基于特征子集区分度与支持向量机的特征选择算法

谢娟英[1)] 谢维信[2)]

[1)](陕西师范大学计算机科学学院 西安 710062)

[2)](深圳大学信息工程学院 ATR 国家重点实验室 广东 深圳 518060)

摘 要 考虑特征之间的相关性对于其类间区分能力的影响,提出了一种新的特征子集区分度衡量准则——DFS(Discernibility of Feature Subsets)准则.该准则考虑特征之间的相关性,通过计算特征子集中全部特征对于分类的联合贡献来判断特征子集的类间辨别能力大小,不再只考虑单个特征对于分类的贡献.结合顺序前向、顺序后向、顺序前向浮动和顺序后向浮动 4 种特征搜索策略,以支持向量机(Support Vector Machines,SVM)为分类工具,引导特征选择过程,得到 4 种基于 DFS 与 SVM 的特征选择算法.其中在顺序前/后向浮动搜索策略中,首先根据 DFS 准则加入/去掉特征到特征子集中,然后在浮动阶段根据所得临时 SVM 分类器的分类性能决定刚加入/去掉特征的去留.UCI 机器学习数据库数据集的对比实验测试表明,提出的 DFS 准则是一种很好的特征子集类间区分能力度量准则;基于 DFS 与 SVM 的特征选择算法实现了有效的特征选择;与其他同类算法相比,基于 DFS 准则与 SVM 的特征选择算法具有非常好的泛化性能,但其所选特征子集的规模不一定是最好的.

关键词 特征选择;支持向量机;相关性;特征子集区分度;特征区分度

中图法分类号 TP18 **DOI 号** 10.3724/SP.J.1016.2014.01704

Several Feature Selection Algorithms Based on the Discernibility of a Feature Subset and Support Vector Machines

XIE Juan-Ying[1)] XIE Wei-Xin[2)]

[1)](*School of Computer Science, Shaanxi Normal University, Xi'an* 710062)

[2)](*School of Information Engineering, National Laboratory of ATR, Shenzhen University, Shenzhen, Guangdong* 518060)

Abstract To consider the influence of the correlation between features on their discernibility between classes, a new criterion was proposed in this paper to evaluate the discernibility of a feature subset. We referred to this criterion as DFS for the short of the discernibility of feature subsets. DFS considers the correlation between features by computing the discernibility of the whole feature subset between classes, so that it can measure the contribution of the whole feature subset to the classification not only that of one feature. Four feature selection algorithms were put forward by combining the DFS, respectively, with the sequential forward search, sequential backward search, sequential forward floating search, and the sequential backward floating search strategies where support vector machines (SVM) were used as a classification tool to guide the feature selection procedure, especially in the sequential forward/backward floating search procedures where a feature was first added to/deleted from the feature subset using the DFS criterion, then it was deleted from/called back during the floating procedure depending on the accuracy of the corresponding temporary SVM classifier went down/up on training subset after adding/deleting the feature to/from the feature subset. Our algorithms were tested on 10 datasets from UCI

收稿日期:2012-09-28;**最终修改稿收到日期**:2014-04-24.**本课题得到国家自然科学基金**(31372250)、**陕西省科技攻关项目**(2013K12-03-24)、**中央高校基本科研业务费专项资金项目**(GK201102007)**资助**.**谢娟英**,**女**,1971 **年生**,**博士**,**副教授**,**中国计算机学会**(CCF)**高级会员**,**主要研究领域为机器学习与数据挖掘**.E-mail:xiejuany@snnu.edu.cn;juanyingxie@gmail.com.**谢维信**,**男**,1941 **年生**,**教授**,**主要研究领域为智能信息处理和模糊信息处理**.

machine learning repository. The experimental results demonstrate that the proposed DFS is a good criterion to evaluate the discernibility of a feature subset. The feature selection algorithms based on DFS and SVM can reduce the dimension of a dataset without compromising its classification capacity, and the generalization of these DFS and SVM based algorithms are much better than that of the available algorithms based on the discernibility of a feature subset. However, the cardinality of the selected feature subsets by them may not be the best ones.

Keywords feature selection; support vector machines; correlation; discernibility of a feature subset; discernibility of a feature

1 引　言

特征选择是模式识别、数据挖掘等领域的重要研究内容，它通过选择原始特征集合中的重要特征构成特征子集，达到降低数据维数，同时保持或提高系统分类性能的目的. 与特征提取不同，特征选择保留的是原始物理特征，因此，可以真正地降低存储需要、测量需求、计算开销等；而特征提取所保留的特征是原始特征的线性组合，降维后的特征与原始每一个特征都相关，因此特征提取保留下的特征没有任何物理意义，这对癌症等疾病诊断研究没有任何意义. 另外，特征提取也不能降低存储需要、测量需求、计算开销等，不具有特征选择的这些优点[1]. 特征选择对于构建一个简洁、高效的分类系统有重要作用. 它不仅可以揭示系统所隐藏的潜在模式和规律，并使分类结果的可视化成为可能[1]. 因此，特征选择，特别是基于最具有泛化能力和最小错分率的学习机——支持向量机(Support Vector Machines, SVM)的特征选择研究备受关注[2-9].

现有特征选择研究主要着眼于选择最优特征子集所需要的两个主要步骤[10]：特征子集搜索策略和特征子集性能评价准则. 经典的特征子集搜索策略包括：顺序前向搜索(Sequential Forward Search, SFS)[11]、顺序后向搜索(Sequential Backward Search, SBS)[12]、顺序前向浮动搜索(Sequential Forward Floating Search, SFFS)[13]、顺序后向浮动搜索(Sequential Backward Floating Search, SBFS)[13] 4 种. 特征子集性能评价准则有独立于分类器的评价方法、以分类器分类准确率评价的方法，以及将两者相结合的评价方法.

特征选择方法依据其与分类器的关系分为 Filter 方法、Wrapper 方法和 Embedded 方法 3 类[14-16]. Filter[14]方法根据每一个特征对分类贡献的大小，定义其重要度，选择重要的特征构成特征子集. 该方法独立于学习过程，时间效率较高，但是该方法需要一个阈值作为特征选择的停止准则. Filter 方法选择的特征子集的分类性能不仅与特征重要性计算方法，即特征排序准则有关；而且与特征搜索策略，以及特征选择过程的停止准则密切相关.

Wrapper[15]方法依赖于学习过程，将训练样本分成训练子集和测试子集两部分. Wrapper 方法中的学习算法完全是一个“黑匣子”，仅以每一组特征子集训练所得分类器的分类准确率作为该组特征分类性能的度量. 为选择出性能最好的特征子集，Wrapper 算法需要的计算量巨大；而且该方法所选择的特征子集依赖于具体学习机；容易产生“过适应”问题，推广性能较差. 另外，确定搜索策略以搜索所有可能的特征组合，评价学习机的性能以引导或停止搜索，以及选择具体的学习算法是 Wrapper 方法的关键.

Embedded[16]方法将特征选择集成在学习机训练过程中，通过优化一个目标函数在训练分类器的过程中实现特征选择. 该方法不用将训练数据集分成训练集和测试集两部分，避免了为评估每一个特征子集对学习机所进行的从头开始的训练，可以快速地得到最佳特征子集，是一种高效的特征选择方法，但是构造一个合适的函数优化模型是该方法的难点.

Filter 算法和 Wrapper 算法相结合的混合特征选择方法集成了 Filter 方法的高效与 Wrapper 方法的高准确率，可得到更优的特征子集[17-19]，是目前特征选择方法研究的一个新趋势. 然而，现有特征选择研究多集中于考虑单个特征对于分类的贡献，以此来评价特征的重要性，从而依次选择重要的特征构成特征子集. 这样进行特征选择忽略了特征之间的

相关性对于分类的影响，或者对特征之间的相关性考虑不足，特征之间的联合作用对于分类的贡献经常被忽略.

Hall的CFS(Correlation based Feature Selector，CFS)[20]特征选择方法基于特征之间的相关性考虑，通过计算特征之间的相关性，以及特征与类标的相关性，来实现特征选择，从而使得被选特征子集的特征之间尽可能不相关，而与类标高度相关. 但是HALL的CFS计算特征相关性的方法只适用于离散型数据，非离散的数据需要进行离散化预处理；另外，特征之间的相关性只注重了特征两两之间的相关性，对于多个特征对于分类的联合贡献考虑不足. 本文基于特征之间相关性的考虑，提出DFS(Discernibility of Feature Subsets，DFS)特征子集区分度衡量准则，通过计算特征子集中特征的联合F-score值来判断特征子集的类别间区分能力大小，以此作为特征子集对分类贡献大小的度量，并结合SFS、SBS、SFFS和SFBS 4种特征搜索策略，以SVM为分类工具引导特征选择过程，实现特征选择. UCI机器学习数据库数据集的实验测试表明，本文提出的DFS准则是一种有效的特征子集类间区分能力度量准则；基于DFS与SVM的特征算法实现了有效的特征选择，具有很好的泛化性能；但是所选择的特征子集的规模不一定是最好的.

