代　号	10701	学　号	0410310215
分类号	TP18	密　级	

题（中、英文）目 基于 SVM 的特征选择方法研究

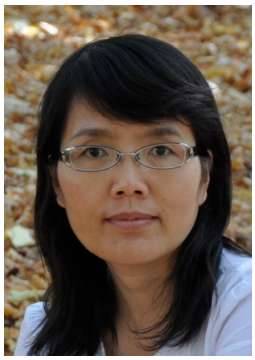
SVM Based Feature Selection Algorithms for Classification

作 者 姓 名 谢娟英 指导教师姓名、职 谢维信 教授

学 科 门 类 工学 学科、专业 信号与信息处理

提 交 论 文 日 二○一二年四月

作者简介



谢娟英，陕西西安人。1993 年毕业于陕西师范大学、获学士学位。2004 年毕业西安电子科技大学获硕士学位。博士导师：谢维信教授。

主要研究方向：智能信息处理、机器学习、数据挖掘和模式识别等。

代表性成果及经历：（获奖、专利、专著、论文等信息、参与或完成实际工程、访学经历）已在《Expert Systems with Applications 》、《 JMLR: Workshop and conference Proceedings》、《LNCS》、《Journal of Computers》、《计算机应用》、《中国生物医学工程学报》、《山东大学学报》（理学版）、《南京大学学报》（自然科学版）、《陕西师范大学学报》（自然科学版）等权威、核心刊物和国际重要学术会议发表学术论文 10 多篇。2010.11～2011.11 受国家留学基金委资助于英国 Brunel 大学做访问学者，从事基因选择研究。

Juanying XIE, was born in Xi’an, Shaanxi Province, PR China, in 1971. She received her bachelor's degree in Computer Science from Shaanxi Normal University, Xi’an, PR China, in 1993, and the M.S. degree in Computer Science & Technology from XiDian University, Xi’an, PR China, in 2004. Her PhD supervisor is professor Weixin Xie.

Her research interests include Intelligent Information Processing, Machine Learning, Data Mining and Pattern Recognition.

From November of 2010 to the November of 2011 she has been awarded a scholarship under the China Scholarship Council (CSC) to pursue her research at Brunel University in the United Kingdom of British as an academic visitor. She has published over 10 journal and conference papers. One of them is belong to the 2nd field of SCI according to the JCR of SCI.

西安电子科技大学
学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中做了明确的说明并表示了谢意。

申请学位论文与资料若有不实之处，本人承担一切的法律责任。

本人签名：_____　　日　　期：_____

西安电子科技大学
关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属西安电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。同时本人保证，毕业后结合学位论文研究课题再撰写的文章一律署名单位为西安电子科技大学。

（保密的论文在解密后遵守此规定）

本学位论文属于保密，在____年解密后适用本授权书。

本人签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘　　要

特征选择通过选择一个最优的特征子集降低数据维数，构造一个简洁的分类系统，提高分类预测的准确性，揭示隐藏的潜在模式和规律，得到一个快速、高效的分类器，并使分类结果可视化成为可能。

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：特征子集搜索策略和特征子集性能评价准则。基于 SVM 的特征选择方法研究存在如下问题：如何评价特征的重要性，即如何判断特征对于分类的贡献？如何考虑特征之间的相关性？如何确定最佳的被选择特征数目？如何选择合适的 SVM 分类器模型、合适的 SVM 参数？对超高维、小样本的基因数据集进行分类分析时，如何实现基因选择？另外，现有基于 SVM 的特征选择方法主要基于后向剔除思想，而后向剔除相对于前向选择时间效率较差。

本研究针对基于 SVM 的特征选择算法研究存在的以上问题，提出分别基于 4 种不同特征重要性评价准则与 SVM 的特征选择算法；并针对基因数据集的高维小样本特点，提出了基于 SVM 分类模型的随机基因选择算法。所取得的主要研究成果包括：

1. **提出基于 G-score 与 SVM 的适用于任意类分类问题的特征选择算法，所提算法弥补了基于 F-score 与 SVM 的特征选择算法只适用于两类分类问题的不足**。其中，G-score 将 F-score 特征重要性评价准则由评价两类分类问题的特征区分度推广到可以衡量任意类分类问题的特征区分度；算法的特征搜索策略采用推广的前向顺序搜索策略 GSFS（Generalized Sequential Forward Search，GSFS）、推广的前向顺序浮动搜索策略 GSFFS（Generalized Sequential Forward Floating Search，GSFFS），以及推广的后向顺序浮动搜索策略 GSBFS（Generalized Sequential Backward Floating Search，GSBFS）。UCI 机器学习数据库数据集的实验显示：基于 G-score 与 SVM 的三种混合特征选择算法可以实现有效的特征选择，其中就特征子集规模来看，基于 G-score 与 SVM 的前向顺序浮动特征选择算法效果最佳；但就分类正确率，即分类器的泛化性能来看，相应的前向顺序特征选择算法最优。

2. **提出基于 D-score 与 SVM 的特征选择算法，该算法避免了基于 G-score 与 SVM 的特征选择算法在衡量特征的类间辨别能力大小时，没有考虑不同特征的测量量纲对特征区分度大小的影响问题**。UCI 机器学习数据库的 9 个特征选择常用数据集实验测试，以及与相应的基于 G-score 与 SVM 特征选择算法的实验比较表明：提出的基于 D-score 与 SVM 的特征选择算法所选择的特征具有较好的分类效果，其分类性能优于基于 G-score 与 SVM 的特征选择方法，实现了保持数据集辨识能力不变情况下进行维数压缩的目的。基于 D-score 与 SVM 的 3 种

混合特征选择算法相比，就特征子集规模来看，前向顺序浮动特征选择算法最好；但就分类器的泛化性能相比，前向顺序特征选择算法具有最好的泛化性能。

3. **提出基于 DFS（Discernibility of Feature Subsets，DFS）与 SVM 的特征选择算法，该算法解决了基于 G-score 与 SVM、D-score 与 SVM 的特征选择算法在衡量特征的类间辨别能力大小时候，没有考虑特征之间的相关性对于单个特征的类间辨别能力大小的影响问题**。其中，DFS 是一种新的特征子集区分度评价方法，通过计算多个特征构成的特征子集的 G-score 值，判断特征子集的类间区分度大小，考虑了特征子集中特征的联合作用，即特征子集中所有特征对于分类的联合贡献。同时根据特征子集评价方法 CFS（Correlation based Feature Selector, CFS）中 Pearson 相关系数度量特征相关性的正、负相关之分，提出不区分特征之间的正、负相关，只考虑其是否相关的 CFSPabs（Correlation based Feature Selector based on the absolute of Pearson's correlation coefficient，CFSPabs）方法。特征搜索策略分别采用经典的顺序前向搜索（Sequential Forward Search, SFS）、顺序后向搜索（Sequential Backward Search, SBS）、顺序前向浮动搜索（Sequential Forward Floating Search, SFFS）、顺序后向浮动搜索（Sequential Backward Floating Search, SBFS）4 种搜索策略，区别在于在浮动搜索策略中，判断特征加入／剔除应用特征子集的区分度 DFS，而浮动剔除／加入特征应用分类器的训练准确率。UCI 机器学习数据库中 10 个经典数据集的 5 折交叉验证实验表明：提出的基于 DFS 特征子集评价准则与 SVM 的特征选择算法是一种有效的特征子集选择方法，该方法所选特征子集的分类性能优于分别基于 CFS 与 SVM、CFSPabs 与 SVM 的特征选择方法；但是就特征子集规模来看，基于 CFSPabs 与 SVM 的方法最优。

4. **鉴于 SVM 对于非线性可分问题的最大泛化性能，提出基于 SVM 分类模型的适用于多类分类问题的特征选择方法 SVM RFE（SVM Recursive Feature Elimination）和 SVM RFA（SVM Recursive Feature Addition），避免分别基于 G-score、D-score 和 DFS 与 SVM 的特征选择算法在非线性可分问题中有可能误剔除有效区分特征的缺陷；同时克服 Guyou 的 SVM-RFE 特征选择算法只适用于两类分类问题的缺陷**。UCI 机器学习数据库的 9 个经典数据集的 5 折交叉验证实验表明：提出的 SVM RFA 和 SVM RFE 特征选择算法能在保持或提高分类正确率的前提下，实现有效的特征选择；9 个数据集的实验测试，SVM RFA 算法在 8 个数据集上优于 SVM RFE 算法。实验还证明，对于较低维数据集，该两个特征选择算法的效率差别不大，但是对于维数比较高的数据集进行特征选择时，SVM RFA 特征选择算法的效率明显优于 SVM RFE 算法。

5. **针对基因数据集的高维小样本特点，并结合上一研究结论，提出基于 SVM 分类模型的基因选择算法——SVM RRFA（SVM Recursive Random**

Feature Addition, SVM RRFA），该算法引入随机思想，针对具体的基因数据集，在每次迭代中同时加入若干个随机数确定的基因。为了减少算法时间开销，提出了简化的 SVM RRFA 基因选择算法。普林斯顿大学基因表达工程的 3 个基因数据集实验测试和比较表明：提出的 SVM RRFA 基因选择算法实现了有效的基因选择，发现了基因数据集的关键区分基因，实现了有效的癌症分类诊断；简化 SVM RRFA 算法提高了 SVM RRFA 基因选择算法的分类正确率、特异性和 Matthews 相关系数；但是对于癌症患者的分类正确率并没有提高。

关键词：SVM 特征选择 特征区分度 特征子集区分度 基因选择

ABSTRACT

This thesis analyzes the problems of the available researches on feature selection based on Support Vector Machines (SVM) firstly. Then the feature selection algorithms are presented based on the specific criterion and the SVM where the specific criterion includes four new kinds of criteria to evaluate the discrimination of features between classes. Among them three criteria focus on the discernibility of a feature between classes, and the other one is about that of a feature subset between classes in classification problems. Finally, according to the properties of gene datasets, the special feature selection algorithms for gene selection are introduced. The author's major contributions are outlined here.

1. **The G-score and SVM based feature selection algorithms are proposed to overcome the disadvantages of the feature selection algorithms based on F-score and SVM that can only be used to deal with a binary classification problem**. Where, the G-score is the generalization of F-score, so that the criterion can be used to measure the discrimination of features between more than two sets of real numbers. At the same time, the sequential forward search strategy, and the sequential forward floating search strategy, and the sequential backward floating search strategy are generalized. These generalized strategies are referred to as GSFS, GSFFS, and GSBFS in short, respectively, in this thesis. These proposed feature selection algorithms are tested on the datasets from UCI machine learning repository. The experimental results prove the validation of the G-score and SVM based feature selection algorithms, and also show that the algorithm using GSFFS search strategy is the best one according to the size of the selected feature subset, while the one using GSFS search strategy is the optimal when the generalization of the classifier is considered.

2. **The D-score and SVM based hybrid feature selection algorithms are proposed to conquer the deficiency of the feature selection algorithms based on G-score and SVM where the influence of different measurement units on different features when measuring their discriminability is not considered**. This D-score criterion not only has the property as the G-score in measuring the discrimination between more than two sets of real numbers, but also is not influenced by different measure for features when calculating their discriminabilities. D-score is used as a criterion to measure the importance of a feature, and GSFS, GSFFS, and GSBFS strategies are, respectively, adopted as search strategies to select features, whilst SVM is used as the classification tool, so that three new hybrid feature selection methods have

been got. The three new hybrid feature selection methods combine the advantages of filters and wrappers where SVM plays the role to evaluate the classification capacity of the selected feature subset via the classification accuracy, and leads the feature selection procedure. These new hybrid feature selection algorithms are tested on nine datasets from UCI machine learning repository and compared with the corresponding algorithms that are based on G-score and SVM. Experimental results show that the D-score and SVM based hybrid feature selelction algorithms outperform the ones based on the G-score and SVM, and can implement the dimension reduction without compromising the classification capacity of datasets. Among the three hybrid feature selection algorithms based on D-score and SVM, the one using the GSFFS search strategy is the best one according to the size of the selected feature subset, and the one based on the GSFS search strategy is best when considering the generalization of a classifier.

