ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 126－136
DOI:10.16337/j.1004-9037.2015.01.012

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于特征间距的二次规划特征选取算法

刘全金[1,2,3]　赵志敏[1,3]　李颖新[4,5]

(1.南京航空航天大学理学院,南京,210016; 2.安庆师范学院物理与电气工程学院,安庆,246011;
3.江苏省光谱成像与智能感知重点实验室,南京理工大学,南京,210094; 4.北京市轻纺机械机器视觉工程技术研究中心,北京,100176;5.北京经纬纺机新技术有限公司,北京,100176)

摘　要:提出一种基于特征间距的二次规划特征选取算法。首先,将特征在类内样本间和异类样本间的距离分别作为二次规划算法目标函数的二次项和一次项参数,用以搜索类内紧密、内间分离的分类特征;同时,通过对二次项和一次项的归一化来均衡特征在同类样本和异类样本之间的关系;然后,将二次规划算法优化后的最优解向量作为衡量特征对分类贡献的权重向量,再根据特征权重高低选取分类特征。特征选取方法在6个数据集中的特征选取实验结果表明了该方法的可行性和有效性。

关键词:特征选取;特征间距;分类;二次规划

中图分类号:TP391　　**文献标志码**:A

Feature Selection Algorithm Based on Quadratic Programming with Margin Between features

Liu Quanjin[1,2,3], Zhao Zhimin[1,3], Li Yingxin[4,5]

(1. College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China; 2. Department of Physics, Anqing Normal College, Anqing, 246011, China; 3. Jiangsu Key Laboratory of Spectral Imaging & Intelligent sense, Nanjing University of Science and Technology, Nanjing, 210094, China; 4. Beijing Light Industry and Textile Machinery Engineering Research Center for Machine Vision, Beijing, 100176, China; 5. Beijing Jingwei Textile Machinery New Technology Co., Ltd., Beijing, 100176, China)

Abstract: A feature selection algorithm using quadratic programming is proposed based on feature margins. Firstly, the inner-class distance of features is taken as the coefficient of the quadratic terms in the objective function and the inter-class distance of features is used as the coefficient of the linear terms for searching informative features. The elements of the quadratic terms and the linear terms are normalized to balance the feature relation between inner class and inter-class. Then, the optimal solution vector is taken as the feature weight vector for selecting informative features. Finally, experiments on six different datasets show the effectiveness and feasibility of the proposed method.

Key words: feature selection; distance between features; classification; quadratic programming

基金项目:国家自然科学基金(61475071,10172043,61173068资助项目;教育部博士点基金(20093218110024)资助项目;江苏省光谱成像与智能感知重点实验室基金(30920130122003)资助项目;安徽省教育厅自然科研基金重点(KJ2010A226)资助项目。
收稿日期:2014-07-14;**修订日期**:2014-10-12

引 言

特征选取是样本分类之前的预处理过程,对样本(特别是高维特征数据集样本)分类至关重要[1,2]。特征选取被广泛应用于图像处理、文本分类以及生物信息学等领域,从样本特征中剔除与样本类别无关的噪声和冗余特征,选出含有丰富类别信息的分类特征组合,以降低特征空间维数,提高分类性能[3,4]。

特征选取方法分为 filter,wrapper 和 embedded 三类:filter 方法独立于分类模型,根据个体的类别可分性指标进行特征选取;wrapper 方法在特征空间搜索特征子集,并基于分类结果考察特征子集的分类性能,选取最佳的特征子集作为分类特征集合;embedded 方法将特征选取植入到分类学习等算法中,与分类模型的目标函数相结合进行特征搜索[5-8]。filter 方法运行效率高,但因可分性指标仅孤立地考察特征个体对样本可分性的贡献,所以选取特征的组合分类能力相对较弱[9],特征可分性指标有 fisher score、互信息和 ReliefF 等[6,10]。wrapper 方法的效率决定于特征搜索算法的复杂度,高维特征空间中搜索特征子集属于 NP 难题,常用浮动搜索算法、分支边界法、遗传算法等贪婪搜索算法。embedded 方法基于分类学习等算法,通过目标函数提高特征集合的分类性能并降低特征集合维数[11]。从某种程度上讲,wrapper 方法和 embedded 方法是特征搜索和分类模型的相互融合。wrapper 方法和 embedded 方法常采用交叉验证方法测试特征集合分类性能,分类模型选用决策树、贝叶斯分类器、K 近邻(K nearest neighbor,KNN)、BP 神经网络和支持向量机等[7-9,12]。

数学规划在特征选取中起着极其重要的作用[13]。二次规划算法被用于从高维特征数据集中选取具有高识别率的特征集合。Rodriguez-Lujan 提出二次规划特征选取方法(Quadratic programming feature selection,QPFS),并用文献[8]方法简化高维数据特征选取时的计算复杂度。QPFS 二次规划目标函数中,二次项基于特征间的互信息考察特征对样本类别的贡献,一次项基于特征与样本类别的相关性考察特征对样本类别的贡献。在文献[14]中,Rodriguez-Lujan 又提出基于核空间的二次规划特征选取方法(Kernel quadratic programming feature selection,KQPFS)。