2 特征子集评价准则

现有特征选择研究中特征重要性的评价多是计算单个特征的类间区分能力，选择对分类贡献大的特征构成特征子集，而忽略了特征之间的相关性对于特征类间区分能力大小的影响，或者对特征之间的相关性考虑不足. Hall[20]通过计算特征之间的相关性、特征与类标的相关性提出CFS特征子集评价准则，CFS计算整个特征子集的类间区分能力实现特征选择，使得被选特征子集中的特征之间尽可能不相关，而与类标高度相关. 但CFS[20]只适用于离散型数据；另外，CFS更注重了特征两两之间的相关性，对于多个特征对于分类的联合贡献考虑不足. 为此，本文首先将Pearson相关系数引入CFS，以计算特征两两之间的相关性，使得CFS可应用于实值特征；然后将Pearson相关系数所表达的正、负相关性进行统一，不区分正、负相关，只考虑相关，从而得到CFSPabs(Correlation based Feature Selector based on the absolute of Pearson's correlation coefficient，CFSPabs)特征子集评价准则. 其次，本文在文献[19]对单个特征的类间区分能力，即对分类的贡献的研究基础上，考虑多个特征对于分类的联合贡献，提出DFS特征子集评价准则，并结合4种特征搜索策略SFS、SBS、SFFS和SFBS，提出4种基于DFS特征子集评价准则的特征选择算法.

2.1 CFS特征子集评价准则

CFS特征子集评价准则[20]的定义如式(1)所示：

$$Ms=\frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\,\overline{r_{ff}}}} \tag{1}$$

式(1)中的Ms度量了包含k个特征的特征子集S的类别辨识能力，$\overline{r_{cf}}$表示特征$f(f\in S)$与类别c的相关系数的均值，$\overline{r_{ff}}$是特征之间的相关系数均值. 式(1)的分子表达了特征子集S的类预测能力；分母表示了特征子集S中特征的冗余程度. 因此分子越大表示特征子集S的类预测能力越强，分母越小表示该特征子集的冗余性越小. 特征选择，就是选择一组特征构成特征子集，该子集与类别高度相关，但是子集中的特征之间高度不相关[12]. 由此可见Ms的值越大，当前特征子集S对于分类的贡献越大，是优良的特征子集. 我们使用Pearson相关系数计算相应特征之间的相关性，以便CFS准则可应用于任意类型的特征值. Pearson相关系数的计算公式如式(2)所示.

$$r=\frac{\sum \boldsymbol{XY}-\frac{\sum \boldsymbol{X}\sum \boldsymbol{Y}}{N}}{\sqrt{\left(\sum \boldsymbol{X}^2-\frac{\left(\sum \boldsymbol{X}\right)^2}{N}\right)\left(\sum \boldsymbol{Y}^2-\frac{\left(\sum \boldsymbol{Y}\right)^2}{N}\right)}} \tag{2}$$

其中，$\boldsymbol{X}$，$\boldsymbol{Y}$表示待求相关系数的两个向量，可以是两列特征向量，或者一列特征向量与一列类标向量，N是样本个数.

2.2 CFSPabs特征子集评价准则

由于Pearson相关系数可能为正值，也可能为负值，也即待判断相关程度的两向量可能正相关，也可能负相关. 而无论正相关还是负相关，相关系数的绝对值越大，也即相关系数越接近+1或−1，则相关性越强；相关系数越接近于0，则相关度越弱. 因此，我们将式(1)所示CFS准则中计算相关系数的Pearson相关系数进行改进，取其绝对值，得到特征子集评价准则CFSPabs. CFSPabs准则中的相关系数计算公式如式(3)所示.

$$r'=\frac{\left|\sum \boldsymbol{XY}-\frac{\sum \boldsymbol{X}\sum \boldsymbol{Y}}{N}\right|}{\sqrt{\left(\sum \boldsymbol{X}^2-\frac{\left(\sum \boldsymbol{X}\right)^2}{N}\right)\left(\sum \boldsymbol{Y}^2-\frac{\left(\sum \boldsymbol{Y}\right)^2}{N}\right)}} \tag{3}$$

2.3　DFS 特征子集评价准则

设 m 维实空间 $\boldsymbol{R}^m$ 中的两类分类问题，训练集样本的规模为 n，正类和负类的样本数分别为 n_+ 和 n_-. 即训练集是 $\{(\boldsymbol{x}_k, y_k) \mid \boldsymbol{x}_k \in \boldsymbol{R}^m, m>0, y_k \in \{1,-1\}, k=1,\cdots,n\}$，$\|\{y_k \mid y_k=+1, k=1,\cdots,n\}\|=n_+$，$\|\{y_k \mid y_k=-1, k=1,\cdots,n\}\|=n_-$. 则含有 $i(i=1,2,\cdots,m)$ 个特征的特征子集的区分度 DFS 的定义如式(4)所示.

$$DFS_i=\frac{\sum_{j=1}^{i}\left((\bar{x}_j^{(+)}-\bar{x}_j)^2+(\bar{x}_j^{(-)}-\bar{x}_j)^2\right)}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}\left(\sum_{j=1}^{i}(x_{k,j}^{(+)}-\bar{x}_j^{(+)})^2\right)+\frac{1}{n_--1}\sum_{k=1}^{n_-}\left(\sum_{j=1}^{i}(x_{k,j}^{(-)}-\bar{x}_j^{(-)})^2\right)} \tag{4}$$

其中，$\bar{x}_j$，$\bar{x}_j^{(+)}$，$\bar{x}_j^{(-)}$ 分别为第 j 个特征在整个数据集上的均值，在正类数据集上的均值和在负类数据集上的均值；$x_{k,j}^{(+)}$ 为正类第 k 个样本点的第 j 个特征的值；$x_{k,j}^{(-)}$ 为第 k 个负类样本的第 j 个特征的值. 式(4)可以简记为式(5).

$$DFS_i=\frac{\|\bar{\boldsymbol{x}}^{(+)}-\bar{\boldsymbol{x}}\|^2+\|\bar{\boldsymbol{x}}^{(-)}-\bar{\boldsymbol{x}}\|^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}\|\boldsymbol{x}_k^{(+)}-\bar{\boldsymbol{x}}^{(+)}\|^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}\|\boldsymbol{x}_k^{(-)}-\bar{\boldsymbol{x}}^{(-)}\|^2} \tag{5}$$

其中，$\|\mathbf{X}-\mathbf{Y}\|$ 表示向量 $\boldsymbol{X}$，$\boldsymbol{Y}$ 之间的距离；$\bar{\boldsymbol{x}}$，$\bar{\boldsymbol{x}}^{(+)}$，$\bar{\boldsymbol{x}}^{(-)}$ 分别为包含 i 个特征的特征子集在整个数据集上的均值向量，在正类数据集上的均值向量和在负类数据集上的均值向量；$\boldsymbol{x}_k^{(+)}$ 为正类第 k 个样本点对应于当前 i 个特征的特征值向量；$\boldsymbol{x}_k^{(-)}$ 为第 k 个负类样本对应当前 i 个特征的特征值向量.

分析式(4)和(5)可知，分子表示正类、负类对应当前含有 i 个特征的特征子集的均值向量与整个样本集对应当前 i 个特征的特征子集的均值向量的距离平方之和；分母表示正类、负类对应当前 i 个特征的特征子集的方差之和；分子值越大表示对应当前 i 个特征的特征子集的类间越疏；分母值越小表示对应当前 i 个特征的特征子集的类内越聚. 因此 DFS_i 值越大，表明包含当前 i 个特征的特征子集的类间区分能力越强，也即类别辨识能力越强.

对于 $l(l\geq 2)$ 类分类问题，设训练样本集规模为 n，样本空间维数为 m. 即训练样本集为 $\{(\boldsymbol{x}_k, y_k) \mid \boldsymbol{x}_k \in \boldsymbol{R}^m$（$m$ 维实空间），$m>0, y_k \in \{1,\cdots,l\}, l\geq 2, k=1,\cdots,n\}$. 其中，第 j 类的样本数为 n_j，即 $\|y_k \mid y_k=j, k=1,\cdots,n\|=n_j, j=1,\cdots,l$，则含有 $i(i=1,\cdots,m)$ 个特征的特征子集的区分度 DFS_i 定义为式(6).

$$DFS_i=\frac{\sum_{j=1}^{l}\|\bar{\boldsymbol{x}}^{(j)}-\bar{\boldsymbol{x}}\|^2}{\sum_{j=1}^{j}\frac{1}{n_j-1}\sum_{k=1}^{n_j}\|\boldsymbol{x}_k^{(j)}-\bar{\boldsymbol{x}}^{(j)}\|^2} \tag{6}$$

其中 $\bar{\boldsymbol{x}}$，$\bar{\boldsymbol{x}}^{(j)}$ 分别为包含当前 i 个特征的特征子集在整个数据集上的均值向量，在第 j 类数据集上的均值向量；$\boldsymbol{x}_k^{(j)}$ 为第 j 类中第 k 个样本对应当前 i 个特征的特征值向量. 式(6)的分子表示 l 个类别中各类别对应包含当前 i 个特征的特征子集的样本中心向量与整个样本集对应当前 i 个特征的中心向量的距离平方和，其值越大，类间越疏. 式(6)的分母表示各个类别对应包含当前 i 个特征的特征子集的类内方差. 方差越小，类内越聚. 因此，式(6)定义的 DFS_i 表示了当前 i 个特征的特征子集的类间距离和与类内方差之比，其值越大表明包含当前 i 个特征的特征子集的类别辨识力越强.