3. **The DFS (Discernibility of Feature Subsets) and SVM based hybrid feature selection algorithms are brought forward ot avoid the deficiencies in the G-score and SVM based or the D-score and SVM based hybrid feature selelction algorithms where the influence of the correlation between features is not considered when evaluating the importance of features between classes in classification problems**. The DFS criterion considers the contribution of each feature in a feature subset to classification together. It computes the combination G-score of features in a feature subset, and uses the combination G-score as the discernibility of the feature subset. The strategies for searching features are the four popular and classic search strategies including sequential forward search (SFS), sequential backward search (SBS), sequential forward floating search (SFFS), and sequential backward floating search (SBFS). However there are some differences to classic SFFS and SBFS when we use them. We add a feature in SFFS, or delete a feature in SBFS, according to the value of DFS of the feature subset, and make a juge to delete or bring back that feature when floating using the accuracy of the classifier on training subset. In addition, we put forward the improved CFS (Correlation based Feature Selector) criterion, named as CFSPabs (Correlation based Feature Selector based on the absolute of Pearson's correlation coefficient). The CFSPabs does not consider the positive or negative correlation between features such as the CFS does, it only considers whether the features are correlated or not. The DFS and SVM based hybrid feature selection algorithms are tested on 10 UCI machine learning repository datasets. Experimental results show that DFS and SVM based hybrid feature selection algorithms are better

than the ones based on CFS and SVM or on CFSPabs and SVM. Howevere, the CFSPabs and SVM based feature selection algorithms are the best one when considering the size of a feature subset.

4. **Considering the generalization of SVM on nonlinear classification problems, the new feature selection algorithms are introduced here based on the SVM classifiers to defeat the potential deficiencies of the feature selection algorithms based on G-score and SVM, or on D-score and SVM, or on DFS and SVM where the active features may be deleted when dealing with the nonlinear classification problems**. **At the same time the disadvantages of the very popular SVM based feature selection algorithm SVM-RFE proposed by Guyon is solved as well**. The new algorithms are SVM RFA (SVM Recursive Feature Addition) and SVM RFE (SVM Recursive Feature Elimination). They calculated the importance of a feature according to the weights of it in SVM models. SVM RFE generalized SVM-RFE proposed by Guyon for binary classification problems to dealing with any classification problems. SVM RFA is based on the forward search strategy compared to the SVM RFE that relies on the backward strategy. These two feature selection algorithms are tested on nine classic datasets from UCI machine learning repository. The experimental results demonstrate that these two SVM model based feature selection algorithms can reduce the dimension of datasets without compromising the classification capacity of them, and the SVM RFA outperforms SVM RFE on eight datasets among the nine ones. Furthermore, from the experimental results we can see that SVM RFA is much more efficient than SVM RFE on high dimension datasets.

5. **According to the characteristic of gene data sets where the samples of a data set is often a few dozen while the dimension or features of a sample is usually several thousands to tens of thousands, and referring to the conclusion of the above study we propose the gene selection algorithm---SVM RRFA (SVM Recursive Random Feature Addition, SVM RRFA)**. In SVM RRFA, we randomly determine the number of genes that should be added at once iteration in the gene selection procedure according to the dimension of a specific data set. In order to reduce the run time of SVM RRFA, we developed the simplified SVM RRFA for gene selection. We tested our SVM RRFA and our simplified SVM RRFA on three gene data sets from the gene expression project of Princeton University. The experimental results show that our SVM RRFA can find the discernibility genes between cancer patients and normal people, and can classify the cancer people from normal people efficiently, and the

simplified SVM RRFA outperformed SVM RRFA in accuracy, specificity and Matthews correlation coefficient, but the sensitivity is not improved.

Keywords: SVM Feature selection Discernibility of a feature Discernibility of a feature subset Gene selection

目　录

第一章 绪论

特征选择是从原始特征集合中选择一部分特征构成一个特征子集。该特征子集能够保持能保持原系统的分类识别性能，同时只包含有原始特征集的最少的特征，特征选择是系统设计人员经常要遇到的问题[1]。

特征选择作为一种特征维约简方法，不同特征提取。特征选择是在保持样本分类能力不变的前提下，通过选择样本原始特征集的一个特征子集实现样本特征维的约简。其实质是：寻求原始输入空间的一个最优子空间，使样本在该子空间中的分类特性与在原空间中一样。也即，在样本的原始特征集中寻找一个能保持样本分类能力不变且包含特征数最少的最优特征子集。这个特征子集的各特征构成最优子空间的基，从而将原系统的样本很好的表达出来。而特征提取是将原有特征空间通过某种变换，映射到一个低维的新空间。低维空间的每一个特征，是原始空间特征的组合。主成分分析 (Principle Component Analysis, PCA)和独立成分分析(Independent Components Analysis, ICA)是特征提取中最常用的方法[2]。特征提取对学习任务可以实现较好的降维，但是特征提取后得到的新特征的理解性很差，因为即使简单的线性组合也会使构造出的特征难以理解，而在很多情况下，我们需要关注特征的实际意义。另外，特征提取的新特征通常由全部原始特征变换得到，使得数据测量的工作量依然存在，且数据的存储空间需求也没有降低。

特征选择的形式化描述如下：假如原始样本空间的维数是 n ， 则特征选择就是从 n 个特征中选择出 m（$m << n$）个最重要的特征，使分类器的期望泛化误差最小；或者在给定的泛化误差下，选择最少的 m 个特征，使这 m 个特征足以表述原问题。其本质就是从原始特征集的 n 个特征中选择足以保持原系统可分性的一个最小的特征子集，该子集的特征线性无关，构成原空间的子空间，特征子集的特征构成子空间的基。因此，特征选择是寻找原始 n 维空间的等价 m 维子空间，并将样本用等价 m 维子空间的 m 个坐标轴来表示。理论上，从包含 n 个特征的集合中选择出包含 m 个特征的特征子集，有 C_n^m 种可能的情况。因此，特征选择需要从 C_n^m 种可能的特征子集中选择一个最优的特征子集，这是一个 NP 问题[3~6]。

1.1 特征选择的分类

特征选择方法依据其与分类器的关系可分为三类[7, 8]：Filter 方法、Wrapper 方法和

Embedded 方法。其中 Filter 方法和 Wrapper 方法最常用。Filter 方法（如图 1 所示）的特点是：独立于学习过程。该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样本的特征子集。常用的特征重要性计算方法有卡方检验、信息增益、基尼系数等[9]。Filter 方法需要一个阈值作为特征选择的停止准则。该方法因为其独立于学习过程而时间效率较高，但是依据其所选特征子集构造的分类器，其分类性能不仅与计算特征重要性的方法，即特征的排序准则有关；而且与选择特征的策略，以及特征选择过程的停止准则密切相关。 经典的 Relief 特征选择算法[10]就是 Filter 方法。

Wrapper 方法[8]（如图 2 所示）自 Kohavi 和 John 研究之后成为一种广为流行的特征选择方法。其最大特点是：依赖于学习过程，需要将训练样本分成训练子集和测试子集两部分。Wrapper 方法中，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子集分类性能的度量。为选择出最优的特征子集，Wrapper 算法需要的计算量巨大；而且选择的最优特征子集依赖于具体的学习机；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成为 Wrapper 方法的关键。

Wrapper 方法由于将学习机作为一个“黑匣子”，使得它成为一种简单通用的特征选择方法，且能比 Filter 方法获得更精确的解，但是一般需要的计算量很大[8]，且泛化性能较差。Filter 算法和 Wrapper 算法结合，集成 Filter 方法的高效与 Wrapper 方法的高准确率与一起，进行特征选择，可得到最优的变量或特征子集[9, 11, 12]。该方法被称为混合的特征选择方法，也是目前特征选择方法研究的一个新趋势。

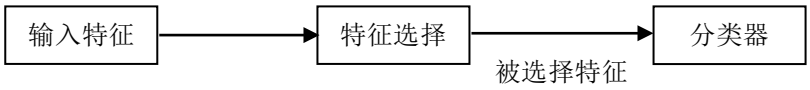


图 1　Filter 特征选择方法

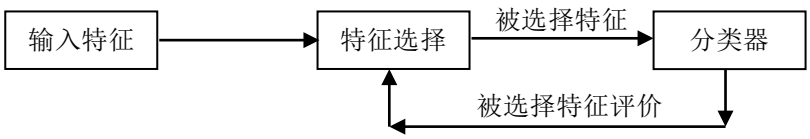


图 2　Wrapper 特征选择方法

Embedded 方法[13]将特征选择集成在学习机的训练过程中，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。该方法与 Wrapper 方法的最大区别在于：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，从而可以快速地得到最优的特征子集，是一种高效的特征选择方法。该方法的难点在于构造一个优化函数模型。Breiman 等人[14]提出的 CART (Classification and Regression Trees，CART)决策树方法就是一种 Embedded 特征选择方法。

1.2 特征选择的搜索策略

在特征选择中常用的特征搜索策略包括：首次适应、分支界定等[8]。Reunanen[15]研究指出，贪婪搜索策略既可以减少计算量，又可以克服“过适应”。常用的贪婪搜索策略包括：顺序前向选择（Sequential Forward Selection， SFS）[16]和顺序后向剔除（Sequential Backward Selection, SBS）[17]。但是 SFS 和 SBS 都有“子集嵌套”的缺点。Pudil 等人[18]提出的两种浮动搜索策略进行特征选择——顺序前向浮动（Sequential Forward Floating Selection, SFFS）和顺序后向浮动策略（Sequential Backward Floating Selection, SBFS）——改进了 SFS 和 SBS 的“子集嵌套”问题。随后，Somol[19~22]对搜索策略进行了诸多改进研究。Nakariyakul[23]提出了一种改进的浮动搜索策略。但在考虑计算效率和有效性的前提下，Pudil 等人[18]的浮动搜索策略仍然是一种很好的搜索策略。

1.3 特征选择的意义与面临的挑战

特征选择在构造一个分类系统中有重要作用[24]。它不仅可以降低数据的维数，而且可以减少计算开销、降低测量和存储需求，同时获得很好的分类性能。通过特征选择可以发现与分类最相关和重要的特征，剔除冗余特征，从而构造一个简洁的分类系统，提高分类预测的准确性，揭示隐藏的潜在模式规律，最终得到一个快速、高效的分类器，并使分类结果可视化成为可能。

当代分子生物技术的发展，加剧了高维小样本癌症基因数据集的产生[9, 25~31]。在这些基因数据集中，样本经常只有几十个，而描述每一个样本特性的基因却成千上万，这构成了高维稀疏分布空间。该空间的维数比空间里的样本数多得多。因此，传统的特征选择面临基因选择的挑战。

比如，结肠癌数据集[25]样本只有 62 个，而作为样本变量的基因却有 2,000 个。包含 72 个样本的白血病患者数据集[26]，每个样本的基因特征多达 7,129 个。另一个包含 72 个样本的白血病数据集[32]，每个样本的基因个数有 12,582 个之多。47 个样本的淋巴

瘤数据集[28]，每个样本的基因有 4,026 个。乳腺癌数据集[29]的样本规模为 97 个，但是每个样本的基因多达 24,481 个。因此，在癌症基因数据分析中，我们不得不面临具有成千上万，乃至上亿基因特征的小样本数据集。样本特征数与样本数之比变得非常巨大，以至于对这些癌症基因数据集进行分类分析，基因选择成为必不可少的关键和首要步骤[9,33,34]。因此，基因选择研究得到高度关注[35~59]。

另外，迅速发展的网络技术使得高维文本数据集急剧产生，这些文本数据具有非常高的维数；还有一些疾病诊断问题[60~64]，数据集不仅维数较高，而且经常包含有冗余特征。保留过多样本特征不仅使计算复杂度增加，降低计算效率，而且会导致“过适应”问题，影响分类的准确性。因此，对这些实际数据集进行分类分析，设计高效的分类模型，特征选择成为必不可少的首要步骤[9,33,34]。

这些实际应用需求，以及特征选择所能带来的优势，使得特征选择成为分类系统设计的必要和重要步骤；成为这些应用领域的研究热点问题；也对传统的特征选择提出了挑战。

1.4 SVM 的优越性使其成为特征选择的有力工具

支持向量机（Support Vector Machines, SVM）是 Boser[65]、Vapnik[66]和 Cristianini[67]分别于 1992、1998 和 2000 年提出的一种在高维空间中寻找最优分类超平面作为决策函数的两类分类算法。作为一种小样本学习机，SVM 在解决小样本、非线性及高维模式识别中表现出许多特有的优势，是迄今为止具有最小化分类错误率和最大化泛化能力的一种强有力的分类工具[68]，已经在模式识别、回归分析等机器学习领域得到广泛应用[69,70]，成为机器学习领域研究的热点。