本文提出基于特征间的距离运用二次规划算法进行特征选取(Quadratic programming feature selection based on feature distance,DQPFS)。首先,将特征在类内和类间的加权距离作为二次规划算法目标函数中的二次项和一次项参数,并通过对二次项和一次项的归一化来均衡特征在同类和异类之间的关系;然后,把优化后的最优解向量作为衡量特征对分类贡献的权重向量,根据特征权重选取有利于分类的特征集合。

本文将 DQPFS 方法、SVM-RFE 方法[11]和 KQPFS 方法[14]在 6 个数据集上进行了特征选取实验。DQPFS 方法选取特征集合的分类能力优于其他 2 种方法选取的特征集合的分类能力,表明 DQPFS 方法在 6 个数据集中的特征选取性能优于其他 2 种方法,说明该方法在特征选取中的有效性,可应用于生物信息学和文本分类等机器学习领域的特征选取中。

1 基于二次规划算法的特征选取方法

二次规划的目标函数是二次实函数,其约束为线性约束。对于二次规划问题

$$\min f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}$$

$$\text{s.t. } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{1}$$

式中:$\boldsymbol{H}$ 是对称矩阵,$\boldsymbol{A}$ 为约束矩阵。通过求解二次归划,搜索最优解 x 使目标函数 $f(\boldsymbol{x})$ 在可行域内达到极小。求解二次规划的典型算法有 Lagrange 方法、Active set 方法、Lemke 方法和路径跟踪法等[15]。

1.1 KQPFS 特征选取方法

对于数据集 $\boldsymbol{D}\in\mathbf{R}^{M\times N}$($M$ 是特征数,N 是样本数),Rodriguez-Lujan 提出 QPFS 特征选取方法[8]。定义基于二次规划的目标函数和约束条件为

$$\min \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}x-\boldsymbol{F}^{\mathrm{T}}\boldsymbol{x}$$

$$\text{s. j.}\quad x_i\geqslant 0,\sum_{i=1}^{M}x_i=1,i=1,\cdots,M \tag{2}$$

式中:$\boldsymbol{Q}\in\mathbf{R}^{M\times M}$为对称正定矩阵,$Q(i,j)$表示数据集 D 中第 i 个特征与第 j 个特征间的相似程度;$\boldsymbol{F}$ 为 M 维向量,表示数据集 D 的特征与样本类别的相关性;M 维向量 $\boldsymbol{x}$ 满足约束条件。

KQPFS 方法用高斯核函数 $\Phi(\boldsymbol{d})$ 修改 QPFS 方法目标函数中二次项 $Q=\sum_{\boldsymbol{d}\in\boldsymbol{D}}\{[\Phi(\boldsymbol{d})-m^{\Phi}]\times[\Phi(\boldsymbol{d})-m^{\Phi}]^{\mathrm{T}}\}$ 和一次项 $F=\sum_{\boldsymbol{d}\in\boldsymbol{D}}(y_d-m^{y})(\Phi(\boldsymbol{d})-m^{\Phi})$,其中 $\boldsymbol{d}$ 为数据集 $\boldsymbol{D}$ 中的一个样本,y_d 为样本 $\boldsymbol{d}$ 的类别标签,$m^{\Phi}=\frac{1}{N}\sum_{\boldsymbol{d}\in\boldsymbol{D}}\Phi(\boldsymbol{d})$,$m^{y}=\frac{1}{N}\sum_{\boldsymbol{d}\in\boldsymbol{D}}\boldsymbol{y}_d$。

KQPFS算法通过最小化目标函数获取特征的权重向量 $\boldsymbol{x}$,其目标是寻找特征间相似程度低、特征与样本类别相关性强的特征。权重向量 $\boldsymbol{x}$ 的元素值表示每个特征的权重,权重值高低反映了它与其他特征间的相似程度以及它与样本类别的相关性,权重值越高的特征对样本分类越重要[8,14]。

1.2 基于特征间距的二次规划特征选取方法

具有丰富样本类别信息的特征在同类样本中分布紧密、异类样本中分布松散[16-19]。本文将特征在同类样本间距离和异类样本间距离分别作为二次规划算法目标函数的二次项和一次项参数,通过二次规划算法寻找类内紧密、类间分散的具有丰富类别信息的特征。

定义二次规划问题

$$\min f(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{x}-\boldsymbol{C}^{\mathrm{T}}\boldsymbol{x}$$

$$\text{s. t.}\quad x_i\geqslant 0,\sum_{i=1}^{M}x_i=1,i=1,\cdots,M \tag{3}$$

式中 $\boldsymbol{H}$ 矩阵的元素为特征在同类样本间的加权距离,有

$$H(i,j)=\frac{1}{\mathrm{card}(D)}\sum_{k=1}^{P}(\mathrm{card}(D_k)\cdot\mathrm{Dist}_k(i,j)) \tag{4}$$

式中 $\mathrm{Dist}_k(i,j)$表示在第 k 类样本中第 i 个特征与第 j 个特征间的距离

$$\mathrm{Dist}_k(i,j)=\frac{|\mu_k(i)-\mu_k(j)|}{\sigma_k(i)+\sigma_k(j)} \tag{5}$$