当 $i=1$ 时，DFS_i 蜕化为单个特征的类间区分能力度量准则——改进的 F-score 准则[19]，如式(7)所示.

$$F_i=\frac{\sum_{j=1}^{l}(\bar{x}_i^{(j)}-\bar{x}_i)^2}{\sum_{j=1}^{l}\frac{1}{n_j-1}\sum_{k=1}^{n_j}(x_{k,i}^{(j)}-\bar{x}_i^{(j)})^2} \tag{7}$$

式(7)中 $\bar{x}_i$，$\bar{x}_i^{(j)}$ 分别为第 i 个特征在整个数据集上的均值，在第 j 类数据集上的均值；$x_{k,i}^{(j)}$ 为第 j 类第 k 个样本点的第 i 个特征的值. 式(7)分子表示 l 个类中第 i 个特征的各类中心到整个样本集中心的距离平方和，其值越大，类间越疏. 式(7)的分母表示各个类第 i 个特征的类内方差. 方差越小，类内越聚. 因此，F_i 近似表示了第 i 个特征的类间距离与类内方差之比，其值越大表明第 i 个特征的辨识力越强.

3　基于 DFS 与 SVM 的特征选择算法

将本文提出的特征子集区分度衡量准则 DFS 结合 SFS、SBS、SFFS 和 SFBS 特征搜索策略，以 SVM 为分类工具，提出 4 种不同的特征选择算法.

算法 1. 基于 DFS 与 SVM 的顺序前向混合特征选择算法.

该算法采用 SFS 搜索策略,以 DFS 度量相应特征子集的类间区分能力,以 SVM 分类模型的分类正确率作为最终依据选择相应特征子集. 特征选择从空集开始,第一次加入类间区分能力最强的一个特征,然后依次迭代加入与已选择特征组合构成类间区分能力最大的特征子集对应的那个特征. 该过程一直进行,直到所有特征都被加入. 最后选择训练集分类正确率最高时对应的特征子集为被选择特征子集. 算法的伪代码描述如下.

输入:当前的训练集和测试集.
输出:特征子集 C
设 $S=\{f_i \mid i=1,\cdots,m\}$ 为全部特征构成的集合,C 为被选择特征构成的子集,初始为空集,即 $C=\varnothing$;
根据式(7)计算训练集上每个特征的类间区分能力;
选择最重要的特征 $f_{\max}=\max\{F_i, i=1,\cdots,m\}$;
令 $S=S-\{f_{\max}\}$; $C=C\cup\{f_{\max}\}$;
使用 C 中特征训练 SVM,得到一个 SVM 分类模型;
以该模型对训练集、测试集进行分类,记录相应的分类正确率;
WHILE $S\neq\varnothing$ DO
BEGIN
  $maxDFS=min$; $maxtempC=\varnothing$;
  FOR each $f\in S$ DO
  BEGIN
    $tempC=C\cup\{f\}$;
    根据式(6)计算 $DFS(tempC)$;
    IF $DFS(tempC)>maxDFS$ THEN
    BEGIN
      $maxDFS=DFS(tempC)$;
      $maxtempC=tempC$;
      $selectedfeature=f$;
    END //end of if
  END //end of for
  $C=maxtempC$;
  $S=S-\{selectedfeature\}$;
  使用 C 中特征训练 SVM,得到一个 SVM 分类模型;
  以该模型对训练集、测试集进行分类,并记录相应的分类正确率;
END //end of while

算法 2. 基于 DFS 与 SVM 的顺序后向混合特征选择算法.

该算法利用 SBS 搜索策略,以 DFS 度量相应特征子集的类间区分能力,SVM 的分类准确率引导特征选择过程. 算法从特征全集开始,每次剔除当前余下特征中最差的一个,即剔除该特征后的剩余特征构成最具有类间区分能力的特征子集. 以训练集分类正确率不再上升时对应的规模最小的特征子集为被选择特征子集. 算法的伪代码描述如下.

输入:当前的训练集 $Xtrain$ 和测试集 $Xtest$
输出:特征子集 S
设 $S=\{f_i \mid i=1,2,\cdots,m\}$ 为全部特征构成的集合;
WHILE $S\neq\varnothing$ DO
BEGIN
  以 S 中特征训练 $Xtrain$,得到一个 SVM 分类模型;
  以该模型分类 $Xtrain$ 和 $Xtest$,并记录分类正确率;
  $maxDFS=min$;
  $featuresubset=S$;
  FOR each $f\in S$ DO
  BEGIN
    $tempsubset=S-\{f\}$;
    根据式(6)计算 $DFS(tempsubset)$;
    $tempDFS=DFS(tempsubset)$;
    IF $tempDFS>maxDFS$ THEN
    BEGIN
      $deletefeature=f$;
      $featuresubset=tempsubset$;
      $maxDFS=temDFS$;
    END //end of if
  END //end of for
  $S=S-\{deletefeature\}$;
END//end of while

算法 3. 基于 DFS 与 SVM 的顺序前向浮动混合特征选择算法.

该算法将 DFS 特征评价准则与顺序前向浮动搜索策略 SFFS 结合,以 SVM 为分类工具,首先加入最具有类间区分能力的一个特征,然后每次迭代加入与已选择特征组合最具有类间区分能力的相应特征,之后浮动部分依据加入特征之后的特征子集对应 SVM 分类器的分类正确率判定刚刚加入的特征是否保留. 若当前特征子集训练所得 SVM 分类模型的训练分类正确率上升,则保留刚加入的特征,否则删除刚刚加入的特征. 该过程一直进行,直到所有特征都被测试过. 最后留在特征子集中的特征构成最佳被选择特征子集. 算法伪代码描述如下.

输入:当前训练集和测试集
输出:特征子集 C
设 $S=\{f_i \mid i=1,2,\cdots,m\}$ 为全部特征构成的集合,C 为被选择特征构成的子集,初始为空集,即 $C=\varnothing$;
根据式(7)在训练集上计算每个特征的区分能力;
选择最重要的特征 $f_{\max}=\max\{F_i, i=1,2,\cdots,m\}$;
令 $S=S-\{f_{\max}\}$;
令 $C=C\cup\{f_{\max}\}$;
使用 C 中特征训练 SVM,得到一个 SVM 分类模型;

记录该模型对训练集、测试集的分类正确率 $acctrain$ 和 $acctest$；
WHILE $S\neq\varnothing$ DO
BEGIN
　$maxDFS=min$；
　$maxtempC=C$；
　FOR each $f\in S$ DO
　BEGIN
　　$tempC=C\cup\{f\}$；
　　根据式(6)计算 $tempC$ 的 DFS 值 $DFS(tempC)$；
　　IF $DFS(tempC)>maxDFS$ THEN
　　BEGIN
　　　$maxDFS=DFS(tempC)$；
　　　$maxtempC=C\cup\{f\}$；
　　　$selectedfeature=f$；
　　END //end of if
　END//end of for
　$C=maxtempC$；
　$S=S-\{selectedfeature\}$；
　$preacctrain=acctrain$；
　$preacctest=acctest$；
　使用 C 中特征训练 SVM，得到一个 SVM 分类模型；以该模型对训练集、测试集进行分类，记录相应的分类正确率 $acctrain$ 和 $acctest$；
　IF $acctrain\leq preacctrain$ THEN
　BEGIN
　　$C=C-\{selectedfeature\}$；
　END//end of if
END //end of while

算法 4.　基于 DFS 与 SVM 的顺序后向浮动混合特征选择算法.

该算法将 DFS 特征子集区分度评价准则与 SVM 结合，以顺序后向浮动搜索策略 SBFS 进行特征搜索，特征选择从全集开始，根据 DFS 值剔除当前余下特征中最差的一个特征，即剔除该特征使得剩余特征构成的特征子集对应的 DFS 值最大. 浮动部分以被选择特征子集的 SVM 分类模型的分类正确率判断刚刚剔除的特征是否需要收回，若剔除该最差特征后导致训练集的分类正确率下降，则将刚剔除的最差特征召回；否则剔除. 该过程一直迭代，直到所有特征都被测试过，最后留下的特征构成被选特征子集. 算法的伪代码描述如下.