SVM 是一种最小化结构风险的机器学习算法[65~67]，克服了经验风险最小化机器学习算法所带来的分类函数推广能力差，即分类器泛化能力差的缺憾。SVM 建立在统计学习理论的 VC 维和结构风险最小化原理基础上[65~67]，能根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，获得最好的推广能力也即泛化能力。

对于线性可分的两类样本，SVM 算法可以学习得到一个将这两类样本分开的最优分类超平面 $f(x)=\langle w\cdot x\rangle+b=0$。该最优分类超平面只需要少数几个样本就可以确定，这几个样本构成支持向量[65~67]。该最优分类超平面是具有最大几何间隔的超平面，即样本集到该最优分类超平面的欧氏距离最大，就是样本集中距离超平面最近的样本到超平面

的欧氏距离达到最大，因此这个最优分类超平面一定在两类中间，使得两类样本集到超平面的距离相等。此时，样本被误分的次数$\leq(2R/\gamma)^2$，γ是几何间隔，R是样本的最长向量长度值，R代表了样本的分布广度。因此，最大化几何间隔作为SVM学习的目标，就是最小化样本错分率。

对于线性不可分的两类样本，SVM通过核函数将低维输入空间线性不可分的样本映射到高维特征空间成为线性可分的样本，在高维特征空间求解线性可分的两类样本的最优分类超平面[65~67]。

常用的核函数有：线性核函数$K(x,x')=x\cdot x'$，多项式核函数函数$K(x,x')=(x\cdot x'+1)^d$，径向基核函数$K(x,x')=\exp(-\|x-x'\|^2/\sigma^2)$，S 型核函数$K(x,x_i)=\tanh(v(x\cdot x_i)+c)$。其中径向基核函数最受研究者们亲睐，因为它几乎可以解决所有的分类问题。但是在具有高维特征的问题中，比如基因选择、文本分类等，线性核函数经常被使用[33, 71~75]。

对于带有噪音的近似可分两类样本，SVM通过引入松弛变量得到一阶或二阶的软间隔分类器[65~67]，并通过惩罚因子调控噪音点对最优分类超平面的影响，惩罚因子越大噪音点的影响越强。

SVM 作为一种典型的两类分类器，对于多类分类问题，通常将原问题转换成多个两类分类问题解决[66]。同时，为了解决实际问题中经常遇到的一个样本不完全属于某一个类别的情况，Lin 等人[76]提出了模糊支持向量机（Fuzzy Support Vector Machines，FSVM）。Inoue 等人[77]提出了一对其余的多类模糊支持向量机， Abe 等人[78]提出了一对一的模糊支持向量机。Jayadeva 等人[79]提出了 Twin-SVM（TWSVM）。Peng 等人[80]对 Twin-SVM 进行了改进，提出了 υ-Twin SVM。Xie 等[81]将 TWSVM 推广到多类分类问题。Xu[82]通过引入零类标提出了一种有效的解决多类分类问题的 SVM。

SVM 通过一系列转化，将求解一个最优分类超平面的问题转化为一个凸二次规划问题[65~67]。凸二次规划具有全局最优解[83]。这个全局最优解对应的那几个样本就是支持向量。

模式识别（分类）问题中特征选择的原则是：从原始特征集中选择尽可能少的特征构成特征子集，该特征子集能使某种准则达到最优[1]。

分类问题的研究目标是：从原始特征集的诸多特征中选择出对于分类有重要贡献的特征，剔除不重要的、甚至冗余的特征，在保持系统分类性能的前提下保留尽可能少的特征构成特征子集，得到原始样本空间的最优子空间，使得样本在该子空间中具有可分

性。

对于疾病诊断、癌症患者与非患者分类分析，研究目标是从原始特征集中，或者成千上万的基因中选择出关键的癌症区分基因，或者称为主要的致病基因，以期获得对患者的高准确率诊断，促进癌症治疗及相关生物医学研究的发展。

那么，SVM 作为迄今为止的具有最小化错分率和最大化泛化能力的一种分类器，应用于特征选择研究将能发现对于分类分析起主要作用的特征，能发现疾病诊断、基因分类分析中的主要区分特征和基因，在此基础上构造的相应分类器，将获得最佳的分类效果，达到最小的错分率和最大的泛化性能。

因此，选择 SVM 作为分类工具，研究基于 SVM 的特征选择及其应用不仅有理论意义，更具有现实应用意义。

1.5 基于 SVM 的特征选择研究进展

毛勇等人[84]对特征选择算法进行了综述，指出特征选择研究主要着眼于选择优化特征子集所需要的特征子集搜索策略和特征子集性能评价准则。

因此，粗糙集理论以其特有的性质被应用于特征选择研究[85~100]。近年，作为粗糙集理论发展的粒子计算也被应用于特征选择研究[101, 102]。然而，粗糙集理论不能提供分类器错分率最小化的理论保障。

支持向量机（Support Vector machines，SVM)提供了最小化分类错误率和最大化泛化能力的理论保障[65~67]，因此，基于 SVM 的特征选择研究得到众多研究者关注[68, 103~129]，并出现了基于 FSVM 的特征选择研究[130]。同时，分子生物技术的发展，使得基于 SVM 的基因选择研究备受关注和亲睐[33, 63, 71~75, 131~140]。

现有基于 SVM 的特征（基因）选择算法可以归为四大类：基于 Filter 方法，以 SVM 为分类工具的特征（基因）选择算法；以 SVM-RFE 为基本框架的基因选择算法；以函数优化为目标的 embedded 型的基因选择算法；混合特征（基因）选择算法。

1.5.1 基于 SVM 的 Filter 特征选择研究进展

基于 SVM 的 Filter 特征选择方法，首先计算样本变量或特征的重要性，选择若干最重要的特征，使用 SVM 分类器进行分类。该类基于 SVM 的特征选择算法常用的样本特征重要性计算方法有：相关性分析、信息增益、基尼系数、F-score、卡方检验、t-统计以及根据 SVM 分类器的错分率定义的特征重要性等[9, 63, 75, 105]。Chen 等[105]对 F-score 所进行的不同数据集上的对比实验证实：实际应用中，选用哪种方法度量特征的重要性，

需要分析具体问题来定。另外，基于 SVM 的变量或特征选择，是在样本输入空间进行变量选择，还是在经过核函数映射后的高维特征空间进行特征选择，也需要根据问题而定。Ponsa 和 López[110]提出一种新的 MRMR 准则评价特征的重要性，使用线性 SVM 进行训练，实验结果证明新的 MRMR 准则优于 Peng 等人[47]的 MRMR 准则。Bonev 等[141]提出了一种基于信息论方法的 Filter 特征选择算法，避开维数灾难，图像分类数据和维阵列基因数据集分类实验验证了该算法优于 Peng 等人的 MRMR[47]基因选择算法。Polat 等人[63]对心脏病、Escherichia coli promoter 基因序列数据库的数据在经过核函数映射后的高维特征空间进行特征选择，计算各个特征在高维特征空间中的 F-score 值，保留 F-score 值大于平均 F-score 值的特征，剔除其余特征，用最小平方 SVM 进行分类，对比实验显示基于 SVM 径向基核函数的 F-score 特征选择算法最有效。Wiliński 等[75]针对肿瘤识别问题，应用不同的特征排序准则对基因进行重要性排序，选择排在前面的指定数目的基因，应用这些基因构造分类器，他所选用的分类器是高斯核函数 SVM 分类器。Wiliński 选用的基因（特征）排序准则包括：相关性分析[142]；KS（Kolmogorov-Smirnov）检验[143]，并使用了三种不同的 KS 测量；Wilcoxon-Mann-Whitney 检验[143]；基于均值和方差的类别区分度统计测量[26,142,144]；以及线性核函数的 SVM[24]。其中基于线性 SVM 的特征排序准则是：每次选择一个特征训练线性 SVM，这样训练特征数个 SVM，每一个特征的区分能力以相应的只含有该特征的线性 SVM 的错分率来衡量，错分率越低则该特征越重要。Wiliński 在 7 个基因数据集上的实验证明基于 SVM 的 Filter 特征选择方法效果最佳。

然而 Wiliński 的研究所使用的特征排序准则只是对特征进行了排序，不能指出最优的被选择特征数目。确定最佳的被选择基因数目，以期获得最好的分类准确率依然是一个有待进一步研究的问题。

Chandra 和 Gupta[145]提出了基于统计的 ERGS（Effective Range based Gene Selection）基因选择算法，基于统计思想给类间区分能力强的特征定义高的权重，实验证实该基因选择算法能对基因进行有效的排序，并选择出与相应疾病最相关的基因。Huang[146]提出了独立于分类模型的最小期望误分代价特征选择算法（minimum expected cost of misclassification，MEMC），该算法不同于评价单个特征重要性的 Filter 特征选择算法，它评价一个特征子集的区分度，结合特征搜索策略 SFS 实现特征选择。

基于 SVM 的 Filter 特征（基因）选择算法，特征选择过程独立于 SVM 分类器，计算效率较高。该类特征选择算法的关键是：特征的评价准则。选用哪一种方法评价特征

（基因）重要性，对基因进行排序没有普适性原则。选择恰当的基因重要性度量准则，实现基因排序依然是一个开放性问题。同时，选择多少个基因可以获得最佳的分类准确率，即特征选择的阈值问题，也是一个有待研究的问题。

1.5.2 基于 SVM-RFE 框架的特征选择研究进展

Guyon 等[33]提出的 SVM-RFE（SVM Recursive Feature Elimination）特征选择方法奠定了基于 SVM 的 Embedded / Wrapper 特征选择算法与基因选择研究的基础。该方法在每一次递归迭代时剔除掉排序在最后的那个特征，对于基因选择，每次可以剔除排序在最后的若干个基因。特征（基因）排序的原则是：训练 SVM 学习机，得到当前的最优分类超平面，计算权向量 $w = \sum_k \alpha_k y_k x_k$，则第 i 个特征的排序准则（重要性）为 $c_i = (w_i)^2$。Guyon 在该研究中发现：在相同的学习机下，分类器的性能取决于所采用的变量或特征选择方法；基于 SVM 特征选择算法的分类器，其分类性能优于使用别的特征选择算法的分类器。同时，Guyon 应用 SVM-RFE 对结肠癌数据集进行特征选择发现：所选特征数和 SVM 学习机的支持向量个数相等，都是 7；对白血病数据集进行特征选择所得特征个数为 4，支持向量的个数为 5，相差 1。

因为基因数据集的高维稀疏特点，Goyon 等[33]指出，每次迭代可以剔除若干个基因。但是，在具体实验中到底剔除多少个才能得到最佳的基因选择结果？基因选择的结果是否受每次剔除的基因个数不同的影响？

继 Guyon 之后，Rakotomamonjy[104]提出了三种基于 SVM 的特征选择算法，作为 Guyon 等的 SVM-RFE 方法的扩展，分别考察了权向量 $\|w\|^2$，误差界 $R^2\|w\|^2$，以及范围估计 $\sum_p \alpha_p^* S_p^2$ 三个特征排序标准，并将梯度应用于这些标准，应用顺序后向剔除贪心算法来搜索若干个最优的特征构成特征子集。与 Guyon 的 SVM-RFE 的不同之处在于，Rakotomamonjy 以每个特征的梯度作为特征排序的准则，每次迭代剔除具有最小梯度的那个特征。Huang 等[71]通过结肠癌[25]和淋巴癌[28]两个基因数据集，研究了 SVM-RFE 特征选择算法进行基因选择时，SVM 分类器的惩罚参数 C 和不同的数据预处理方法对癌症诊断准确率的影响。Huang 等研究发现样本标准化优于特征标准化；小的惩罚参数带来小的分类器错分率。Li 等[73]通过白血病微阵列数据集[27]、乳腺癌[29]数据集，以及 GCM 数据集[147]中的淋巴瘤和白血病数据集，以 SVM、Ridge Regression 和 Rocchio 三个分类