式中:$\mu_k(i)$表示第 i 个特征在第 k 类内样本中的均值;$\sigma_k(i)$表示第 i 个特征在第 k 类内样本中的标准差;定义两个特征在第 k 类的均值差 $|\mu_k(i)-\mu_k(j)|$ 为两个特征的类内间距;用 $\sigma_k(i)+\sigma_k(j)$对特征间距进行标准化,标准化后的特征的均值差异体现了两个特征在类内的距离;$\mathrm{card}(D)$表示数据集中样本总数;$\mathrm{card}(D_k)$表示数据集 D 中第 $k\in(1,\cdots,P)$类样本的数量,则$\frac{\mathrm{card}(D_k)}{\mathrm{card}(D)}$表示第 k 类样本在数据集中所占比例,用该比例对特征间距 $\mathrm{Dist}_k(i,j)$加权,通过距离加权降低因样本类别分布不均匀对分析特征间距的影响。

式(3)中 $\boldsymbol{C}$ 向量的元素为特征在异类样本间的加权距离

$$C(i)=\frac{1}{\mathrm{card}(D)}\sum_{k=1}^{P}(\mathrm{card}(D_k)\mathrm{Dist}_k(i)) \tag{6}$$

式中:$\mathrm{Dist}_k(i)=\dfrac{|\mu_k(i)-\mu_{k-}(i)|}{\sigma_k(i)+\sigma_-(i)}$表示第 i 个特征在第 k 类样本与其他类别样本间的距离。其中,$\mu_{k-}(i)$表示第 i 个特征在第 k 类以外样本中的均值,$\sigma_{k-}(i)$表示第 i 个特征在第 k 类以外样本中的标准差。

由式(3)知,二次规划算法优化后的最优解向量与特征向量一一对应,最优解向量元素值的高低反映了对应特征在类内与其他特征间的亲疏关系以及特征在异类样本间的距离关系。本文将最优解向量视为衡量特征对样本分类贡献大小的特征权重向量。

为使优化后的最优解能均衡体现特征在类内和类间关系,本文对二次项和一次项作归一化处理;同时,为满足二次规划对二次项 $\boldsymbol{H}$ 矩阵的正定性要求,还对 $\boldsymbol{H}$ 矩阵的对角线元素最大化等处理。具体算法如下:

(1)将 $\boldsymbol{H}$ 矩阵和 $\boldsymbol{C}$ 向量的元素减去各自的最小值,使其元素为非负数。

(2)对 $\boldsymbol{H}$ 矩阵和 $\boldsymbol{C}$ 向量元素做归一化处理:

$$H(i,j)=H(i,j)/\sum_{m=1}^{M-1}\Big(\sum_{n=m+1}^{M}(H(m,n))\Big),C(i)=C(i)/\sum_{m=1}^{M}(C(m))$$

(3)最大化 $\boldsymbol{H}$ 矩阵的对角线元素:$\boldsymbol{H}=\boldsymbol{H}+\gamma\cdot\boldsymbol{E}$。$\sum_{m=1}^{M-1}\Big(\sum_{n=m+1}^{M}(H(m,n))\Big)$为 $\boldsymbol{H}$ 矩阵上三角元素的和;$\boldsymbol{E}$ 为与 $\boldsymbol{H}$ 矩阵同阶的单位矩阵,γ 为大于等于 1 的标量。

归一化处理二次项和一次项参数,使优化后的特征权重能均衡体现特征在同类和异类样本中的作用,有利选取出类内紧密、类间分散的具有丰富类别信息的分类特征。

对于样本数大于特征数的数据集,$\boldsymbol{H}$ 矩阵的奇异程度相对较低,二次规划的计算复杂度也较低;而对于特征数大于样本数的数据集,$\boldsymbol{H}$ 矩阵是奇异的,二次规划的计算复杂度较高。本文借鉴文献[20]中特征空间向样本空间转换的线性变换方法,建立特征权重向量 $\boldsymbol{x}$ 与数据集 D 样本 $d_i,i\in(1,\cdots,N)$之间的线性关系

$$\boldsymbol{x}=\sum_{i=1}^{N}\beta_i\boldsymbol{d}_i=\boldsymbol{D}\cdot\boldsymbol{\beta} \tag{7}$$

式中 $\boldsymbol{\beta}=[\beta_1,\beta_2,\cdots,\beta_N]^{\mathrm{T}}$ 为向量 $\boldsymbol{x}$ 在各样本上的权重系数。

该线性变换相当于样本在原特征空间的张集[20]。将这种变换用于式(3)的二次规划中,分析高维数据集的特征重要性,二次规划的目标函数和约束条件为

$$\min f(\boldsymbol{x})=\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\cdot[\boldsymbol{D}^{\mathrm{T}}\cdot\boldsymbol{H}\cdot\boldsymbol{D}]\cdot\boldsymbol{\beta}-[\boldsymbol{C}^{\mathrm{T}}\cdot\boldsymbol{D}]\cdot\boldsymbol{\beta}$$

$$\text{s.t.}\quad \beta_i\geqslant 0,\sum_{i=1}^{N}\beta_i=1,i=1,\cdots,N \tag{8}$$

变换后的$[\boldsymbol{D}^{\mathrm{T}}\cdot\boldsymbol{H}\cdot\boldsymbol{D}]$是半正定矩阵,$[\boldsymbol{D}^{\mathrm{T}}\cdot\boldsymbol{H}\cdot\boldsymbol{D}]$的奇异程度大大降低,二次规划算法的计算复杂度也大大降低。优化后的最优解向量 $\boldsymbol{\beta}$ 经式(7)变换得到特征权重向量 $\boldsymbol{x}$。因为式(7)是线性变换,所以特征权重向量 $\boldsymbol{x}$ 仍能反映特征在类内的紧密度和类间的分散度,权重高的特征承载的样本类别信息比权重低的特征丰富。