输入：当前训练集 $Xtrain$ 与测试集 $Xtest$
输出：特征子集 C
设 $S=\{f_i\mid i=1,2,\cdots,m\}$ 为包含全部特征的集合，$C=\{f_i\mid i=1,2,\cdots,m\}$ 为被选择特征构成的子集；
WHILE $S\neq\varnothing$ DO
BEGIN
　以 C 中特征在 $Xtrain$ 训练得到一个 SVM 分类模型；以该模型对 $Xtrain$ 和 $Xtest$ 进行分类，记录相应的分类正确率 $acctrain$ 和 $acctest$；
　$maxDFS=min$；$featuresubset=S$；
　FOR each $f\in S$ DO
　BEGIN
　　$tempsubset=S-\{f\}$；
　　根据式(6)计算 $tempsubset$ 的 $DFS(tempsubset)$；
　　$tempDFS=DFS(tempsubset)$；
　　IF $tempDFS>maxDFS$ THEN
　　BEGIN
　　　$deletefeature=f$；
　　　$featuresubset=tempsubset$；
　　　$maxDFS=tempDFS$；
　　END //end of if
　END //end of for
　$S=S-\{deletefeature\}$；
　$preacctrain=acctrain$；
　$preacctest=acctest$；
　使用 S 中特征训练 SVM，得到一个 SVM 分类模型；以该模型对 $Xtarin$、$Xtest$ 进行分类，记录相应的分类正确率 $acctrain$ 和 $acctest$；
　IF $acctrain\geq preacctrain$ THEN
　BEGIN
　　$C=C-\{deletefeature\}$；
　END//end of if
END//end of while

4　实验结果与分析

实验采用 UCI 机器学习数据库[21] 的 10 个数据集① iris，dermatology，glass，handwrite，ionosphere，WDBC(Wisconsin Diagnostic Breast Cancer)，WPBC(Wisconsin Prognostic Breast Cancer)，wine，thyroid-disease 和 heart disease. 其中，dermatology 数据集，删去了 8 个含有缺失数据的样本；glass 数据集分成了 window glass 和 non-window glass 两类；handwrite 数据集只选择了 semeion handwritten digit 数据集的前两类；WPBC 数据集删去了 4 个含有缺失数据的样本；thyroid-disease 数据集是其中的 new-thyroid，即 thyroid gland data 数据集；heart disease 数据集使用了其中的 processed cleveland 数据集，并删去了 6 个含有缺失数据的样本. 数据集详细信息如表 1 描述. 实验环境为 Dell Vostro

① http://www.ics.uci.edu/~mlearn/MLRepository.html

1450 笔记本电脑，Intel Core i5-2450 2.50GHz CPU，4 GB 内存，Win7 64 位操作系统，Matlab 应用软件.

表 1 实验所用 UCI 数据集描述

数据集	样本个数	特征数	类别数
iris	150	4	3
dermatology	358	34	6
glass	214	9	2
handwrite	323	255	2
ionosphere	351	34	2
WDBC	569	30	2
WPBC	194	33	2
wine	178	13	3
thyroid-disease	215	5	3
heart disease	297	13	5

为得到具有统计意义的实验结果，采用 5 折交叉验证实验. 同时，为了得到随机的实验数据，采用将样本顺序随机打乱，每一类样本依次逐个加入到 5 个初始为空的样本集合，直到这一类的每一个样本都被加入，实现将样本随机均匀划分为 5 份的目的. 以 1 份样本为测试样本集，其余 4 份为训练样本集，并以每 1 份都做过测试集结束 5 折交叉验证实验. 样本随机打乱的方法是：随机生成一个 5000×2 的 2 维数组，数组每一元素的值在 1～数据集规模之间；交换数组每一行两个元素值对应的两个样本. 实验采用林智仁教授等开发的 SVM 工具箱[22]，核函数采用 RBF(Radial Basis Function)核函数[23]. 为了更客观地比较基于不同特征子集评价准则的特征选择算法的性能，对各特征子集评价准则进行性能比较，本文各算法的核函数参数均采用默认值.

为测试提出的 DFS 特征子集评价准则的有效性，将其与 CFS 以及提出的 CFSPabs 特征子集评价准则进行实验比较. 3 种分别基于 DFS、CFS 和 CFSPabs 不同特征子集重要性(区分度)评价准则与 SVM 的特征选择算法的实验结果如表 2～表 5 所示. 表中数值为各算法 5 折交叉验证实验对应实验结果的平均值，加重和加下划线的被选择特征数、测试集分类正确率、运行时间(以 s 为单位)分别表示最小的特征子集规模、最高的分类正确率、最快的运行速度.

为了更清楚地展示特征依次被选入(或依次被剔除)时，相应 SVM 的正确率变化情况，以说明选择到多少个特征时才是最好的，图 1、图 2 分别展示了基于 SFS、SBS 搜索策略，并分别以 DFS、CFS 和 CFSPabs 为特征子集重要性(区分度)评价准则，以 SVM 为分类工具的相应特征选择算法的训练正确率和测试正确率变化曲线. 其中的训练正确率与测试正确率变化曲线是 5 折交叉验证实验的平均正确率曲线.

表 2 DFS、CFS 和 CFSPabs 特征子集评价准则的顺序前向特征选择算法的 5 折交叉验证实验结果

数据集	原始特征数	被选择特征数			测试集分类正确率/%			运行时间/s		
		DFS	CFS	CFSPabs	DFS	CFS	CFSPabs	DFS	CFS	CFSPabs
iris	4	3.8	3.8	3.8	**96.66667**	96	96	0.010858	0.005068	**0.004732**
dermatology	34	33.2	31.4	**24.2**	**95.52295**	92.51244	91.67825	**1.059512**	1.525841	1.553808
glass	9	8.8	7.8	**7.2**	93.45364	**93.90818**	93.44257	0.045224	0.024390	**0.022436**
handwrite	255	37.2	139.4	**13.6**	**98.77841**	98.75947	98.45178	**41.26044**	1965.369	1922.932
ionosphere	34	20.8	18.8	18.8	92.29376	92.00805	**93.15091**	**0.740245**	0.979745	0.945958
WDBC	30	28.0	**14.0**	15.0	**64.68026**	63.09042	63.09042	**1.316294**	1.584882	1.532996
WPBC	33	29.4	**5.6**	6.2	76.29555	**77.33536**	75.78273	**0.466218**	0.794759	0.797613
wine	13	11.8	12.0	**7.0**	47.76405	**53.91849**	41.61504	0.099037	**0.063909**	0.081331
thyroid-disease	5	4.8	4.2	4.2	**77.67442**	76.27907	76.27907	0.031925	**0.019746**	0.021291
heart disease	13	13.0	**11.8**	12.4	53.88194	53.88194	53.88194	0.199113	0.1853434	**0.161913**

表 3 DFS、CFS 和 CFSPabs 特征子集评价准则的顺序后向特征选择算法的 5 折交叉验证实验结果

数据集	原始特征数	被选择特征数			测试集分类正确率/%			运行时间/s		
		DFS	CFS	CFSPabs	DFS	CFS	CFSPabs	DFS	CFS	CFSPabs
iris	4	2.8	2.8	2.8	**96.66667**	96	96	0.010435	0.004894	**0.004679**
dermatology	34	32.2	**31.2**	31.8	95.52295	93.61604	**96.09438**	**1.171889**	3.029078	2.533540
glass	9	7.8	**3.6**	7.2	93.45364	**94.3733**	93.91876	0.050377	0.030151	**0.025696**
handwrite	255	**36.0**	127.8	110.8	**98.77841**	98.45659	97.23499	**52.50070**	5800.821	5585.649
ionosphere	34	19.8	**19.2**	21.2	92.29376	**93.14688**	91.73038	**0.796304**	2.116588	1.974555
WDBC	30	27.0	**11.4**	24.4	64.68026	63.61216	**64.69257**	**1.325363**	2.398988	1.910676
WPBC	33	28.4	**8.2**	10.8	76.29555	76.29555	75.79555	**0.498147**	1.687616	1.542187
wine	13	10.8	**7.6**	12.0	**47.76405**	43.91756	43.91849	0.105099	0.103214	**0.065303**
thyroid-disease	5	3.8	4.0	**3.2**	**77.67442**	75.34884	76.27907	0.025749	**0.016401**	0.018775
heart disease	13	12.0	**10.6**	11.2	53.88194	53.88194	53.88194	0.215978	0.217254	**0.173576**