器对 SVM-RFE 方法在基因选择中的有效性进行了理论分析和实验比较。Li 等人发现在递归地剔除特征的过程中，分类器对于冗余特征的惩罚能力决定了其所选择特征的分类有效性。为了减少偏差，Duan 等[72]提出在 SVM-RFE 的每步迭代中训练多个 SVM 分类器，并根据所训练的多个分类器的训练结果定义了计算每个特征重要性的新指标，改进了 SVM-RFE 特征选择算法，但是增加了很大的计算量。Xia 等[130]将模糊支持向量机（FSVM）应用于特征选择，提出了基于 FSVM 的特征选择算法 FSVM-SBS。Mundra 和 Rajapakse[148]提出了基于支持向量的 t-score 基因排序准则，与 SVM-RFE 类似对特征进行后向剔除，3 个基因数据集上的实验证实该算法优于 SVM-RFE 和基于 t-score 的特征选择算法。Maldonado 等人[116]提出一种新的类似于 SVM-RFE 的另一种基于 SVM 的特征选择算法，该算法以去掉某一属性所引起的分类器错分样本数为依据对特征进行排序，采用后向顺序选择，每一次剔除掉排在最后的那个特征，即剔除掉去掉该特征所引起的错分样本个数最少的那个特征。迭代一直进行，直到去掉当前排在最后的那个特征所引起的 SVM 错分样本数大于等于使用全部特征训练时，所得分类器对于测试集的错分样本数时结束。Mundra 和 Rajapakse[149]提出了增强的 SVM-RFE 特征选择算法用于基因选择，将 MRMR 引入到 SVM-RFE 对特征进行排序。Mundra 和 Rajapakse 指出，与 MRMR 或 SVM-RFE 基因选择算法相比，增强 SVM-RFE 基因选择算法在多数基因数据集上选择到更小规模的特征子集；该算法也提供了将 Filter 基因选择算法和 Wrapper 基因选择算法结合的框架。

以上基于 SVM-RFE 框架的基因选择算法归于不同的基因重要度计算原则，特征的剔除原则均以其重要度为依据。然而，基因选择的结果会随每次迭代剔除的基因个数的不同而有差异[150,151]。另外，选择结果受到 SVM 核函数参数的影响[122~124]。还与事先确定的最终选择的基因子集的规模相关[152]。

为此，Tang 等[150]提出了两步 SVM-RFE 基因选择策略，第一步训练多个 SVM-RFE，将各个 SVM-RFE 选择的特征子集做并集，第二步在并集特征上再运行 SVM-RFE 算法，每次迭代只剔除一个基因。两步 SVM-RFE 克服了 SVM-RFE 的特征选择结果对每次迭代剔除基因个数敏感的缺陷。Abeel[151]提出了集成多个基于 Bootstrap 思想构造的 SVM-RFE 基因选择算法用于癌症诊断，提高了基因选择结果的鲁棒性。Tapia 等[140]深入研究了基因选择算法 SVM-RFE 的局限性，提出了利用 SVM-RFE 实现稳定的基因选择的方法。

Zhou 等[74]将 SVM-RFE 特征选择算法推广到多类基因分类问题，基于不同的多类

SVM 框架，提出了四种不同 MSVM-RFE 算法实现多类基因分类问题中的基因选择问题。Zhou 等[153]提出了针对多类分类问题特征排序的集成空间 SVM-RFE 特征选择算法。

以上分析可见，基于 SVM-RFE 的框架的特征（基因）选择方法研究依然倍受关注，构成了基因选择研究的一个基石。这些研究的共同特点是：基于 SVM-RFE 框架，以分类器的反馈信息来剔除最不重要的特征，基于后向搜索，需要的计算量大；且不可避免地存在子集嵌套问题。

目前的研究趋势是：针对 SVM-RFE 的不稳定性而提出的改进基因选择算法研究；基于信息融合思想的集成多个 SVM-RFE 的集成特征选择算法研究；以及针对多类分类问题的 SVM-RFE 特征（基因）选择算法研究。

1.5.3 以函数优化为目标的基于 SVM 的 Embedded 特征选择方法研究进展

该类基于 SVM 的 Embedded 特征选择方法[50~52,68,103,137,138]不同于其他基于 SVM 的特征选择方法之处在于：该方法的特征选择和学习交互作用，学习和特征选择不分开。

Weston 等人[103]的基于 SVM 的 Embedded 特征选择算法将 $R^2\|w\|^2$ 最小化作为目标函数实现特征选择。Miranda 等人[68]提出构造一个线性惩罚的支持向量机（Linearly Penalized Support Vector Machines, LP-SVM），特征选择在构造这个模型的过程中实现。Hochreiter 等[50]提出了 P-SVM（Potential Support Vector Machine）模型，通过求解样本空间的约束优化函数得到特征子集。Li 等[51]提出了 FVM（Feature Vector Machine），通过重构 Lasso 回归模型为 SVM 模型，构造非线性特征核函数实现特征选择。Cheng 等[52]提出了 RFVM（Relevance Feature Vector Machine），在样本空间应用相关向量机（Relevance Vector Machine, RVM），以及特征核函数实现线性相关和非线性相关特征的特征选择。Choi 等[137]提出了一种新的机器学习算法实现基因选择，同时得到高准确率的分类器。Choi 等提出的学习算法通过最小化一个带有拒绝选择的 l_1 范数 SVM，构造了一个分类器，通过参数调整实现分类器的优化，在分类器优化过程中实现显著基因的选择，并得到一个高度精确的预测模型。Li 等[138]引入两个调整参数，定义了自适应的可变网络惩罚项，从而提出 AHSVM（Adaptive Huberized Support Vector Machines，AHSVM），在实现分类的同时实现基因选择。Choi 和 Li 方法都需要较多的参数调整。

基于 SVM 的以函数优化为目标的这类 Embedded 特征（基因）选择算法，构造合适的待优化目标函数是实现特征（基因）选择的关键。如何构造这样的目标函数是一个具有挑战性的问题。

1.5.4 基于 SVM 的混合特征选择算法研究进展

该类特征选择算法集成 Filter 特征选择算法的快速和 Wrapper 特征选择算法的准确与一起，实现高效准确的特征选择。

Liu 等人[107]提出基于 SVM 的混合特征选择算法，先对样本原始特征进行 Filter 算法预处理，再对这些预选择后的特征子集进行 SVM 前向顺序 Wrapper 特征选择算法，在不降低分类准确率的条件下提高了计算效率。Lee 等[135] 提出遗传算法结合卡方检以及 SVM 的混合特征选择算法实现基因选择，用于基因微阵列分类研究。Lee[154]提出 F-score 与前向顺序选择相结合的 SVM 混合特征选择方法，成功应用于股市趋势分析。谢娟英等[119]将度量样本特征区分度的 F-score 方法进行推广，提出推广的 F-score，用于度量多类分类问题的特征区分度，结合前向顺序选择，以 SVM 分类器的分类准确率指导特征选择过程，得到基于 SVM 的 Filter 和 Wrapper 方法相结合的混合特征选择算法；将该算法应用于皮肤病的诊断研究，取得了很好的研究效果[122,123]。在此研究基础上，为了避免类偏斜问题，提出新的准确率定义，进一步提出基于改进 F-score 和新准确率标准的 4 种浮动混合特征选择算法，以 SVM 分类器的分类准确率变化指导特征被选择与否，既克服了顺序特征选择的“子集嵌套”问题，又避免了特征选择的类偏斜问题，在皮肤病诊断研究中取得了更好的效果[124]；同时对针对 F-score 的缺憾，提出 D-score 特征评价准则[120]，并应用于皮肤病的诊断研究[121]。Zheng 等[139]提出应用 *t*-统计和独立变量组分析的混合 SVM 基因选择算法用于肿瘤诊断。Monirul Kabir 等[155]提出了一种新的混合特征选择，以蚁群算法进行最优特征子集的搜索，以学习机评价特征子集。

总之，混合特征选择方法既有 Filter 特征选择方法的优势，又具有 Wrapper 特征选择方法的优势，是一种高效准确的特征选择方法，也是特征选择方法研究的一种新趋势。

1.6 基于 SVM 的特征选择研究趋势

由以上基于 SVM 的特征选择研究进展分析可见，基于 SVM 的特征选择研究可以概括为如下几个研究方向：第一，基于 SVM 的 Filter 型特征选择算法的基因重要性评价准则研究；第二，以 SVM-RFE 为基本框架的基因选择研究；第三，基于信息融合思想的多个基因选择算法的集成实现基因选择；第四，基因选择算法的稳定性或鲁棒性研究；第五，多类分类问题的基因选择算法研究；第六，将 Filter 基因选择算法的基因排序准则引入 SVM-RFE 的混合基因选择研究，以及将 Filter 方法与 Wrapper 方法结合的，以 SVM 分类器的分类准确率指导特征选择过程的混合特征选择研究；第七，基于目标

优化的基因选择算法研究；第八，智能计算理论与 SVM 结合的基因选择算法研究。研究趋势呈现以下几种状态。

1st 集成 Filter 特征选择算法的高效与 Wrapper 特征选择算法的准确的混合特征选择算法研究。

Filter 特征选择算法因为独立于分类器，具有较高的计算效率，Wrapper 特征选择方法依赖于分类器，分类的准确率较高，因此基于 SVM 的特征选择研究趋于集成 Filter 特征选择算法的快速和 Wrapper 特征选择算法的准确与一起，以 Filter 方法的特征重要度评价准则对特征排序，以 SVM 分类器的反馈信息指导 Filter 特征选择过程，或者在 SVM-RFE 特征／基因选择算法中引入 Filter 特征／基因选择算法的特征／基因排序准则对特征／基因进行排序，通过考虑特征／基因之间的相互关系，或者特征／基因与类标的关系等信息，改善特征／基因排序准则，实现更好特征／基因排序，从而实现高效高准确率的基因（特征）选择。 同时，为了确定 SVM 学习机的最佳参数，智能计算方法被引入基于 SVM 的混合基因选择算法研究。

2nd 特征重要性评价准则研究。基因（特征）重要性评价准则在基因（特征）选择中起重要作用。特征选择的目的是选择到重要的起关键作用的特征，剔除冗余的不重要的特征／基因。因此，如何评价特征的重要性，在特征重要性评价中如何考虑特征之间的相关性依然是特征／基因选择研究的热点。另外，特征子集评价受到关注[146]。

3rd 集成多个特征选择算法的融合特征选择算法研究。基于信息融合思想，集成多个特征选择方法与一起进行特征／基因选择，融合各方法的优势，实现更优的特征／基因选择，提高特征／基因选择结果的稳定性和鲁棒性。

4th 与智能方法相结合的特征选择研究。借助智能方法的全局搜索最优解优势，实现最优特征子集选择，以及特征选择过程中 SVM 学习机的最佳参数选择。

5th 构建新的基于样本空间的 SVM 模型，通过求解优化模型直接进行基因（特征）选择。

6th 多类问题的特征／基因选择研究。虽然多数基因数据集是两类分类问题，但是也有一些基因数据集是多类分类问题。如,白血病数据集 Leukemia-MLL[32]就是一个三类分类问题。

1.7 本研究的意义与研究内容

从以上研究进展与趋势的分析可见，我们面临当代分子生物技术发展所带来了高维

小样本癌症基因数据集、网络技术所带来的高维文本数据集，以及各种疾病的高维数据集，对这些高维数据集进行分类分析，特征选择成为首要任务。特征选择已经成为了这些应用研究领域的必要和重要步骤，在构建高准确率的分类系统任务中发挥着不可替代的重要作用。

以 SVM 这一迄今为止的具有最小化分类错误率、最大化泛化能力的分类器为分类工具，以其分类效果为准则指导特征选择过程，进行特征选择，必将选择出最具有区分能力的特征子集，必将在这些应用领域的分类分析研究中发挥重要作用，提高分类的准确率。

然而现有基于 SVM 的特征选择算法研究中，仍有以下问题有待进一步研究。首先，如何选择合适的特征重要性排序准则？如何考虑特征之间的相关性？如何确定最佳的被选择特征数目？以及如何选择合适 SVM 分类器模型、合适的 SVM 参数？在癌症等基因数据集的分类分析中，面对超高维小样本的数据集， 特征搜索过程中，逐个处理单个特征对于这样的数据集显然已经不可能，那么如何进行特征选择？另外，现有基于 SVM 的特征选择方法主要基于后向剔除思想，即基于 SVM-RFE 框架，然而后向剔除相对于前向选择时间效率较差。

为此，本研究针对以上问题，主要研究基于 SVM 的前向特征选择方法，以期探索基于 SVM 的特征选择方法中的特征排序准则、特征相关性等理论意义和应用价值；期望发现分类问题中的关键区分特征，剔除冗余特征，构建高效的分类器模型，提高分类预测的准确率；对癌症等疾病的诊治起到一定的指导作用。