2 特征选取实验

2.1 实验数据集

本文用 DQPFS 方法选取 6 个数据集的分类特征。其中有 2 个数据集是来自于加州大学欧文分校机器学习数据库中的 Ionosphere 和 Promoters 数据集[21];另外 4 个数据集为基因表达谱数据集 Acute Leuke-

mia[22], Multiple myeloma[23], Colon[24] 和 DLBCL[25]。数据集的基本信息如表 1 所示。

表 1 实验数据集结构信息

Table 1 Information of datasets used for experiments

数据集	特征数	样本数(+/−)	训练集/校验集
Ionosphere	34	351(126/225)	211/140
Promoters	57	106(53/53)	64/42
Acute Leukemia	7129	72(47/25)	44/28
Multiple myeloma	7129	105(74/31)	63/42
Colon	2000	62(40/22)	38/24
DLBCL	7129	77(58/19)	47/30

2.2 递归特征选取过程

本文还将 KQPFS 方法和 SVM-RFE 方法在这 6 个数据集中进行了的特征选取实验。SVM-RFE 方法基于支持向量机分类模型在递归特征剔除(Recursive feature eliminate, RFE)过程中分析特征对分类的重要性,选取分类特征集合[11]。KQPFS 方法在核空间中获取使 QPFS 算法中二次规划目标函数最大化的特征权重向量[14]。

在 6 个数据集中分别重复进行 40 次特征选取实验,通过实验结果比较 3 种特征选取方法的性能。每次实验都将数据集样本按 3:2 随机分配至训练集和校验集,然后,在递归特征选取过程中[11],特征选取方法基于训练集样本分析特征所含的类别信息,依次选取对分类影响最多的若干个特征组成分类特征集合;最后,基于这些特征集合用训练集训练分类模型识别校验集样本类别。分类能力最强的低维特征集合为最佳特征集合[11,17]。

3 种方法在相同的训练集中分析特征重要性、生成嵌套的分类特征集合;用相同的校验集检验所选特征集合的分类性能。3 种方法 40 次特征选取实验得到的分类统计结果反映其在对应数据集中的特征选取性能。

2.3 特征选取实验参数设置

特征选取实验是在装配 Intel Core i5-3470 CPU 和内存 4.00 GB 的 PC 机上,用 MATLAB 软件(version 8.0.0.783)完成;二次规划函数选用是 MATLAB 的"optim"工具箱里的 quadprog 函数。

对于 2 个特征个数低于样本个数的 Ionosphere 和 Promoters 数据集,DQPFS 方法以式(3)作为二次规划算法的目标函数优化特征权重向量;对于特征个数高于样本数的 4 个基因表达谱数据集,DQPFS 方法先用式(8)作为二次规划算法的目标函数,然后再用式(7)将优化后的最优解向量线性变换为特征权重向量。

经在 6 个数据集上实验验证,目标函数的二次项 $\boldsymbol{H}$ 矩阵的对角线元素最大化参数 γ 取 1～9 之间的整数值时,均能得到较好的特征选取结果。本文所列实验结果为 γ 取 5 时得到。

3 种方法在递归特征选取过程中特征选取率均为 50%,即从前一轮特征集合中选取权重最高的 50%特征组成新的特征集合。选用 SVM 和 KNN[19] 两个分类器作为检验特征集合分类性能的分类模型。

考虑到数据集不同类别样本数量不均匀,为了较客观地分析 3 种特征选取方法性能,本文同时以被选分类特征集合的分类准确率和 AUC(Area under ROC curve)值[26]作为衡量特征集合分类能力的标准。分类准确率和 AUC 值高低反映了特征集合的分类能力,进而体现特征选取方法的性能。