表 4　DFS、CFS 和 CFSPabs 特征子集评价准则的顺序前向浮动特征选择算法的 5 折交叉验证实验结果

数据集	原始特征数	被选择特征数			测试集			运行时间/s		
		DFS	CFS	CFSPabs	DFS	CFS	CFSPabs	DFS	CFS	CFSPabs
iris	4	**3.2**	3.4	3.4	**97.33333**	96.66667	96.66667	0.009394	**0.004198**	0.004282
dermatology	34	14.2	12.0	**11.0**	94.73595	92.01030	**96.66538**	0.984429	0.653162	**0.505920**
glass	9	4.6	5.0	**3.8**	92.03614	**92.96637**	91.54938	0.043883	0.020172	**0.019138**
handwrite	255	9.0	9.6	**5.4**	**98.77375**	95.67716	95.35999	28.73756	16.17666	**9.902686**
ionosphere	34	9.6	9.4	**8.2**	92.58350	91.44064	**93.44064**	0.653062	0.352574	**0.334499**
WDBC	30	8.8	6.2	6.2	69.18045	74.96114	**74.96114**	1.103749	**0.723270**	0.737452
WPBC	33	8.8	4.4	**3.8**	76.29555	**77.33536**	75.78273	0.429413	0.229571	**0.212909**
wine	13	7.4	6.8	**5.2**	47.87625	**50.50392**	46.07793	0.097809	**0.051235**	0.063475
thyroid-disease	5	4.4	4.2	4.2	**77.20930**	76.27907	76.27907	0.024129	0.017887	**0.016580**
heart disease	13	10.0	**8.8**	9.0	53.88194	53.88194	53.88194	0.192441	**0.150492**	0.155060

表 5　DFS、CFS 和 CFSPabs 特征子集评价准则的顺序后向浮动特征选择算法的 5 折交叉验证实验结果

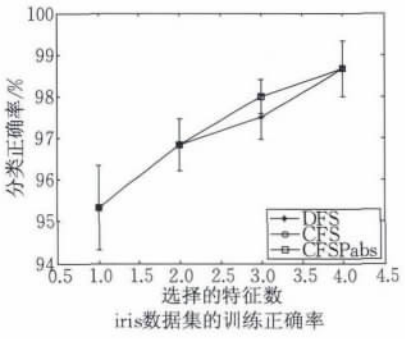
数据集	原始特征数	被选择特征数			测试集分类正确率/%			运行时间/s		
		DFS	CFS	CFSPabs	DFS	CFS	CFSPabs	DFS	CFS	CFSPabs
iris	4	2.8	2.6	2.6	**96.66667**	95.33333	95.33333	0.010638	0.004828	**0.004016**
dermatology	34	**13.8**	16.4	14.4	**92.71300**	88.29022	90.75268	**1.392221**	3.907337	3.215710
glass	9	3.4	3.6	**2.4**	93.02074	93.91926	**94.84949**	0.054530	0.028757	**0.026175**
handwrite	255	**9.6**	30.6	21.0	**98.46125**	97.83625	98.4375	**52.18866**	5838.756	5656.66
ionosphere	34	15.2	15.4	15.2	92.01207	**94.00805**	94.00402	**0.836711**	2.601833	2.362460
WDBC	30	4.0	3.0	**2.6**	**62.91958**	62.74259	**62.91958**	**1.909339**	2.521980	2.434949
WPBC	33	4.2	**1.8**	2.8	76.29555	76.29555	**76.79555**	**0.541972**	1.636516	1.524585
wine	13	2.0	**1.8**	2.8	**46.04704**	42.75789	41.61504	0.121354	0.105184	**0.102492**
thyroid-disease	5	**2.2**	3.2	2.4	**78.60465**	**78.60465**	75.34884	0.030379	**0.017902**	0.018788
heart disease	13	**3.4**	4.6	3.6	53.54861	**53.88194**	53.54861	0.305235	0.262979	**0.244258**

表 2 基于 SFS 搜索序策略的 5 折交叉验证实验结果显示，提出的 DFS 特征子集评价准则在 iris，dermatology，handwrite，WDBC，thyroid-disease 和 heart disease 6 个数据集上选择的特征子集的分类正确率高于等于 CFS、CFSPabs 准则选择的特征子集的分类正确率；CFS 准则在 glass，WPBC 和 wine 3 个数据集上的分类正确率超过 DFS 和 CFSPabs 准则；CFSPabs 准则只在 ionosphere 一个数据集上的分类正确率超过 DFS 和 CFS 准则. 特征子集规模比较显示：CFSPabs 准则选择的特征子集规模优于 DFS 和 CFS 准则. 其中，在 dermatology，glass，handwrite 和 wine 4 个数聚集上优于 DFS 和 CFS 准则，在 ionosphere 和 thyroid-disease 数据集上等于 CFSPabs 准则，优于提出 DFS 准则选择的特征子集规模；DFS 准则只在 iris 数据集上与其他两个准则持平，在其他 9 个数据集上都不如其他两个准则；CFS 准则在 WDBC，WPBC 和 heart disease 共 3 个数据集上优于其他两个准则. 运行时间比较显示：提出的 DFS 特征子集评价准则优于其他两个特征子集评价准则，特别体现在 handwrite 这类较高维数据集的特征选择上，基于 DFS 特征子集评价准则的特征选择算法明显优于其他两个准则. 10 个数据集的运行时间显示，DFS 准则在 5 个数据集上运行时间优于 CFS 和 CFSPabs 准则；而 CFSpabs 准则在 3 个数据集上较优；CFS 准则在 2 个数据集上较优.

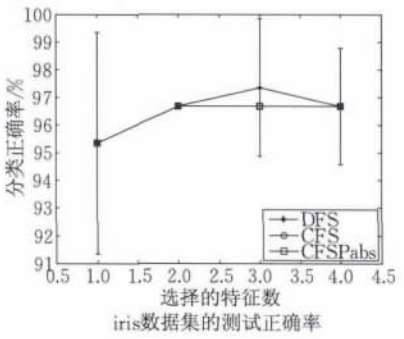
以上对于表 2 实验结果的分析得出：提出的 DFS 特征子集评价准则无论运行时间还是所选择特征子集的分类性能都较优，但所选择的特征子集规模不是最优的. 然而从 handwrite 这一较高维数据集的实验结果来看，提出的 DFS 特征子集评价准则优于 Hall 提出的 CFS 特征子集评价准则.

表 3 基于 SBS 搜索策略的实验结果显示，在所选特征子集的分类正确率、相应特征选择算法的运行时间两个指标的平均值比较上，提出的 DFS 特征子集评价准则优于 CFS 和 CFSPabs 准则；另外，CFS 和 CFSPabs 准则相比，后者优于前者. 从选择的特征子集规模来看，虽然提出的 DFS 特征子集评价准则只在 handwrite 数据集上较优，但是 3 个特征子集评价准则的比较可见，DFS 准则在 handwrite 数据集上选择的特征数远远少于 CFS 和 CFSPabs 准则选择的特征数. 虽然 CFS 准则在 7 个数据集上选择的特征子集规模优于其他两准则，但是特征子集的规模差别不像 DFS 与 CFS、CFSPabs 在 handwrite 数据集选择的特征子集规模差别那么突出.