本研究的内容组织如下，第一章，综述基于 SVM 的特征选择研究现状；第二章，论述基于 G-score 与 SVM 的特征选择方法；第三章，论述基于 D-score 与 SVM 的特征选择算法；第四章，阐述基于 DFS 特征子集评价准则与 SVM 的特征选择算法；第五章，论述基于 SVM 分类模型的特征选择算法；第六章，介绍基于 SVM 分类模型的基因选择；第七章, 对本研究进行了总结和展望。

第二章 基于 G-score 与 SVM 的特征选择

本章主要将提出基于 G-score 与 SVM 的三种混合特征选择算法，并以 UCI 机器学习数据库的数据集对算法进行测试。其中，G-score 是 F-score 特征评价准则的推广，是可以用于任意类分类问题的特征评价准则；算法中的特征搜索策略是将几种经典的特征搜索策略进行推广，提出的推广前向顺序搜索 GSFS（Generalized Sequential Forward Search）、推广前向顺序浮动搜索 GSFFS（Generalized Sequential Forward Floating Search）和推广后向顺序浮动搜索 GSBFS（Generalized Sequential Backward Floating Search）。作为基本知识，本章首先介绍支持向量集 SVM（Support Vector Machine）的基本原理。

2.1 SVM 原理

SVM 是 Vapnik 等在统计学习理论的 VC 维和结构风险最小化理论基础上提出的小样本机器学习算法[66, 156]。SVM 在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势，能根据有限的样本信息在模型复杂性（即对特定训练样本的学习精度，Accuracy）和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，获得最好的推广能力。

所谓 VC 维是对函数类的一种度量，函数越复杂，VC 维约高。更直观的定义可以表述为[157]：函数集的 VC 维就是用这个函数集中的函数所能够打散的最大样本集规模。就是说如果一个函数集能够打散的样本集的势是 h，包含 h+1 个样本的样本集不能被该函数集打散，则该函数集的 VC 维就是 h。这里所谓的打散意指函数集中的函数能够将势为 h 的样本集按照所有可能得 2^h 种形式分成两类。

传统机器学习基于使学习机的输出值与真实值之间的误差最小化，即经验误差最小化原则。但是该原则会带来学习机在真实分类时效果很差，也即学习机的推广能力，或泛化能力很差。这种情况就是学习机通过学习得到了一个足够复杂的分类函数，该函数的 VC 维足够高，可以正确分类所有的学习样本，但是对于测试样本却不能进行正确分类。

统计学习理论引入泛化误差界，从而使得：真实风险＝经验风险＋置信风险。其中，经验风险，代表了分类器在给定样本上的误差； 置信风险，代表了可以在多大程度上信任分类器对未知样本的分类结果。显然，置信风险不能精确计算，只能给出一个估计区间。因此整个误差只能计算上界，即泛化误差界。

置信风险与两个两有关，一是学习样本的数量，显然学习样本数量越大，学习结果越有可能正确，此时置信风险越小；二是分类函数的 VC 维，显然 VC 维越大，推广能力越差，置信风险将变大。

这样，统计学习的目标从经验风险最小化变成了经验风险与置信风险和的最小化，也即结构风险最小化。

SVM 的基本思想是：将输入空间线性不可分的样本通过核函数映射到一个高维特征空间，在特征空间求解一个线性约束二次规划，得到一个能将样本线性分割的具有最大间隔的分类超平面。

对于线性可分的两类分类问题，SVM 旨在寻求将两类样本分开且保证分类间隔最大的最优分类面。

假设线性可分样本集为$\{(x_i,y_i)\big|x_i\in R^m, y_i\in\{+1,-1\}, i=1,2,\cdots,n,\}$,SVM 分类器的线性判别函数一般形式为$g(x)=\langle w\cdot x\rangle+b$，相应的分类面为$\langle w\cdot x\rangle+b=0$，要使分类间隔最大化，则必须使分别属于正、负类的样本到分类面的最小距离最大化，因此，最优的分类面一定处在两类的中间。因此，属于正类的样本使得$g(x)\geq+1$；属于负类的样本使得$g(x)\leq-1$。 处在$\langle w\cdot x\rangle+b=+1$上的样本，以及处在$\langle w\cdot x\rangle+b=-1$的样本称为支持向量。

样本x_i到分类面的几何距离为：$\delta_i=\dfrac{1}{\|w\|}\left|g\left(x_i\right)\right|$。那么要使这个距离最大化，即就是需要使$\|w\|$最小化。为了计算方便，改求使$\dfrac{1}{2}\|w\|^2$最小化。因此，寻求最优分类面转化为下面的优化问题：

$$\begin{aligned}&\min\varphi(w)=\frac{1}{2}\|w\|^2\\&s.t.\quad y_i[\langle w\cdot x_i\rangle+b]-1\geq0, i=1,2,\cdots,n\end{aligned}\tag{2-1}$$

式（2-1）的对偶优化为

$$\begin{aligned}&\max Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\langle x_i\cdot x_j\rangle\\&s.t.\quad\sum_{i=1}^{n}\alpha_i y_i=0\\&\qquad\alpha_i\geq0, i=1,2,\cdots,n\end{aligned}\tag{2-2}$$

求解式(2-2)，可得到最优分类函数为

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i y_i\langle x_i\cdot x\rangle+b\} \tag{2-3}$$

对应 $\alpha_i\neq 0$ 的样本称为支持向量。

在线性不可分情况下，引入松弛变量 ξ_i 和惩罚参数 C，式(2-1)转化为

$$\begin{aligned}&\min\varphi(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i\\&s.t.\qquad y_i[\langle w\cdot x_i\rangle+b]\geq 1-\xi_i,i=1,2,\cdots,n\\&\qquad\quad \xi_i\geq 0,i=1,2,\cdots,n\end{aligned} \tag{2-4}$$

其对偶问题为

$$\begin{aligned}&\max Q(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\langle x_i\cdot x_j\rangle\\&s.t.\qquad \sum_{i=1}^{n}\alpha_i y_i=0\\&\qquad\quad C\geq\alpha_i\geq 0,i=1,2,\cdots,n\end{aligned} \tag{2-5}$$

对于非线性可分的分类问题，可通过定义适当的核函数实现非线性变换，将输入空间映射到一个高维特征空间，在特征空间求解最优线性分类面。引入核函数后，以上的内积可用核函数代替。此时，SVM 的决策函数为

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b\} \tag{2-6}$$

不同的内积核函数导致不同的支持向量机算法，目前采用的内积核函数主要有以下四类：

(1) 线性核函数 $K(x,x')=x\cdot x'$；

(2) 多项式核函数函数 $K(x,x')=(x\cdot x'+1)^d$；

(3) 径向基核函数 $K(x,x')=\exp(-\|x-x'\|^2/\sigma^2)$；

(4) S 型核函数 $K(x,x_i)=\tanh(v(x\cdot x_i)+c)$。

对于多类分类问题，SVM 通过将其转换为若干个两类分类问题来实现分类。最经典的方法是一对其余与一对一方法。前者将 l 类分类问题，转换成 l 个两类分类问题，第 i

个两类分类以第 i 类为正类，其余 $l-1$ 类构成负类；样本 x 属于 l 个分类模型输出值最大的那个模型的正类。后者构造 $\frac{1}{2}l(l-1)$ 个分类模型，对于待分类样本 x 每个分类器有一个输出，对应样本 x 属于某个类别，统计样本 x 属于各类别的次数，则样本 x 最终 属于统计次数最大的那个类别。

2.2 特征搜索策略

全局最优搜索、随机搜索和启发式搜索三种特征子集搜索策略中启发式搜索能够达到和前两种搜索策略类似的效果，且具有运算速度快的特点[84]。顺序前向搜索（Sequential Forward Search， SFS）、顺序后向搜索（Sequential Backward Search， SBS）、顺序前向浮动搜索（Sequential Forward Floating Search， SFFS）和顺序后向浮动搜索（Sequential Backward Floating Search， SBFS）是四种常用的启发式搜索策略。

SFS[16]是一种自下而上的搜索方法。该方法从空集开始，首先选择最好的一个特征加入被选择特征子集，然后选择最好的一对特征加入被选择特征子集，其中这个最好的特征对里包含已经入选的那个最好特征。该过程一直进行，每次从未被选择的特征中选择一个最好的特征，该最好的特征和已被选择的特征子集组合将具有最好的分类特性，直到达到指定数目特征或者满足其他搜索停止条件。

与SFS相反，SBS[17]是一种自上而下的搜索方法。该方法从特征全集开始，首先选择最差的一个特征剔除，然后从余下的特征中再选择最差的一个特征剔除。该过程一直进行，每次从剩余特征构成的特征子集中选择最差的一个特征剔除，直到达到指定数目特征或者满足其他搜索停止条件。具体实现过程是：假设特征全集包含有 m 个特征，相应的准则函数是 J，则考查 m 个包含 $m-1$ 个特征的特征子集，计算 $J_i(m-1), i=1,2,\cdots,m$，保留最好的一组特征；然后考察 $m-1$ 个包含 $m-2$ 个特征的特征子集，保留最好的一组；该过程一直进行，直到达到指定数目的特征或者满足某种停止准则。

SFS和SBS构成两种最常用的贪婪搜索策略。他们的共同缺点是：“子集嵌套”，即一旦某个特征被选中，或者被剔除，之后就没法剔除或者再加入该特征，这样仅能得到一个局部最优的特征子集。

Pudil等[18] 提出的顺序前向浮动搜索（SFFS）和后向浮动搜索（SBFS）分别克服了SFS方法和SBS方法的特征子集嵌套包含缺陷。SFFS方法也是一种自底向上的搜索方法，

该方法根据SFS思想逐个加入特征到被选特征子集，每加入一个特征后，进行有条件的剔除，从那些之前加入的特征里选择最不重要的一个特征，若剔除该特征使得准则函数提高，则剔除该特征，直到被选特征子集只剩下两个特征，或者不再存在可以剔除的特征；然后再继续根据SFS加入下一个特征。该过程一直进行， 直到所有特征都被加入。

SBFS是一种自顶向下的搜索过程，该方法首先应用SBS剔除掉两个最不重要的特征，得到包含 $m-2$ 个特征的特征子集；然后从被剔除的特征中选择最重要的特征加入，也即选择与当前特征子集组合使得准则函数升高的特征加入，直到被剔除的特征只剩下两个，或者不存在符合条件的能被加入的特征；则继续应用SBS剔除当前最不重要的特征，之后再从已经删除的特征中选择与当前特征子集组合使得准则函数升高的特征加入，直到被剔除的特征只剩下两个，或者不存在符合条件的能被加入的特征；再应用SBS从当前特征子集剔除最不重要的特征。该过程一直进行，直到所有特征都被考察过。

SFS因为从空集开始，其计算效率通常高于SBS。因此，本研究主要采用SFS和SFFS搜索策略进行特征选择。

2.3 传统 F-score 特征重要性评价准则

传统 F-score 特征评价准则由 Chen 等人提出[105]，是一种基于类间类内距离的特征重要性评价准则，能有效衡量样本特征在两类分类问题中对于实现正确分类的贡献大小，也即特征在类别间辨别能力的大小，是一种简单有效的特征选择方法。其描述如下。

给定训练样本集 $\{(x_k,y_k)\mid x_k\in R^m,\ m\succ 0,\ y_k\in\{1,-1\},\ k=1,2,\cdots,n\}$。其中 m 表示样本空间维数，即每个样本的特征数，$\left\|\{y_k\,|\,y_k=+1,k=1,2,\cdots,n\}\right\|=n_+$，$\left\|\{y_k\,|\,y_k=-1,k=1,2,\cdots,n\}\right\|=n_-$，即正类和负类的样本数分别为 n_+ 和 n_-。则样本第 $i\,(i=1,2,\cdots,m)$ 个特征的区分度 F-score 定义为：

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

其中 $\bar{x}_i$，$\bar{x}_i^{(+)}$，和 $\bar{x}_i^{(-)}$ 分别为第 i 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；$x_{k,i}^{(+)}$ 为正类第 k 个样本点的第 i 个特征的特征值；$x_{k,i}^{(-)}$ 为第 k 个负类样本的第 i 个特征的特征值。F_i 值越大，表明第 i 个特征的辨识力越强。

2.4 推广 F-score 特征重要性评价准则——G-score 准则

传统 F-score 只能用来度量两类分类问题中样本特征区分能力的大小，为了度量多类分类问题中特征辨别能力的大小，我们将传统 F-score 进行推广，提出推广的 F-score 特征重要性评价准则，并称其为 G-score 特征评价准则。G-score 可用于衡量多类分类问题中，样本特征对于实现正确分类的贡献大小。G-score 的定义如下。