2.4 实验结果比较

3 种方法在 6 个数据集上选取的特征集合分类结果如图 1～6 所示。 纵轴代表 40 次特征选取实验

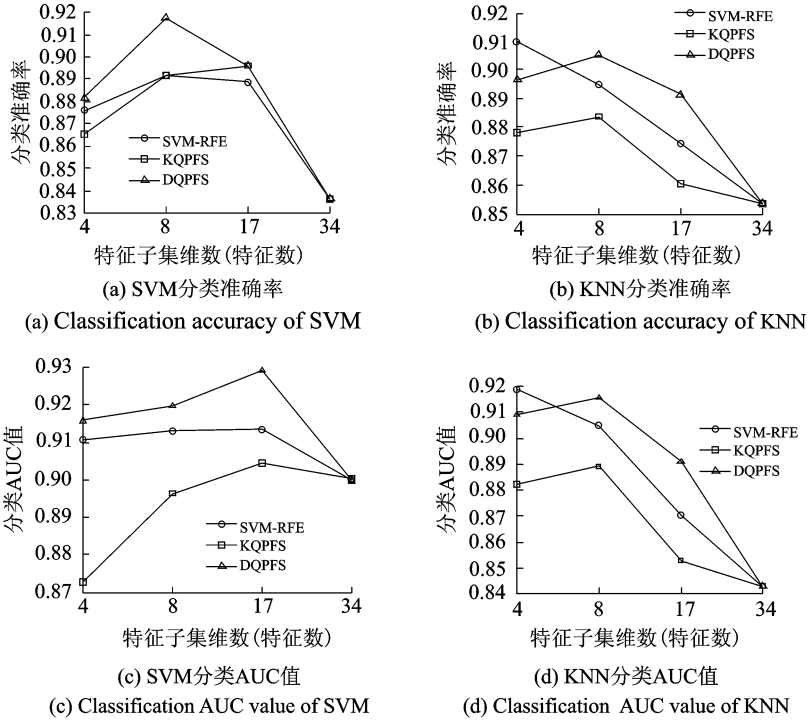

图 1 特征集合在 Ionosphere 数据上的分类结果

Fig. 1 Classification result of feature subsets on Ionosphere validation dataset

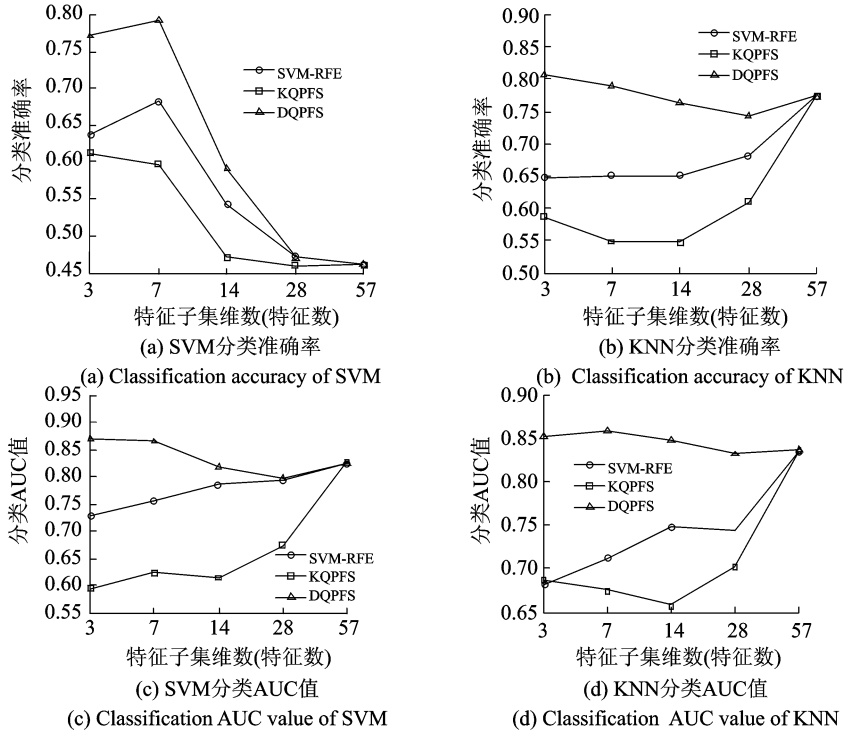

图 2 特征集合在 Promosters 数据上的分类结果

Fig. 2 Classification result of feature subsets on Promosters validation dataset

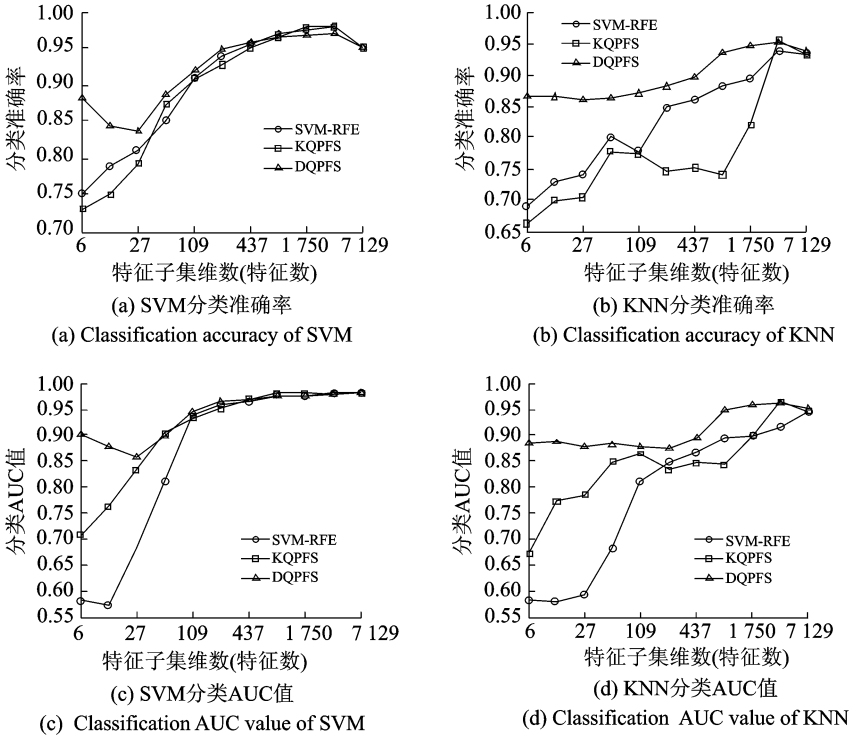

图 3 特征集合在 Acute Leukemia 数据上的分类结果

Fig. 3 Classification result of feature subsets on Acute Leukemia validation dataset

