldsymbol{K} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z} = [\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1.2,\cdots,n},$$

$$\boldsymbol{H} = \boldsymbol{K} - \boldsymbol{l}_n\boldsymbol{K} - \boldsymbol{K}\boldsymbol{l}_n + l_n\boldsymbol{K}\boldsymbol{l}_n,$$

其中$\boldsymbol{l}_n = [l/n]_{n\times n}$.

对式(9)两边分别乘以$\boldsymbol{Z}^{\mathrm{T}}$，如果$(1-a)\boldsymbol{K}\boldsymbol{K} + a\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}$可逆，则式(5)可写为

$$(((1-a)\boldsymbol{K}\boldsymbol{K} + a\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{K}\boldsymbol{\nu}\boldsymbol{K} - 4\gamma^2)\boldsymbol{\beta} = 0. \quad (10)$$

式中$\boldsymbol{\nu} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$是$n\times n$的矩阵. 很显然，针对非线性分类问题，特征选择的优化问题解是$((1-a)\boldsymbol{K}\boldsymbol{K} + a\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H})^{-1}\boldsymbol{K}\boldsymbol{\nu}\boldsymbol{K}$最大特征值对应的特征向量$\boldsymbol{\beta}^*$.

为防止式(10)矩阵求逆时，因矩阵奇异而导致数值求解不稳定问题，可使用$k\boldsymbol{I}$作为正则化项来解决矩阵奇异问题[6]，这里$\boldsymbol{I}$是单位矩阵，k是一个正数. 因此式(10)写为

$$(((1-a)\boldsymbol{K}\boldsymbol{K} + a\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H} + k\boldsymbol{I})^{-1}\boldsymbol{K}\boldsymbol{\nu}\boldsymbol{K} - 4\gamma^2)\boldsymbol{\beta} = 0.$$

虽然用核方法将第 3 节提出的特征选择算法扩展为针对非线性分类问题的特征选择算法，但是，本节针对非线性分类问题特征选择算法的优化问题解$\boldsymbol{\beta}^*$与第 3 节针对线性分类问题特征选择算法的优化问题解$\boldsymbol{w}^*$的含义是不同的. $\boldsymbol{w}^*$可以看成是特征压缩因子，其分量的绝对值大小代表相应特征的重要性，因而可以直接用于特征选择. 然而$\boldsymbol{\beta}^*$的维数与训练样本的数目一致，其分量的绝对值大小可以看成是训练样本在分类器模型中的重要程度. 因此，基于核方法扩展的特征选择算法只能使用第 3 节算法选择特征子集.

5　仿真实验

本文算法是 Filter 方法，与考虑了特征之间相关性的 ABB[9]、RELIEF－F[10] 和 CFS[4]、CPSS[5]共 4 种 filter 方法进行比较. 实验中使用 RBF 核函数，核函数参数$\sigma = 8$，正则化参数$k = 5$，模型参数使用交叉验证方法确定. 实验中使用 UCI 的 breast－w、wine、iris、breast-cancer、sonar 数据集，使用 10-fold 交叉验证确定算法的预测结果. 由于 filter 算法是独立于分类器的特征选择算法，因此本文使用 SVM 对选择特征后的数据集进行分类，且 SVM 同样采用 RBF 核函数，参数与本文算法的核函数参数相同. ABB、RELIEF-F 、CFS 和 CPSS 分别采用核函数参数为 10、15、20 和 10，且使用它们中最好的结果与本文算法进行比较，结果如表 1 所示.

从表 1 可以看出，本文算法总体上取得令人满意的分类精度，但是对个别数据集，其性能不是最好的，原因是数据集含有的特征数目太少，不能体现本算法的优势，如 wine 数据集只有 13 个特征，对特征数量较多、特征之间相关性较大的数据集，如 sonar 数据集(60 个特征)，本文算法能够取得较高的分类精度.

表 1　本文算法与 ABB、RELIEF－F、CFS 分类精度和 CPSS 方式分类精度比较

数据集	分类精度				
	ABB	RELIEF－F	CFS	CPSS	本文算法
breast-w	0.767 ±0.006	0.667 ±0.017	0.881 ±0.009	0.703 ±0.017	0.756 ±0.009
wine	0.889 ±0.009	0.929 ±0.024	0.958 ±0.012	0.896 ±0.013	0.929 ±0.015
iris	0.950 ±0.012	0.935 ±0.025	0.954 ±0.010	0.955 ±0.010	0.958 ±0.012
breast-cancer	0.699 ±0.014	0.708 ±0.008	0.715 ±0.020	0.723 ±0.021	0.729 ±0.016
sonar	0.700 ±0.021	0.700 ±0.020	0.701 ±0.025	0.720 ±0.029	0.730 ±0.031

6　结　论

1)本文从均方误差最小的分类训练准则出发，并借鉴线性鉴别分析的思想，提出了一种类内方差与相关度结合的特征选择算法，并使用核方法将其推广到可以解决非线性分类的特征选择问

题. 该算法不仅同时考虑了样本特征之间以及特征与分类标号之间的相关性,而且使得类内方差最小,有效地提高了分类器的性能.