给定训练样本集 $\{(x_k, y_k) \mid x_k \in R^m, m \succ 0, y_k \in \{1,2,\cdots,l\}, l \geq 2, k = 1,2,\cdots,n\}$。其中 m 表示样本空间维数，$\left\|\{y_k \mid y_k = j, k = 1,2,\cdots,n\}\right\| = n_j$，$j = 1,2,\cdots,l$，即第 j 类的样本数为 n_j。则样本第 i $(i = 1,2,\cdots,m)$ 个特征的区分度 G-score 定义为：

$$F_i = \frac{\sum_{j=1}^{l} (\bar{x}_i^{(j)} - \bar{x}_i)^2}{\sum_{j=1}^{l} \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (x_{k,i}^{(j)} - \bar{x}_i^{(j)})^2} \tag{2-8}$$

其中 $\bar{x}_i$，$\bar{x}_i^{(j)}$ 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 个特征的特征值。分析式（2-8），分子表示 l 个类中各类的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(2-8)的分母表示各个类的类内方差。方差越小，类内越聚。因此，F_i 近似表示了 i 第个特征的类间、类内方差之比，其值越大表明第 i 个特征的辨识力越强。

2.5 基于 G-score 与 SVM 的混合特征选择方法

如 1.6.4 和 1.7 所述，集成 Filter 特征选择方法的高效与 Wrapper 特征选择方法的高分类准确率优势，研究混合的特征选择方法已经成为基于 SVM 的特征选择方法研究的趋势。同时，基于 SVM 与 G-score 的 Filter 特征选择方法，通过剔除 G-score 值较低的特征来实现样本特征的选择，或者选择 G-score 较高的特征来进行特征选择，没有考虑分类器 SVM 的分类准确率对于特征选择过程的引导作用，从而很难实现最优特征子集的选择。 因此，本研究将推广的 F-score，即 G-score 与 SVM 结合，提出基于 G-score 与 SVM 的混合特征选择方法，以便能实现更好的特征子集选择，构造分类准确率更优的分类器，从而得到更好的分类模型。

基于 G-score 和 SVM 的特征选择方法，首先计算每个特征的 G-score 值，以此来衡量各特征对于分类的贡献大小，G-score 值越大的特征辨别能力越强，对分类的贡献就越大。根据各特征的 G-score 值，将所有特征降序排序，分别采用 GSFS，GSFFS 和 GSBFS

特征搜索策略，以 SVM 为分类工具指导特征选择过程，从而实现特征选择。算法整体思想如图 2.1 所示。

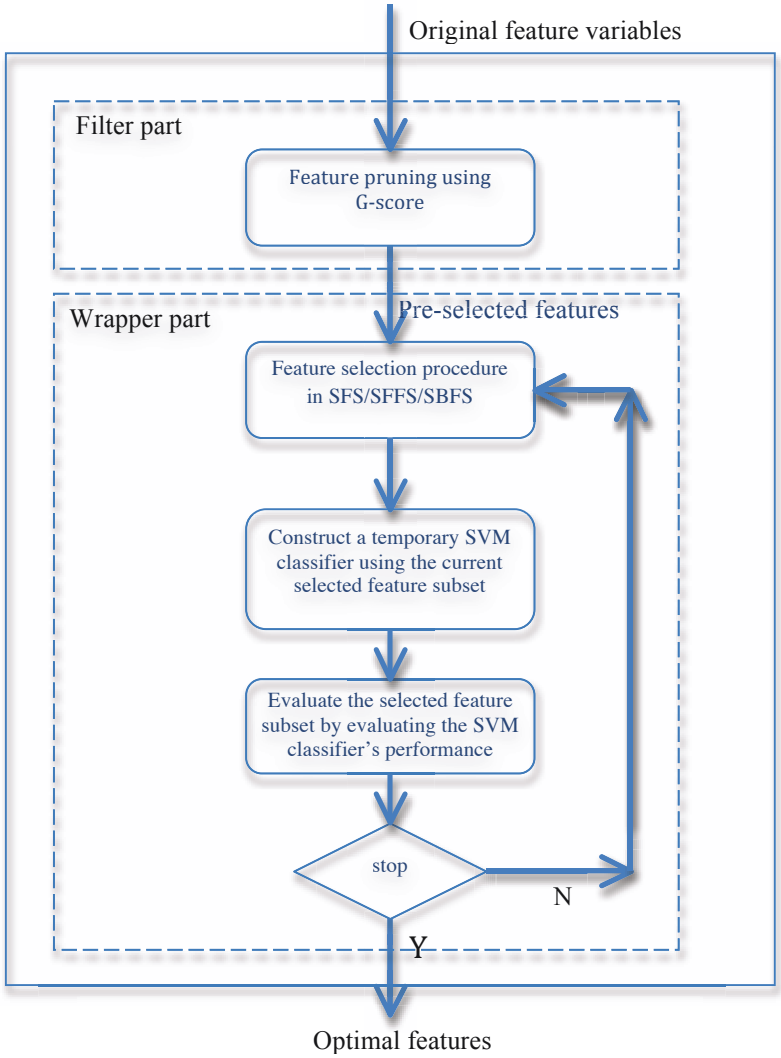


图 2.1 基于 G-score 与 SVM 的混合特征选择模型

Figure 2.1 the hybrid feature selection model based on G-score and SVM

2.5.1 推广的 SFS、SFFS 和 SBFS 特征搜索策略

原始 SFS 如 2.2 所描述，其主要特点是，下一个特征的选择，一定是选择与已入选特征组合具有最优分类性能的特征。本研究使用的 SFS 不同于 2.2 节所描述的传统 SFS。此处的前向顺序搜索 SFS， 是首先选择最优的特征，也即 G-score 值最高的特征，然后在后续循环迭代中，每次选择当前最优的特征，也即在未被选择的特征中，G-score 值最高的特征。我们称其为推广的 SFS，即 generalized sequential forward search，简称为 GSFS。

本研究使用的 SFFS 也不同于 2.2 节所描述的 Pudil 提出的 SFFS。我们称其为推广

的 SFFS，即 generalized sequential forward floating search，简称为 GSFFS。其思想是：首先加入最优的一个特征到被选择特征子集，当用 G-score 作为特征重要性度量准则时，就是加入 G-score 值最大的那个特征；下一个特征的选择过程是，首先进行试加入，将当前未被选择的特征中最重要的特征试加入被选择特征子集，以该特征子集构造分类器，其分类性能若提高，则加入该特征到被选特征子集，否则不加入；继续判断下一个特征。该过程一直进行，直到所有的特征都判断结束。最后被选特征子集中的特征即为所选择的特征，以这些特征表达原空间的样本，不但可以保持原空间样本之间的可区分性，而且降低了空间的维数，降低了存储开销，同时降低了测量开销。

如同推广的 SFFS—GSFFS 不同于 Pudil 提出的 SFFS，本研究使用的 SBFS 也不同于 Pudil 等提出的 SBFS[18]。同样地，我们称其为推广的 SBFS—GSBFS（Generalized Sequential Backward Floating Search）。其思想是：首先对全部特征按照某一评价准则进行降序排序，剔除最差的那个特征。对于下一个最差的特征，先进行试剔除，如果剔除该特征后，训练所得分类器的分类准确率上升，则从特征集中剔除该特征，否则保留该特征在特征子集中，而不剔除。该过程一直进行，直到所有的特征都判断结束。最后保留下来的特征即为所选择的特征。

2.5.2 基于 G-score 与 SVM 的前向顺序混合特征选择

该算法以 G-score 度量每个特征的区分能力，也即对分类的贡献。因此，首先计算每个特征的 G-score 值，然后根据各特征的 G-score 值将特征降序排序，采用推广的前向顺序搜索策略 GSFS 进行有效特征的选择，并以 SVM 作为分类工具，衡量当前被选特征子集的分类性能，以引导特征选择过程。算法的详细步骤以及实验结果与分析分别描述如下。

2.5.2.1 算法步骤描述

step 1：计算每个特征的 G-score 值 F_i，并初始化被选特征子集为空集，即使 selected_feature_subset=Φ 。

step 2：对全部特征根据其 G-score 值降序排序，并构成备选特征子集 candidate_feature_subset。

step 3：若备选特征子集不空，则从备选特征子集的特征中选择当前第一个特征 f_i，也即 G-score 值最大的特征，添加到被选特征集合，并从备选特征子集删掉 f_i 特征。即

if $candidate\_feature\_subset \neq \Phi$ then

$f_i = \max\{F_i \mid G-score(f_i) = F_i, f_i \in candidate\_feature\_subset\}$，

$selected\_feature\_subset = selected\_feature\_subset + \{f_i\}$，

$candidate\_feature\_subset = candidate\_feature\_subset - \{f_i\}$。

step 4：对只含有被选特征子集特征的训练集样本进行 10 折交叉验证训练；以只含有被选特征子集特征的训练集样本的 10 折交叉验证实验的平均分类正确率作为当前被选特征子集分类性能的评价。

Step 5: Goto step 3，直到备选特征子集为空，即知道 candidate_feature_subset＝Φ。

最后选择分类正确率最高，且对应特征子集规模最小的特征子集为被选择的最优特征子集，构造相应的 SVM 分类器，得到只包含最优特征子集特征的分类模型。以该分类模型对测试集样本进行测试，以测试集的分类准确率作为该模型的性能评价，也即被选择特征子集性能的最终评价。

该算法通过将 G-score 和 SVM 相结合，以 G-score 判断每个特征对于分类的贡献，以 SVM 为分类工具，并以其对训练集样本的分类准确率评价相应特征子集，选择分类效果最佳的特征子集， 实现了最佳特征子集的选择。以下 UCI 机器学习数据库的数据集实验测试，证明了该算法的有效性。

2.5.2.2 实验结果与分析

为了验证基于 G-score 与 SVM 的前向顺序混合特征选择算法的性能，我们采用 UCI 机器学习数据库[158]中的 Pima Indians Diabetes，New-thyroid disease，Contraceptive Method Choice，Wine，Dermatology 和 Statlog (Landsat Satellite) 六组数据集对该特征选择方法进行测试。数据集描述见表 2.1。实验中训练数据集与测试数据集采用 6：4 的比例进行随机划分；对于 UCI 数据库中已经分好训练集、测试集的 Statlog (Landsat Satellite) 数据集，直接采用现有的训练集、测试集进行实验。利用训练数据集进行最优特征子集选择，测试数据集对所选择的最优特征子集的分类有效性进行评价。表 2.2 为各特征根据其 G-score 值进行降序排序后的序列，表 2.3 为最终选取的各数据集的最优特征子集。

为了进一步说明 G-score 与 SVM 混合特征选择方法的有效性，将该特征选择方法的实验结果与没有经过特征选择，直接使用 SVM 方法进行分类的分类结果，以及应用

PCA 进行降维，然后使用 SVM 方法进行分类的实验结果进行比较，表 2.4 所示为三种方法的分类正确率比较，表 2.5 所示为三种方法的训练时间比较。

表 2.1 实验所用 UCI 数据集描述

Table 2.1 the description of data sets from UCI machine learning repository

数据集	样本个数	特征数	类别数
Pima Indians Diabetes	768	8	2
New-thyroid disease	215	5	3
Contraceptive Method Choice	1473	9	3
Wine	178	13	3
Dermatology	358	34	6
Statlog (Landsat Satellite)	6435	36	7

表 2.2 各特征的 G-score 值排序结果

Table 2.2 the order of features of data sets according to their G-score value

数据集	原始特征数	根据 G-score 值降序排序的特征号
Pima Indians Diabetes	8	2, 6, 8, 1, 7, 5, 3, 4
New-thyroid disease	5	2, 4, 5, 3, 1
Contraceptive Method Choice	9	2, 8, 1, 7, 3, 4, 5, 9, 6
Wine	13	7, 11, 12, 13, 1, 6, 2, 10, 9, 4, 8, 3, 5
Dermatology	34	31, 30, 33, 7, 27, 15, 29, 6, 12, 25, 8, 21, 20, 22, 28, 9, 16, 10, 14, 34, 5, 3, 11, 4, 24, 23, 19, 2, 26, 1, 17, 32, 18, 13
Statlog(Landsat Satellite)	36	18, 17, 22, 14, 21, 6, 20, 13, 30, 2, 29, 24, 34, 5, 16, 33, 1, 26, 25, 10, 8, 9, 12, 32, 4, 28, 36, 19, 15, 23, 7, 31, 11, 27, 3, 35