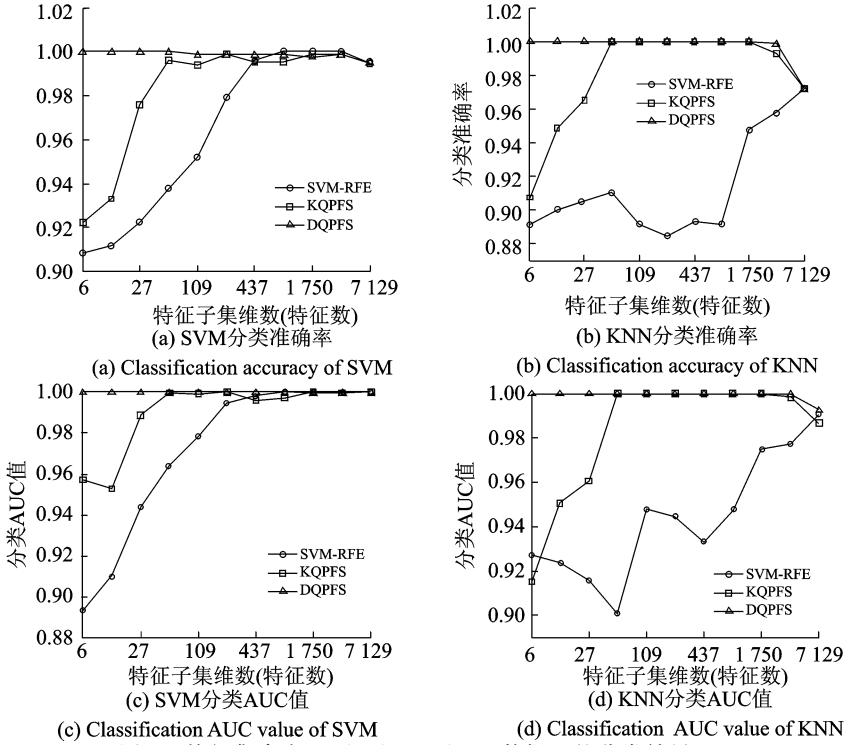

图 4 特征集合在 Multiple myeloma 数据上的分类结果

Fig. 4 Classification result of feature subsets on Multiple myeloma validation dataset