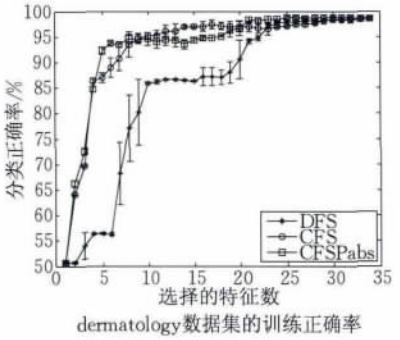
以上表 2 和表 3 实验结果的分析得出：提出的



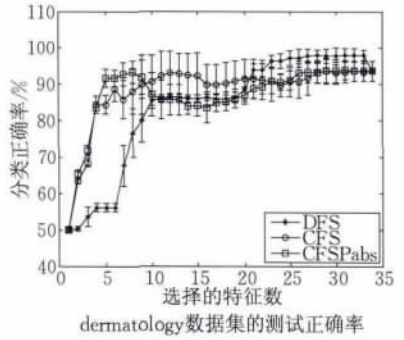
iris数据集的训练正确率



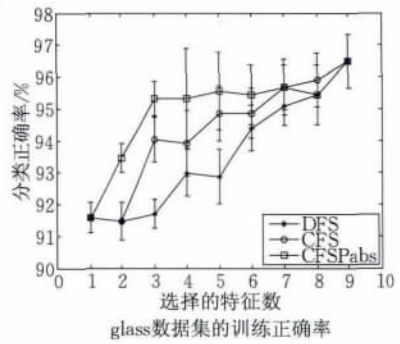
iris数据集的测试正确率



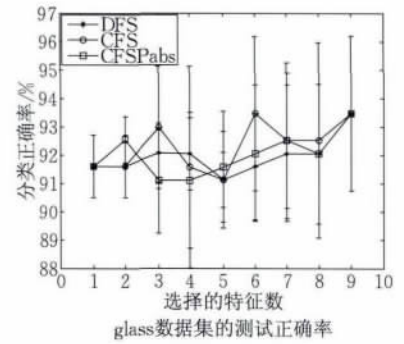
dermatology数据集的训练正确率



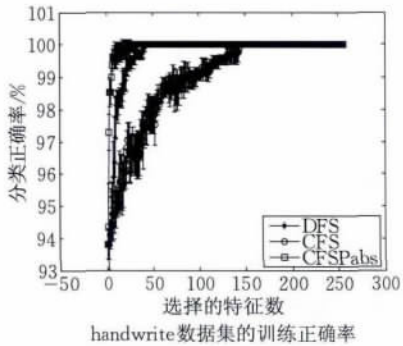
dermatology数据集的测试正确率



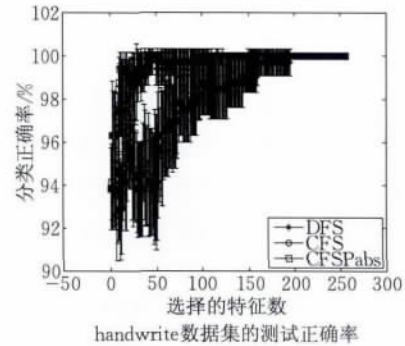
glass数据集的训练正确率



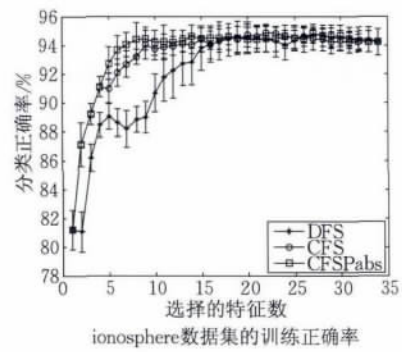
glass数据集的测试正确率



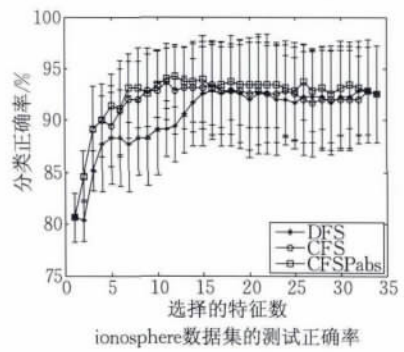
handwrite数据集的训练正确率



handwrite数据集的测试正确率



ionosphere数据集的训练正确率



ionosphere数据集的测试正确率

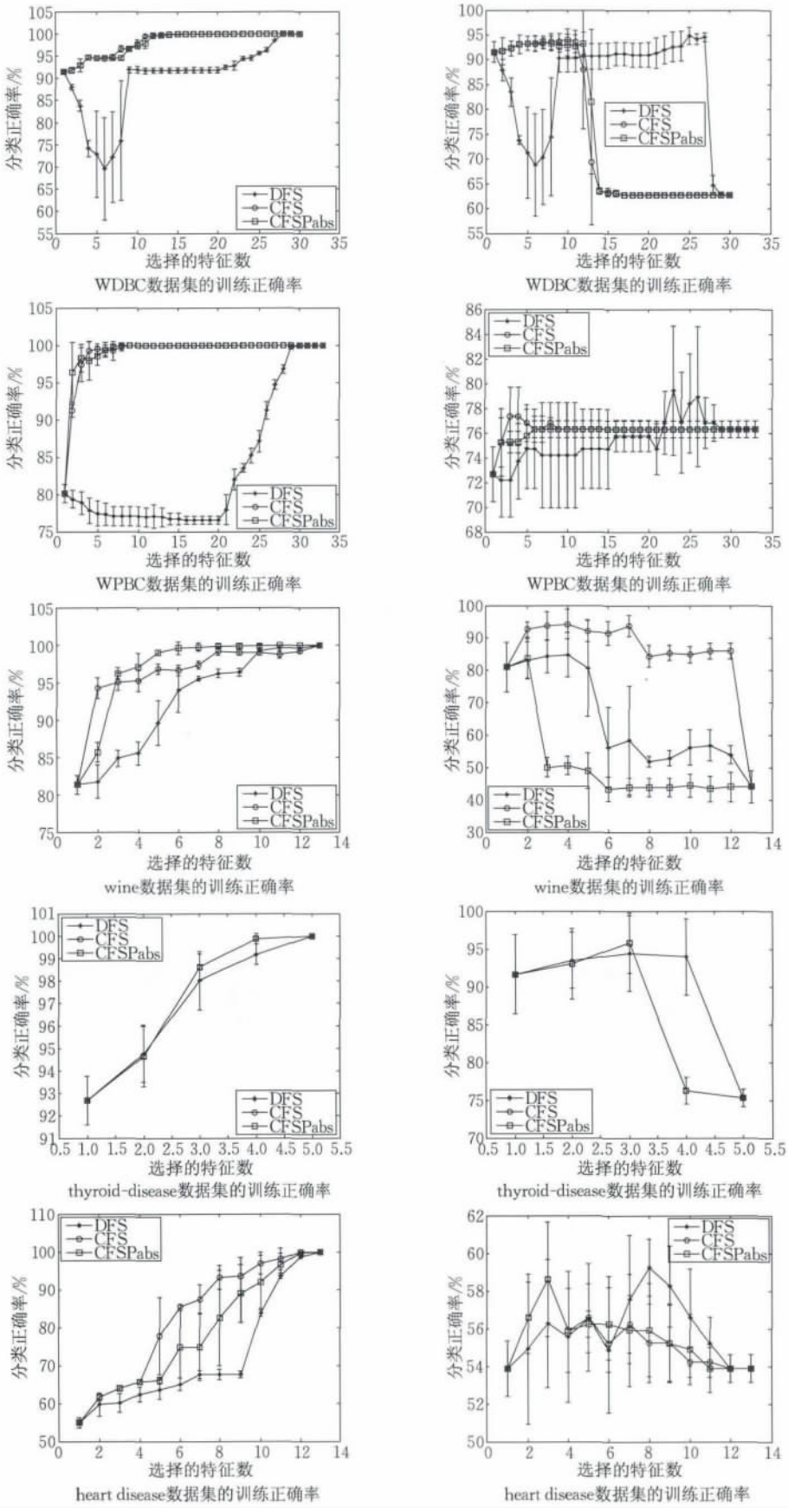
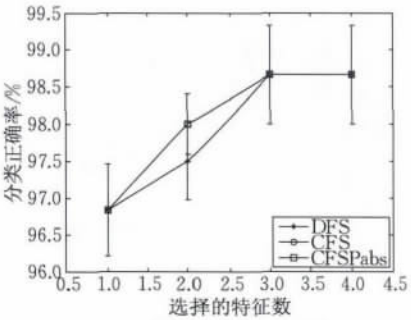
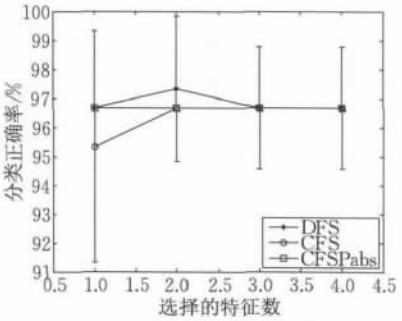


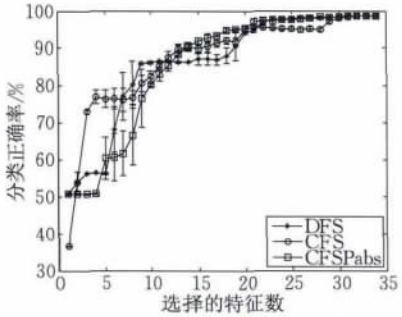
图 1　基于 SFS 搜索策略的 DFS、CFS 和 CFSPabs 特征子集评价准则的 5 折交叉验证实验的平均训练和测试正确率