表 2.3 各数据集的最优特征子集

Table 2.3 the optimal feature subsets of data sets

数据集	所选择特征数	最优特征子集所包含的特征
Pima Indians Diabetes	3	2, 6, 8
New-thyroid disease	3	2, 4, 5
Contraceptive Method Choice	6	2, 8, 1, 7, 3, 4
Wine	5	7, 11, 12, 13, 1
Dermatology	21	31, 30, 33, 7, 27, 15, 29, 6, 12, 25, 8, 21, 20, 22, 28, 9, 16, 10, 14, 34, 5
Statlog(Landsat Satellite)	10	18, 17, 22, 14, 21, 6, 20, 13, 30, 2

表 2.4 G-score＋SVM、SVM、PCA＋SVM 三种方法的分类正确率比较

Table 2.4 the accuracy of G-score+SVM, SVM and PCA+SVM

数据集	分类正确率（%）		
	SVM	PCA+SVM	**G-score +SVM**
Pima Indians Diabetes	75.97	75.32	**79.55**
New-thyroid disease	88.37	90.70	**97.67**
Contraceptive Method Choice	77.83	77.33	**79.36**
Wine	91.78	64.38	**98.63**
Dermatology	83.33	76.71	**98.61**
Statlog (Landsat Satellite)	**100**	**100**	**100**

表 2.5 G-score＋SVM、SVM、PCA＋SVM 三种方法的训练时间比较

Table 2.5 the training time of G-score+SVM, SVM and PCA+SVM

数据集	训练时间（s）		
	SVM	PCA+SVM	**G-score +SVM**
Pima Indians Diabetes	2.172	5.485	**0.391**
New-thyroid disease	0.032	0.046	**0.016**
Contraceptive Method Choice	21.469	72.641	**20.235**
Wine	**1.25**	22.688	**2.242**
Dermatology	0.328	0.547	**0.297**
Statlog (Landsat Satellite)	**18.937**	**12.312**	**12.078**

表 2.4 实验结果表明，基于 G-score 与 SVM 的特征选择方法获得了最好的识别结果，尤其对于 Statlog （Landsat Satellite）数据集，基于 G-score 与 SVM 的混合特征选择方法不足 1 / 3 的特征就达到了 100％的识别率。对于其他 5 各数据集，应用经过 G-score 与 SVM 结合的特征选择方法进行选择后的特征构造的分类器，分类准确率都得到提高，特别是对于皮肤病数据集 Dermatology，诊断准确率得到大幅度第提升。因此，基于 G-score 与 SVM 的前向顺序混合特征选择方法具有很好的泛化能力。

表 2.5 的训练时间比较，可见除 Wine 数据集外，其他五个数据集上的训练时间都优于 PCA+SVM 方法，以及直接使用 SVM 的方法。在 Wine 数据集上，基于 G-score 与 SVM 的前向顺序混合特征选择方法的训练时间远小于 PCA+SVM 方法。

以上分析表明，基于 G-score 与 SVM 的前向混合特征选择算法实现了有效的特征选择，提高了分类器的分类准确率。

2.5.3 基于 G-score 与 SVM 的前向顺序浮动混合特征选择

前向顺序搜索有“子集嵌套包含”的缺陷。某个特征一旦被选中，后继的搜索就没有办法剔除该特征。因此，基于G-score与SVM的前向顺序特征选择仅能得到一个局部最优的特征子集。为了尽可能得到全局最优的特征子集，本研究进一步提出基于G-score与SVM的前向顺序浮动混合特征选择算法。此处的前向浮动搜索不是传统的SFFS，而是2.5.1提出的推广前向浮动搜索GSFFS。算法主要思想是：每次加入某一个特征后，若训练集的分类正确率没有上升，则不加入当前选择的特征；否则就加入该特征。然后尝试加入下一个最好的特征。直到所有的特征被扫描一遍结束。算法详细步骤与实验结果分别描述如下。

2.5.3.1 算法步骤描述

step 1：以训练样本集计算每个特征的 G-score 值，并依据 G-score 值将所有特征降序排序。

step 2：初始化被选特征子集为空集，备选特征子集包含全部特征。即使得

$selected\_feature\_subset = \Phi$,

$candidate\_feature\_subset = \{f_i | f_i \in original\_featreu\_set\}$。

step 3：若备选特征子集不空，则从备选特征子集中选择第一个特征，也即 G-score 值最大的特征，添加到被选特征集合，并从备选特征子集删掉该特征。即，

if $candidate\_feature\_subset \neq \Phi$ then

$f_i = \max\{F_i | G-score(f_i) = F_i, f_i \in candidate\_feature\_subset\}$，

$selected\_feature\_subset = selected\_feature\_subset + \{f_i\}$，

$candidate\_feature\_subset = candidate\_feature\_subset - \{f_i\}$。

step 4：以只含有被选特征子集特征的训练集样本 5 折交叉验证训练 SVM，构造相应的 SVM 分类器，得到临时 SVM 分类模型；以只含有被选特征子集特征的训练集样本测试这个临时 SVM 分类模型，以其分类准确率评价当前被选择特征子集的性能。

step 5：若备选特征子集不空，则从备选特征子集的特征中选择当前第一个特征，试添加到被选特征集合，构成临时的被选择特征子集。即

if $candidate\_feature\_subset \neq \Phi$ then

$f_i = \max\{F_i | G-score(f_i) = F_i, f_i \in candidate\_feature\_subset\}$，

$temp\_selected\_feature\_subset = selected\_feature\_subset + \{f_i\}$，

Step 6：以只含有临时被选特征子集特征的训练样本 5 折交叉训练 SVM 分类器，构造临时的 SVM 分类模型；以训练集集样本来测试这个临时 SVM 分类模型的分类性能。若这个临时 SVM 的分类准确率提高，则将试加入的特征加入被选特征子集；否则，舍弃该特征，不加入其到被选特征子集。即，

if $current\_accuracy > pre\_accuracy$ then

$selected\_feature\_subset = selected\_feature\_subset + \{f_i\}$。

step 7：从备选特征子集删去该特征。

$candidate\_feature\_subset = candidate\_feature\_subset - \{f_i\}$。

step 8：转 step 5，直到备选特征子集为空。

最后留在被选特征子集中的特征即为所选择的特征。以该特征子集构造相应 SVM 分类模型，以其对测试集样本的分类准确率来评价该模型。该特征选择方法能够在一定程度上避免陷入局部最优解，同时选取的特征数目通常会比 GSFS 策略选取的特征数目少。

2.5.3.2 实验结果与分析

为了测试基于 G-score 与 SVM 的前向浮动特征选择算法的性能，采用 UCI 机器学习数据库中的 9 个数据集 dermatology，glass，handwrite，Ionosphere，WDBC (Wisconsin Diagnostic Breast Cancer，WDBC)，WPBC (Wisconsin Prognostic Breast Cancer, WPBC)，wine，thyroid-disease，和 heart disease。实验所用的各数据集描述如表 2.6 所示。其中，dermatology 数据集，删去了 8 个含有缺失数据的样本，因此总的数据集规模由原来的 366 变成了 358；glass 数据集分成了 window glass 和 non-window glass 两类；handwrite，即 Semeion Handwritten Digit 数据集，只选择了前两类进行实验；WPBC 数据集删去了 4 个含有缺失数据的样本，因此数据集规模为 194；关于 thyroid-disease 数据集，实验使用了其中的 new-thyroid，也即 Thyroid gland data 数据集；Heart Disease 数据集使用的是其中的 processed cleveland 数据集，实验中删去了 6 个含有缺失数据的样本，因此数据集规模为 297。

表 2.6 实验所用 UCI 数据集描述

Table 2.6 the description of data sets from UCI machine learning repository

数据集	样本个数	特征数	类别数
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

为了得到具有统计意义的实验结果，我们采用五折交叉验证的方法进行实验。将样本顺序随机打乱；然后对每一类的样本依次逐个加入到五个样本集合中（初始时样本集合为空），直到这一类的每一个样本都被加入。从而实现了样本均匀划分为五份的目的，以每一份作测试样本，其余四份作训练样本，实现五折交叉验证实验。

实验中支持向量机的核函数采用径向基核函数 RBF[159]，为了得到具有较好推广性

能的 SVM 分类模型，需要选择合适的 SVM 惩罚因子C 和 RBF 的核函数参数γ 。为此，采用网格搜索和 5-折交叉验证相结合的方法，来选择最优的参数C和γ。其中，$C\in\{2^{-5},\cdots,2^{15}\}$， $\gamma\in\{2^{-15},\cdots,2^{5}\}$。实验中对每一组参数组合，训练集用 5-折交叉验证确定当前C和γ对应的分类正确率平均值，最后选取训练集平均分类正确率最佳的一组C和γ做为 SVM 的最佳参数，构建具有最大间隔分类面的 SVM 分类模型。以此模型对相应的训练集、测试集进行测试，得到训练集、测试集的分类正确率。实验使用的 SVM 工具箱为台湾林智仁教授等开发的 LibSvm 工具箱[160]。

下面是基于 G-score 与 SVM 的前向顺序浮动搜索算法的实验结果。为了说明该算法的有效性，此处将该实验结果与相同实验设计下的基于 G-score 与 SVM 的前向顺序实验结果进行了比较。表 2.7 是所选择特征数的比较。表中所选特征的数目以平均特征数（最小值，最大值）的形式给出，特征数目中带下划线的表示的是特征数目较小。表 2.8 是训练集与测试集的分类正确率比较。分类正确率中带下划线的表示分类正确率较高。

表 2.7 G-score＋SVM 的前向顺序浮动与前向顺序特征选择所选择特征数比较
Table 2.7 the size of the selected feature subset of G-score and SVM with GSFS and GSFFS search strategy respectively

数据集	原始特征数	G-score+SVM+GSFS	G-score+SVM+GSFFS
dermatology	34	24	12.4 (11, 18)
glass	9	4	4.4 (3, 5)
handwrite	255	13	7.6 (7, 9)
Ionosphere	34	8	11.8 (10, 15)
WDBC	30	11	9 (4, 13)
WPBC	33	10	7.4 (5, 12)
wine	13	7	5.6 (5, 7)
thyroid-disease	5	4	3.6 (3, 4)
Heart Disease	13	7	5.4 (4, 6)

表 2.8 基于 G-score 与 SVM 前向顺序浮动与前向顺序特征选择的分类正确率
Table 2.8 the accuracy of G-score and SVM with GSFS and GSFFS search strategy respectively

数据集	G-score+SVM+GSFS		G-score+SVM+GSFFS	
	训练集(%)	测试集(%)	训练集(%)	测试集(%)
dermatology	98.8121	98.3245	98.9515	98.0463
glass	98.3619	92.5451	99.4166	92.0477
handwrite	99.3041	98.2209	99.6908	98.4471
Ionosphere	95.7989	92.0282	99.8576	92.8853
WDBC	97.6706	95.2520	99.3842	95.0873
WPBC	87.5068	74.7699	93.1694	70.1667
wine	99.1588	97.1569	100	99.4444
thyroid-disease	98.8372	95.8140	99.0698	95.8139
Heart Disease	85.1016	81.1525	90.1489	76.0734

表 2.7 的实验结果显示基于 G-score 与 SVM 的前向浮动特征选择算法除了在 Ionosphere 数据集上所选择的平均特征数比基于 G-score 与 SVM 的前向顺序特征选择算

法略多之外，在其他 8 个数聚集上所选择的平均特征数都比基于 G-score 与 SVM 的前向顺序特征选择算法少。同时，表 2.8 的分类正确率比较显示，基于 G-score 与 SVM 的前向浮动特征选择算法在所有训练集上的分类正确率都优于基于 G-score 与 SVM 的前向顺序特征选择算法；但是测试集的分类正确率比较显示：基于 G-score 与 SVM 的前向浮动特征选择算法在 dermatology，glass，WDBC，WPBC，Heart Disease 五个数据集上略低于基于 G-score 与 SVM 的前向顺序特征选择算法；其他四个数据集上基于 G-score 与 SVM 的前向浮动特征选择算法不仅实现了特征维数的降低，而且分类正确率也得到提高。

以上分析得出：基于 G-score 与 SVM 的前向浮动特征选择算法相对基于 G-score 与 SVM 的前向顺序特征选择算法在所选择的特征子集规模上具有更优的结果；然而以所选择特征子集构造的分类器的泛化性能与前者相比略有降低；特别是对疾病诊断数据集，从特征选择后的分类器的泛化性能来看，基于 G-score 与 SVM 的前向顺序特征选择算法要明显地好。