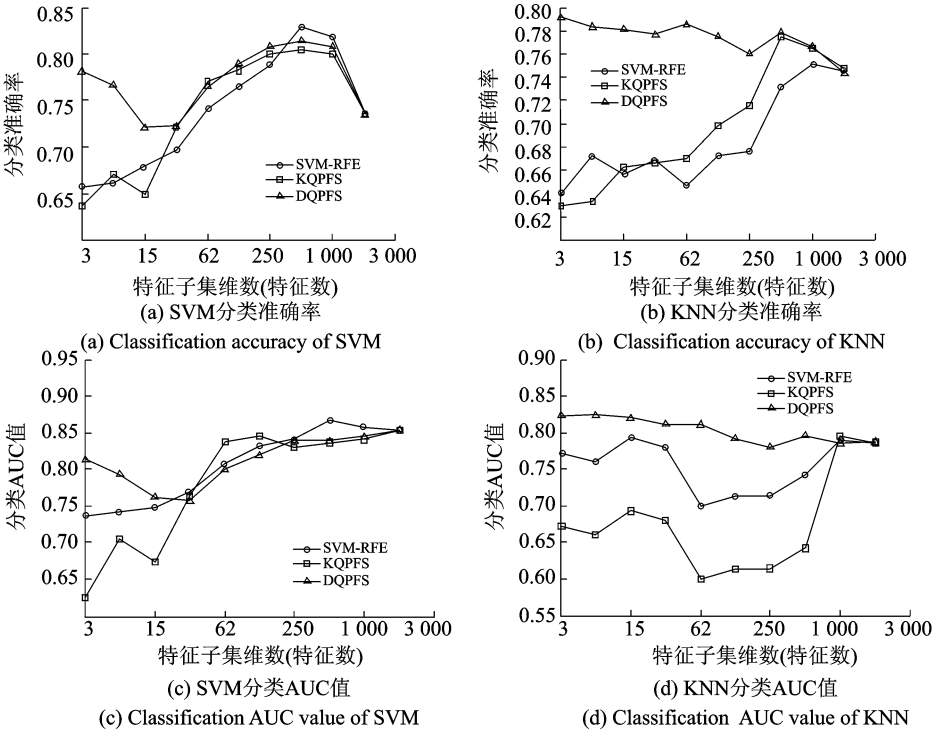

图 5 特征集合在 Colon 数据上的分类结果

Fig. 5 Classification result of feature subsets on Colon validation dataset

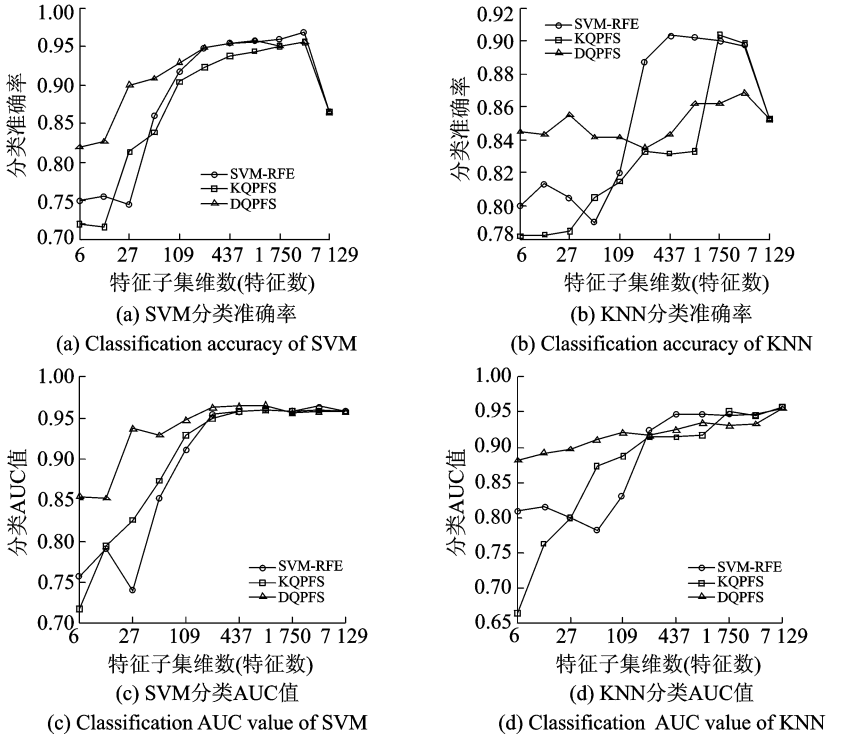

图 6 特征集合在 DLBCL 数据上的分类结果

Fig. 6 Classification result of feature subsets on DLBCL validation dataset

得到的嵌套特征集合对校验集样本分类正确率或 AUC 值的统计平均值，横轴表示特征集合维数，即特征集合中的特征数。

由图 1～3 可知，在 Ionoshpere，Promosters 和 Acute Leukemia 数据集上，DQPFS 选取的特征集合分类性能优于 KQPFS 和 SVM-RFE 方法选取的特征集合分类性能。

如图 4 所示，在 Multiple myeloma 数据集上，3 种方法选取的特征集合都有较好分类能力，其中 DQPFS 方法选取出的低维特征集合仍能准确识别所有校验集样本类别。

由图 5 可知，在 Colon 数据集上，DQPFS 方法选取的低维特征集合的分类结果要优于其他两种方法选取的特征集合分类结果。

如图 6 所示，在 DLBCL 数据集上，DQPFS 方法选取的高维特征集合的 KNN 分类结果比其他两种方法选取的特征集合分类结果差，但 DQPFS 方法选取的低维特征集合的分类能力强于其他低维特征集合分类能力。

4 个基因表达谱数据集的特征比较多，其中有不少“噪声特征”，这些“噪声特征”势必会影响特征选取效果。本文将 3 种特征选取方法置于相同的数据环境中进行特征选取。实验结果表明，本文提出的 DQPFS 方法有较好的抗“噪声”能力，在相同的环境中能选取出分类能力较强的分类特征集合。特征选取方法耗时长短与算法复杂度、特征选取范围及样本数有关。表 2 列出了 3 种特征选取方法在 6 个数据集中选取特征集合耗用的平均时间。

表 2 特征集合选取时间对照表 s

Table 2 Average time of selecting feature subsets by each method on five datasets

数据集	SVM-RFE	KQPFS	DQPFS
Ionosphere	4.298	0.062	0.109
Promosters	0.747	0.012	0.262
Acute Leukemia	1.359	0.086	326.3
Multiple myeloma	2.687	0.116	374.1
Colon	0.861	0.028	61.3
DLBCL	1.560	0.081	350.5

DQPFS 方法所耗时间主要用于计算特征间的距离，耗时随着特征选取范围的增加而增加。KQPFS 方法在 6 个数据集中选取特征耗时最短。低维特征数据集的特征选取实验中，DQPFS 方法用时少于 SVM-RFE 方法。高维特征数据集的特征选取中，DQPFS 方法耗时多于其他两种方法，但其特征选取性能最好。

3 结束语

DQPFS 方法定义了特征间的加权距离，并基于特征间距分析特征所含样本类别信息。将特征的类内距离和类间距离分别作为二次规划算法目标函数的二次项和一次项参数，并通过对二次项和一次项的归一化均衡特征在类内和类间的关系，二次规划算法优化后的最优解被视为衡量特征所含样本类别信息的特征权重。特征选取实验结果表明，DQPFS 方法在 6 个数据集中选取的特征集合对校验样本的识别能力最强，说明该方法所选特征较好地体现了同类样本间的共性和异类样本间的差异性。其在高维和低维数据集中良好的特征选取性能反映了该方法的鲁棒性和有效性。

致谢 感谢 Irene Rodriguez-Lujan 教授给予的帮助，提供了文献[14]的 KQPFS 方法源程序。

参考文献:

[1] Ahmed A-A, Akram A, Rami N-K. Feature subset selection using differential evolution and a wheel based search strategy [J]. Swarm and Evolutionary Computation, 2013(9):15-26.

[2] Howley T, Madden MG, O'Connell ML, et al. The effect of principal component analysis on machine learning accuracy with high dimensional spectral data[J]. Knowledge based Systems, 2006, 19(5): 363-370.