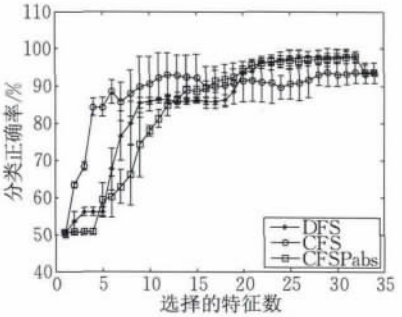
iris数据集的训练正确率



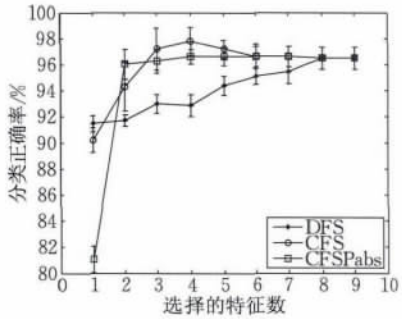
iris数据集的测试正确率



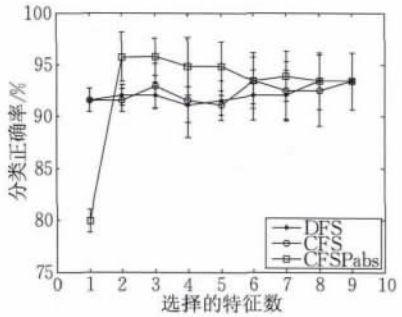
dermatology数据集的训练正确率



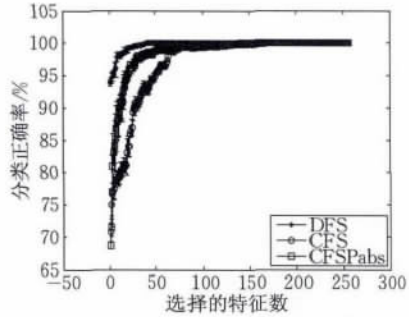
dermatology数据集的测试正确率



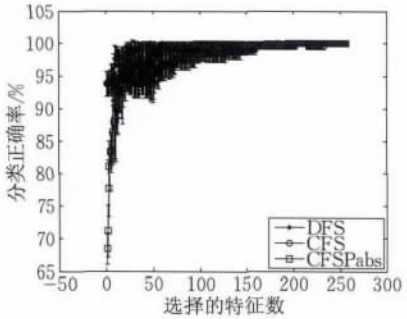
glass数据集的训练正确率



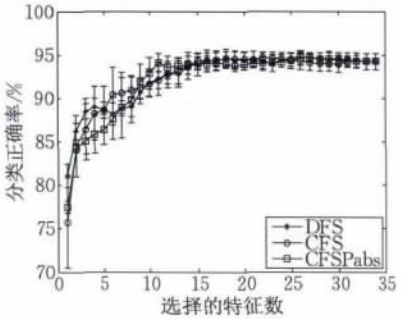
glass数据集的测试正确率



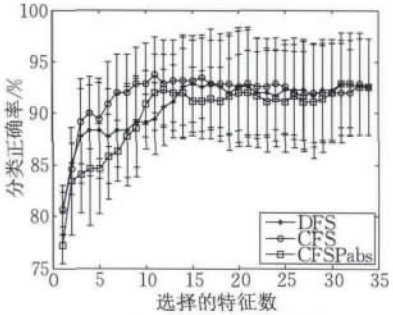
handwrite数据集的训练正确率



handwrite数据集的测试正确率



ionosphere数据集的训练正确率



ionosphere数据集的测试正确率

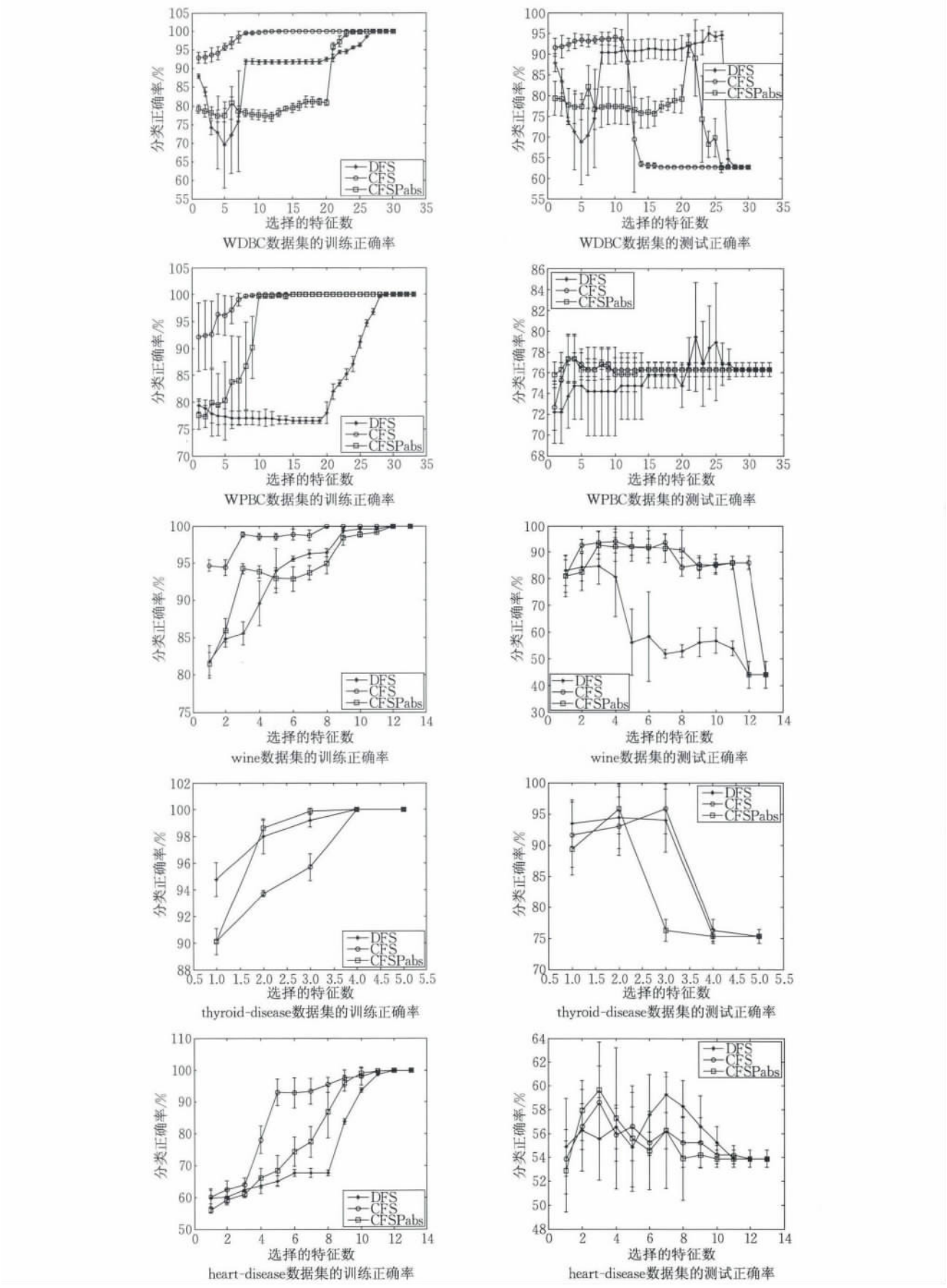


图 2　基于 SBS 搜索策略的 DFS、CFS 和 CFSPabs 特征子集评价准则的 5 折交叉验证实验的平均训练和测试正确率

DFS 特征子集区分度衡量准则最好. 其选择的特征子集不仅泛化性能好,而且对于较高维数据集的降维效果特别突出;另外,基于 DFS 特征子集评价准则与 SVM 的特征选择算法的运行效率较高.

表 4 基于 SFFS 搜索策略的实验结果显示:提出的 DFS 特征子集区分度衡量准则在 3 个数据集 iris、handwrite 和 thyroid-disease 上选择的特征子集的分类正确率最高;CFS 准则在 glass、WPBC 和 wine 3 个数据集上选择的特征子集具有较高的分类正确率;改进的 CFSPabs 准则在 Dermatology、ionosphere 和 WDBC 3 个数聚集上选择的特征子集分类正确率最好;对于 heart disease 数据集,DFS、CFS 和 CFSPabs 3 个特征子集评价准则选择的特征子集具有相同的分类性能. 特征子集规模分析可见,改进的 CFSPabs 准则在 dermatology、glass、handwrite、ionosphere、WDBC、WPBC、wine 和 thyroid-disease 共 8 个数聚集上选择的特征子集规模不超过 DFS 准则和 CFS 准则,其中在 WDBC 和 thyroid-disease 两个数据集上与 CFS 准则相同;CFS 准则只在 heart disease 数据集上最优;提出的 DFS 准则只在 iris 数据集上选择的特征子集规模优于其他两准则.

表 5 基于 SBFS 搜索策略的实验结果显示,提出的 DFS 特征子集区分度评价准则性能最优,因为基于 DFS 准则的顺序后向浮动特征选择算法的 5 折交叉验证实验,无论其选择的特征子集平均规模,还是其选择的特征子集的平均分类正确率,还是其平均运行时间都优于基于其他两个特征子集评价准则 CFS 与 CFSPab 的特征选择算法.

表 2～表 5 的实验结果还显示:对于较高维数据集进行特征选择,SFS 特征搜索策略的运行时间低于 SBS, SFFS 搜索策略需要的时间低于 SBFS. 另外,从 handwrite 数据集的 5 折交叉验证实验的运行时间明显可见,除了采用 SFFS 搜索策略的特征选择算法外,其他以 SFS、SBS、SBFS 为搜索策略的 3 个特征选择算法,DFS 特征子集评价准则明显优于 CFS 和 CFSPabs 准则. 所得分类器在该数据集的泛化性能,无论采用那个搜索策略,DFS 特征子集评价准则选择的特征子集的分类性能都最好.