2)仿真实验表明,本文算法适用于对特征数量多、特征之间相关性强的数据集进行特征选择,其选择的特征子集能够显著提高分类精度,具有较大的优越性.

3)尽管本文算法能够有效选取样本特征,但是如何从理论上确定合适的选取特征数目仍然是一个需要解决的问题.

参考文献:

[1] THEODORIDIS S, KOUTROUMBAS K. 模式识别[M]. 北京: 电子工业出版社, 2006: 171 – 183.

[2] VERRON S, TIPLICA T, KOBI A. Fault detection and identification with a new feature selection based on mutual information[J]. Journal of Process Control, 2008, 18(5): 479 – 490.

[3] CHO H. An orthogonally filtered tree classifier based on nonlinear kernel-based optimal representation of data [J]. Expert Systems With Applications, 2008, 34 (2): 1028 – 1037.

[4] HALL M A, SMITH L A. Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper[C]//Proceedings of the Twelfth International Florida AI Research Society Conference. Menlo Park: AAAI Press, 1999: 235 – 239.

[5] 詹德川,周志华. 基于相关投影分的特征选择算法[J]. 计算机科学与探索, 2007, 1(2): 138 – 145.

[6] XIANG C, FAN X, LEE T H. Face recognition using recursive fisher linear discriminant[J]. IEEE Transactions on Image Processing, 2006, 15 (8): 2097 – 2105.

[7] YE J, XIONG T. Computational and theoretical analysis of null space and orthogonal linear discriminant analysis [J]. Journal of Machine Learning Research, 2006, 7 (7): 1183 – 1204.

[8] HALL M A, SMITH L A. A method of feature selection based on particle swarm optimization algorithm with trans-gene operator[C]// 2008 Chinese Control and Decision Conference. Piscataway: IEEE Computer Society, 2008: 3568 – 3571.

[9] PUDIL P, NOVOVICOVA J, KITTLER J. Floating search methods in feature selection[J]. Pattern Recognition Letters, 1994, 15(11): 1119 – 1125.

[10] LIU H, MOTODA K, DASH M. A monotonic measure for optimal feature selection[C]//Proceedings of the 10th European Conference on Machine Learning. Berlin: Springer-Verlag, 1998: 101 – 106.

(编辑 杨 波)

(上接第 131 页)

成电解加工阴极运动设计,精确地控制实际加工中去除材料的厚度,具有较高的工程实用价值.

参考文献:

[1] 徐家文,云乃彰,严德荣. 数控电解加工整体叶盘的研究、应用和发展[J]. 航空制造技术,2003,(6): 31 – 34.

[2] 康敏,徐家文. 精密展成电解加工整体叶轮的展成运动轨迹确定[J]. 中国机械工程, 2006,17(4): 346 – 349.

[3] 朱永伟,徐家文,赵建社. 大扭曲叶片整体涡轮盘电解加工关键技术[J]. 哈尔滨工业大学学报,2008,40 (5): 783 – 787.

[4] XU Jiawen, YUN Naizhang, TANG Yangxin, *et al*. The modelling of NC-electrochemical contour evolution machining using a rotary tool-cathode[J]. Journal of Materials Processing Technology, 2005,159:272 – 277.

[5] KOZAK J, FUDZYNSKI A F, DOMANOWSKI P. Computer simulation electrochemical shaping (ECM-CNC) using a universal tool electrode[J]. Journal of Materials Processing Technology,1998, 76:161 – 164.

[6] KOZAKA J, CHUCHR M, RUSZAJ A, *et al*. The computer aided simulation of electrochemical process with universal spherical electrodes when machining sculptured surfaces[J]. Journal of Materials Processing Technology,2000,107: 283 – 287.

[7] DOMANOWSKI P, KOZAK J. Direct and inverse problems of shaping by electrochemical generating machining [J]. Journal of Materials Processing Technology, 2000, 107:300 – 306.

[8] RUSZAJ A, ZYBURA-SKRABALAK M. The mathematical modelling of electrochemical machining with flat ended universal electrodes[J]. Journal of Materials Processing Technology, 2001,109:333 – 338.

[9] GONG Hu, CAO Lixin, LIU Jian. Improved positioning of cylindrical cutter for flank milling ruled surfaces[J]. Computer-Aided Design, 2005, 37: 1205 – 1213.

[10] 于源,赖天琴,员敏,等. 基于特征的直纹面 5 轴侧铣精加工刀位计算方法[J]. 机械工程学报,2002,38 (6):130 – 133.

(编辑 杨 波)