[3] 汪友生,胡百乐,陈建新.基于支持向量机的动脉硬化斑块识别[J].数据采集与处理,2012,27(3):283-286.
Wang Yousheng, Hu Baile, Chen Jianxin, et al. Recognition of atherosclerotic plaque based on support vector machine[J]. Journal of Data Acquisition and Processing ,2012,27(3): 283-286.

[4] 张道强,陈松灿.高维数据降维方法[J].中国计算机学会通讯,2009, 5(8): 15-22.
Zhang Daoqiang, Chen Songcan. Reduction method on high-dimensional data set [J]. Communications of the CCF,2009, 5(8):15-22.

[5] 李颖新, 阮晓钢. 基于基因表达谱的肿瘤亚型识别与分类特征基因选取研究 [J]. 电子学报,2005, 33(4):651-655.
Li Yinxin, Ruan Xiaogang. Cancer cubtype recognition and feature selection with gene expression profiles[J]. Acta Electronic Sinica, 2005, 33(4):651-655.

[6] He X, Cai D, Niyogi P. Laplacian score for feature selection[M]. Advances in Neural Information Processing Systems 18, Cambridge, MA: MIT Press,2005.

[7] Bouaguel W, Mufti GB. An improvement direction for filter selection techniques using information theory measures and quadratic optimization[J]. International Journal of Advanced Research in Artificial Intelligence, 2012,1(5):7-11.

[8] Rodriguez-Lujan I, Huerta R, Elkan C, et al. Quadratic programming feature selection[J]. Journal of Machine Learning Research, 2010, 11:1491-1516.

[9] Ferreira A J, Figueiredo M A T. Efficient feature selection filters for high-dimensional data[J]. Pattern Recognition Letters, 2012,33:1794-1804.

[10] Liu B, Fang B, Liu X W, et al. Large margin subspace learning for feature selection [J]. Pattern Recognition,2013,46(10): 2798-2806.

[11] Guyon I, Weston J, Barnhil S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002,46(1/2/3):389-422.

[12] 高亚东,邓升平. 基于支持向量机的直升机旋翼不平衡故障分类研究[J]. 南京航空航天大学学报, 2011,43(3):435-438.
Gao Yadong, Deng Shengping. Unbalance fault identification of helicopter roter using support vector machine[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2011,43(3):435-438.

[13] Bradley P S, Mangasarian O L, Street W N. Feature selection via mathematical programming[J]. Informs Journal on Computing, 1998,10:209-217.

[14] Rodriguez-Lujan I, Santa Cruz C, Huerta R. On the equivalence of kernel fisher discriminant analysis and kernel quadratic programming feature selection[J]. Patter Recognition Letters, 2011, 32:1567-1571.

[15] 陈宝林. 最优化理论与算法[M]. 北京:清华大学出版社,2005:415-431.
Chen Baolin. Theory and algorithm for optimization[M]. Beijing: Tsinghua University Press, 2005:415-431.

[16] Theodoridis S, Koutroumbas K. Pattern recognition[M]. New York: Academic Press, 1999.

[17] Liu Q J, Zhao Z M, Li Y X, et al. Feature selection based on sensitivity analysis of fuzzy ISODATA[J]. Neurocomputing, 2012(85):29-37.

[18] 边肇祺, 张学工. 模式识别[M]. 2 版. 北京:清华大学出版社, 2001.
Bian Zhaoqi, Zhang Xuegong. Pattern recognition[M]. 2nd Edition. Beijing: Tsinghua University Press, 2001.

[19] 孙即祥. 现代模式识别[M]. Ind Edition. 北京:高等教育出版社, 2008.
Shun Jixiang. Modern pattern recognition[M]. Beijing: High Education Press, 2008.

[20] 王文俊. 基于类别保留投影的基因表达数据特征提取新方法[J].电子学报,2012,40(2):358-364.
Wang Wenjun. New method of feature extraction for gene expression data based on class preserving projection[J]. Acta Electronic Sinica, 2012,40(2):358-364.

[21] University of California. Machine learning Repository [EB/OL]. http://archive.ics.uci.edu/ml/datasets,1989.

[22] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: Class discovery and class prediction by gene ex-

pression monitoring [J]. Science, 1999, 286(5439):531-537.

[23] Shipp M A, Ross K N, Tamayo P, et al. Diffuse large B-cell Lymphoma outcome prediction by gene-expression profiling and supervised machine learning[J]. Nat Med, 2002,8(1):68-74.

[24] Alon U , Barkai N, Notterman D A, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays[J]. PNAS USA,1999,96:6745-6750.

[25] Singh D, Febbo P G, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior[J]. Cancer Cell, 2002, 1(2):203-209.

[26] Provost F, Fawcett T. Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions[C]//Third Inter Conf on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press,1997:43-48.

作者简介:**刘全金**(1971-),男,博士,教授,研究方向:机器学习、图像处理和生物信息学,E-mail: liuquanjing666@126.com;**赵志敏**(1955-),女,教授,研究方向:现代测量与控制技术,智能计算,E-mail: nuaazhzhm@126.com;**李颖新**(1972-),男,博士,高级工程师,研究方向:机器视觉,机器学习与数据挖掘、生物信息学。