图 1 关于特征依次被选入时在 10 个 UCI 数据集的 5 折交叉验证实验的平均训练和测试正确率比较显示,提出的 DFS 特征子集重要性评价准则选择的特征子集具有更好的泛化性能,CFSPabs 准则能选择到规模较小的特征子集. 图 1 展示的详细实验结果与表 2 所示的实验结果一致.

图 2 关于特征依次被剔除时的 5 折交叉验证实验的平均训练和测试正确率的详细结果与表 3 展示的实验结果一致. 从图 2 实验结果的比较看出,提出的 DFS 特征子集评价准则选择的特征子集的泛化性能优于 CFS 和 CFSPabs 准则. 另外,在 handwrite 这一较高维数据集,DFS 特征子集评价准则不仅选择的特征子集的泛化性能很好,而且选择的特征子集的规模也小于 CFS 和 CFSPabs 准则.

以上 5 折交叉验证实验的结果分析揭示:本文提出的 DFS 特征子集区分度衡量准则是一个很好的特征子集类间区分能力度量准则,基于该准则的特征选择算法所选择的特征子集具有很好的泛化性能,且对于较高维数据集,该准则不仅大大降低数据维数,还具有很高的运行效率,需要的运行时间最少. 我们对 CFS 准则改进得到的 CFSPabs 特征子集评价准则所选特征子集的泛化性能与 CFS 准则选择的特征子集的泛化性能差别不大,但是基于 CFSPabs 准则的特征选择算法运行时间效率较高,且选择的特征子集规模优于 CFS 准则选择的特征子集的规模.

5 结论与应用前景展望

本文提出了 DFS 特征子集区分度评价准则,该准则充分考虑特征之间的相关性,通过计算整个特征子集对于分类的联合贡献克服了单个特征区分度准则在衡量特征的类间辨别能力大小时,没有考虑特征之间的相关性对于单个特征辨别能力大小影响的缺憾,以及 Hall 提出的 CFS 特征子集评价准则对于特征之间相关性考虑不足和不适于连续性数值的缺陷. 在此基础上,以 DFS 准则作为特征选择依据,提出基于不同特征搜索策略与 SVM 的 4 种混合特征选择算法.

UCI 机器学习数据库的 10 个经典数据集的 5 折交叉验证实验表明:提出的 DFS 特征子集区分度评价准则是一种有效的特征子集辨识能力衡量准则,基于该准则与 SVM 的混合特征选择算法选择的特征子集具有很好的分类效果,其泛化性能优于分别基于 CFS 和 CFSPabs 特征子集评价准则的特征选择算法,达到了保持数据集辨识能力情况下进行维数压缩的目的,特别是对较高维数据集,DFS

特征子集评价准则特别有效.实验结果同时显示:我们改进的 CFSPabs 特征子集评价准则优于 CFS 准则.

本文实验还揭示,提出的 DFS 特征子集区分度衡量准则选择出了分类红斑鳞状皮肤病等疾病的有效特征子集,实现了对相应疾病的诊断.

另外,伴着计算机网络带来的文本大数据,以及医疗诊断大数据和社会大数据,还有随着分子生物技术发展产生的癌症基因大数据,特征选择是对这些大数据进行分类分析的首要和关键步骤.本文提出的特征选择方法及其变种将在这些蓬勃兴起的大数据领域发挥重要作用.继本文之后,我们已经开展了关于癌症基因数据集的特征选择方法研究,并取得了很好的实验结果.

参 考 文 献

[1] Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research, 2003, 3: 1157-1182

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

[3] Rakotomamonjy A. Variable selection using svm based criteria. The Journal of Machine Learning Research, 2003, 3: 1357-1370

[4] Duan K B, Rajapakse J C, Wang H, et al. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Transactions on NanoBioscience, 2005, 4(3): 228-234

[5] Xia H, Hu B Q. Feature selection using fuzzy support vector machines. Fuzzy Optimization and Decision Making, 2006, 5(2): 187-192

[6] Zhou X, Tuck D P. MSVM-RFE: Extensions of SVM-RFE for multiclass gene selection on DNA microarray data. Bioinformatics, 2007, 23(9): 1106-1114

[7] Maldonado S, Weber R. A wrapper method for feature selection using support vector machines. Information Sciences, 2009, 179(13): 2208-2217

[8] Somol P, Novovicova J. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(11): 1921-1939

[9] Tapia E, Bulacio P, Angelone L. Sparse and stable gene selection with consensus SVM-RFE. Pattern Recognition Letters, 2012, 33(2): 164-172

[10] Mao Yong, Zhou Xiao-Bo, Xia Zheng, et al. A survey for study of feature selection algorithms. Chinese Journal of Pattern Recognition and Artificial Intelligence, 2007, 20(2): 211-218(in Chinese)
(毛勇, 周晓波, 夏铮等. 特征选择算法研究综述. 模式识别与人工智能, 2007, 20(2): 211-218)

[11] Whitney A W. A direct method of nonparametric measurement selection. IEEE Transactions on Computers, 1971, 100(9): 1100-1103

[12] Marill T, Green D M. On the effectiveness of receptors in recognition systems. IEEE Transactions on Information Theory, 1963, 9(1): 11-17

[13] Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognition Letters, 1994, 15(11): 1119-1125

[14] Blum A L, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997, 97(1): 245-271

[15] Kohavi R, John G H. Wrappers for feature subset selection. Artificial Intelligence, 1997, 97(1): 273-324

[16] Lal T N, Chapelle O, Weston J, et al. Embedded methods// Guyon I, Nikravesh M, Gunn S, Zadeh L A eds. Feature extraction. Berlin Heidelberg: Springer-Verlag, 2006: 137-165

[17] Uncu Ö, Türkşen I B. A novel feature selection approach: Combining feature wrappers and filters. Information Sciences, 2007, 177(2): 449-466

[18] Hu Q, Pedrycz W, Yu D, et al. Selecting discrete and continuous features based on neighborhood decision error minimization. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2010, 40(1): 137-150

[19] Xie J, Wang C. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Systems with Applications, 2011, 38(5): 5809-5815

[20] Hall M A. Correlation-based feature selection for machine learning [Ph. D. dissertation]. The University of Waikato, Hamilton, New Zealand, 1999

[21] Frank A, Asuncion A. UCI machine learning repository [EB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html. Irvine, CA: University of California, School of Information and Computer Science, 2010

[22] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 1-27

[23] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification. Department of Computer Science, National Taiwan University, Technical Report. Taipei, China, 2003

XIE Juan-Ying, born in 1971, Ph.D., associate professor. Her research interests are machine learning and data mining.

XIE Wei-Xin, born in 1941, professor. His research interests include intelligent information processing and fuzzy information processing.

Background

Support vector machine (SVM) based feature selection has been the very popular research area, especially in cancer patient recognition from normal people. There are many researchers focus on the research in this field. There are three kinds of feature selection algorithms, and they are respectively filters, wrappers and embedded algorithms. The classic feature selection algorithm based on SVM is the very popular SVM-RFE proposed by professor Guyon. It is a well known wrapper algorithm of feature selection, but it has a heavy computational load. The filter feature selection algorithms based on SVM are efficient for the feature selection procedure are independent of the specific classification algorithms. However, there are some deficiencies in the available filter feature selection algorithms, such as the correlation between features is not considered or not considered thoroughly when evaluating the significance of features to classification, which caused the accuracy of this kind of feature selection algorithms are not as well as that of SVM-RFE. The motivation of this paper is to overcome the deficiencies of the available SVM based feature selection algorithms. Therefore, the authors follow their previous study which was published in the Expert system with Applications and propose the criterion to evaluate the discernibility of a feature subset not only of a feature to classification by taking into account the correlation between features and calculate the contribution of the whole feature subset to the classification. The authors combined the proposed criterion and the four popular search strategies whilst using SVM as the classification tool to guide the feature selection procedures and proposed the four hybrid feature selection algorithms in this work. These algorithms were tested on the very popular datasets from UCI machine learning repository. The experimental results demonstrate that the proposed criterion is very promising and the feature subset selected by it has got the high classification accuracy, and the feature selection algorithms based on the proposed criterion are faster than those based on the other criterion which valued the contribution of a whole feature subset to the classification. The work in this paper is very promising. It and its variation can be used in dealing with the big data because the feature selection is the key and the primary procedure in analyzing the big data, such as mining the text big dataset and analyzing the gene cancer dataset, etc.

This paper is supported in part by the National Natural Science Foundation of China under Grant No. 31372250, and is also supported by the Key Science and Technology Program of Shaanxi Province of China under Grant No. 2013K12-03-24, and is supported by the Fundamental Research Funds for the Central Universities of China under Grant No. GK201102007 as well.