2.5.4 基于 G-score 与 SVM 的后向顺序浮动特征选择

为了进一步说明基于 G-score 与 SVM 的前向顺序特征选择，以及基于 G-score 与 SVM 的前向顺序浮动特征选择方法的有效性，本部分将研究基于 G-score 与 SVM 的后向顺序浮动特征选择方法。此处使用的后向浮动搜索不是 Pudil 提出的 SBFS[18]，而是 2.5.1 提出的推广后向浮动搜索 GSBFS。算法实现步骤和实验结果分别如 2.5.4.1 和 2.5.4.2 描述。

2.5.4.1 算法步骤描述

step 1：以训练样本集计算各特征的 G-score 值，并对特征进行降序排序，以全部特征构成被选特征子集和备选特征子集。即，初始化备选特征子集和被选特征子集如下：

$$selected\_feature\_subset = candidate\_feature\_subset = full\_feature\_set.$$

step 2：对包含被选特征子集特征的训练样本进行 5 折交叉验证训练，也即使用包含全部特征的训练样本训练 SVM 分类器，得到一个临时 SVM 分类器，测试该模型对训练集和测试样本集样本的分类正确率 。

step 3：从被选特征子集中试删除当前备选特征子集的最后一个特征对应的特征。

step 4：对只含有被选择特征的训练集样本进行学习，得到新的临时 SVM 分类器；以该 SVM 分类模型对训练集和测试集样本进行测试，记录相应的分类正确率。

step 5：若训练集的分类准确率提高，则从被选特征子集删除该特征，否则不删去该特征。

step 6：将该特征从备选特征子集删除。

step 7：转 step 3，直到备选特征子集为空。

最后被选特征子集中的特征即为所选择的特征。该特征选择算法避免了顺序后向特征搜索策略的“子集嵌套”缺陷，部分考虑了特征之间的相关性，使得最终得到的特征子集避开了局部最优的缺憾。

2.5.4.2 实验结果与分析

为了和基于 G-score 与 SVM 的前向顺序浮动、前向顺序两种搜索策略的特征选择算法进行比较，本部分实验采用和2.5.3 部分相同的实验数据集。数据集描述如表2.6所示。实验同样也采用五折交叉验证方法，并在同样划分的数据集上进行。实验中支持向量机的核函数依然采用径向基核函数 RBF[159]，同样对于核函数参数C和γ的选择也采用在训练集上的网格搜索和 5-折交叉验证相结合的方法进行。其中，参数C和γ的搜索范围和 2.5.3 相同，$C\in\{2^{-5},\cdots,2^{15}\}$，$\gamma\in\{2^{-15},\cdots,2^{5}\}$。对每一组参数组合，根据训练集在当前 C和γ组合的 5-折交叉验证的平均分类正确率，选取平均分类正确率最佳的一组C和γ为 SVM 的最佳参数，用训练集样本构建具有最大间隔分类面的 SVM 分类模型。以此模型对相应的训练集、测试集进行测试，得到训练集分类正确率、测试计分类正确率。实验使用相同的 SVM 工具箱[160]。

下面是基于 G-score 与 SVM 的后向顺序浮动搜索算法的实验结果。表 2.9 给出了基于 G-score 与 SVM，并分别以 GSFS、GSFFS 和 GSBFS 三种搜索策略进行特征选择过程中的特征搜索的特征选择算法所选择特征数的比较。表中所选特征的数目以平均特征数（最小值，最大值）的形式给出，特征数目中带下划线的表示的是特征数目较小。表 2.10 展示了 3 中不同的基于 G-score 与 SVM 的特征选择算法所构造的分类模型的对应的训练集与测试集的分类正确率比较。分类正确率中带下划线的表示分类正确率较高。

表 2.9 G-score 与 SVM 结合三种不同搜索策略的特征选择算法所选特征子集规模比较
Table 2.9 the size of the selected feature subset of G-score and SVM with GSFS, GSFFS and GSBFS search strategy respectively

数据集	原特征数	G-score+SVM+GSFS	G-score+SVM+GSFFS	G-score+SVM+GSBFS
dermatology	34	24	12.4 (11, 18)	14.2 (9, 17)
glass	9	4	4.4 (3, 5)	3.8 (3, 5)
handwrite	255	13	7.6 (7, 9)	13.6 (8, 23)
Ionosphere	34	8	11.8 (10, 15)	14.6 (10, 21)

WDBC	30	11	9 (4, 13)	20 (15, 23)
WPBC	33	10	7.4 (5, 12)	25 (19, 29)
wine	13	7	5.6 (5, 7)	5.8 (4, 7)
thyroid-disease	5	4	3.6 (3, 4)	3.6 (3, 4)
Heart Disease	13	7	5.4 (4, 6)	9.6 (6, 12)

表 2.10 基于 G-score 与 SVM 三种特征选择算法的分类正确率比较

Table 2.10 the accuracy of G-score and SVM with GSFS, GSFFS and GSBFS search strategy respectively

数据集	G-score+SVM+GSFS		G-score+SVM+GSFFS		G-score+SVM+GSBFS	
	训练集(%)	测试集(%)	训练集(%)	测试集(%)	训练集(%)	测试集(%)
dermatology	98.8121	98.3245	98.9515	98.0463	99.5814	94.1169
glass	98.3619	92.5451	99.4166	92.0477	99.5335	89.7206
handwrite	99.3041	98.2209	99.6908	98.4471	100	98.1489
Ionosphere	95.7989	92.0282	99.8576	92.8853	99.8574	92.8773
WDBC	97.6706	95.2520	99.3842	95.0873	99.1211	96.6556
WPBC	87.5068	74.7699	93.1694	70.1667	94.8369	74.7031
wine	99.1588	97.1569	100	99.4444	99.8612	89.3800
thyroid-disease	98.8372	95.8140	99.0698	95.8139	99.0698	89.3023
Heart Disease	85.1016	81.1525	90.1489	76.0734	91.069	76.4181

表 2.9 的实验结果显示基于 G-score 与 SVM 的后向浮动特征选择算法除了在 glass 数据集上的 5 折交叉验证实验选择的平均特征数最少之外，其他 8 个数据集上都不如基于 G-score 与 SVM 的前向浮动特征选择算法选择的特征数少；在 handwrite，Ionosphere，WDBC ，WPBC，Heart Disease 五个数据集上选择的特征数甚至比基于 G-score 与 SVM 的前向顺序特征选择算法选择的更多。

表 2.10 关于分类正确率的比较显示，基于 G-score 与 SVM 的后向浮动特征选择算法只有在 WDBC 一个数据集上的分类正确率比其他两个算法高，达到了最好的分类正确率，在其他 8 个数聚集上，该算法的分类正确率不如其他两个算法，分类性能较差。在训练集上的分类性能，基于 G-score 与 SVM 的后向浮动特征选择算法在 dermatology，glass，handwrite，WPBC，Heart Disease 五个数据集上优于其他两个特征选择算法。但是我们进行特征选择，以所选择的特征为样本特征构造相应的最佳分类模型时，我们更关心的是分类模型的推广性能，也即对于测试集的分类正确率。

以上分析比较可见：就特征子集的规模来看，基于 G-score 与 SVM 的前向浮动特征选择算法具有最优的性能。但就分类正确率，即分类器的泛化性能来看，基于 G-score 与 SVM 的前向顺序特征选择算法性能最优。

2.6 小结

本章提出了基于 G-score 与 SVM 的特征选择算法，该算法克服了基于 F-score 与 SVM 的特征选择算法只适用于两类分类问题的缺憾。其中的 G-score 特征重要性评价准则将

F-score 特征重要性评价准则由判断两类分类问题中的特征区分度，推广到可以评价任意类分类问题的特征区分度。提出的推广前向顺序特征搜索策略 GSFS、前向顺序浮动特征搜索策略 GSFFS，以及后向浮动特征搜索策略 GSBFS，结合 G-score 特征重要性评价准则，以 SVM 为分类工具，实现了基于 G-score 与 SVM 的三种有效混合特征选择算法。UCI 机器学习数据库数据集的实验测试证明了该三种混合的特征选择算法的有效性。其中，就所选择的特征子集规模来看，基于 G-score 与 SVM 的前向顺序浮动特征选择算法效果最佳；但就分类正确率，即分类模型的泛化性能来看，基于 G-score 与 SVM 的前向顺序特征选择算法最优。

第三章 基于 D-score 与 SVM 的特征选择

在第二章提出基于 G-score 与 SVM 的三种混合特征选择算法，并以 UCI 机器学习数据库的数据集对三种特征选择算法进行了测试。其中的特征重要性评价准则 G-score 是将传统 F-score 特征评价准则进行推广，提出的可以用于任意类分类问题的特征重要性评价准则；特征搜索策略是将顺序前向搜索（Sequential Forward Search，SFS）、顺序前向浮动搜索（Sequential Forward Floating Search，SFFS）、顺序后向浮动搜索（Sequential Backward Floating Search，SBFS）进行推广，提出的推广顺序前向搜索 （Generalized Sequential Forward Search，GSFS）、推广顺序前向浮动搜索 （Generalized Sequential Forward Floating Search，GSFFS）和推广顺序后向浮动搜索 （Generalized Sequential Backward Floating Search，GSBFS）。

然而，G-score 是一种基于类内、类间距离的特征评价准则，该准则能够衡量特征在类别间辨别能力的大小，是一种简单有效的特征选择方法，这在第二章已经证实。但是 G-score 特征重要性评价准则没有考虑不同特征的测量量纲对特征重要性，即对特征区分度的影响。为此，本章基于不同特征测量量纲的考虑，提出 D-score 特征重要性评价准则，用以判断特征在类别间的区分度大小，以此作为特征对分类贡献大小的度量，进行特征选择。从而提出基于 D-score 与 SVM 的特征选择算法，以克服基于 G-score 与 SVM 的特征选择算法在衡量特征的类间辨别能力大小时，没有考虑不同特征测量量纲对于特征区分度大小影响的缺陷。

本章内容组织如下，首先介绍 D-score 特征重要度评价准则；然后论述以该准则为特征区分能力度量准则，分别结合 GSFS、GSFFS、GSBFS 三种特征搜索策略，以 SVM 为分类工具，并以 SVM 分类模型的分类正确率引导特征搜索过程的三种混合特征选择算法。

3.1 D-score 特征重要度评价准则

公式（2-8）定义的 G-score 特征重要性评价准则是基于类内、类间距离的类别可分性准则，没有考虑不同的特征测量量纲对特征重要性，即特征区分度的影响。基于特征不同测量量纲的考虑，我们提出 D-score 特征重要性评价准则，其定义如下。

首先将式（2-8）除以 $l-1$，得式（3-1）：

$$F_i' = \frac{\frac{1}{l-1}\sum_{j=1}^{l}(\bar{x}_i^{(j)} - \bar{x}_i)^2}{\sum_{j=1}^{l}\frac{1}{n_j-1}\sum_{k=1}^{n_j}(x_{k,i}^{(j)} - \bar{x}_i^{(j)})^2} \tag{3-1}$$

其中 $\bar{x}_i$， $\bar{x}_i^{(j)}$ 分别为第 i 个特征在整个数据集和第 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 个特征的特征值。因此，式(3.1)中分子表示各类别之间的方差，此处以各类的质心代表相应类，分母表示各类别的类内方差之和。分子越大，表明类间的差异越大；分母越小，表明类内差异越小。因此式(3-1)值越大，表明相应特征的分类能力越强，即类间越疏，类内越密，分类效果越好，也就是说此特征的辨别力越强。

根据统计学原理[161]，方差反映了数据分散程度的绝对值，其数值大小一方面取决于原变量值本身水平，即度量值的高低，因此与变量的均值大小有关，变量值绝对水平高的，离散程度的测度值自然也就大，绝对水平低的离散程度的测度值自然也就小；另一方面，它们与原变量值的计量单位相同，采用不同计量单位计量的变量值，其离散程度的测度值也就不同。为了消除均值和测量单位不同对离散程度统计量的影响，引入了离散系数。离散系数是指一组数据的标准差和其对应均值的比值，也称为变异系数，用 v 表示，其定义如式 (3-2)所示。

$$v = \frac{\sigma}{\bar{x}} \tag{3-2}$$

其中，σ 和 $\bar{x}$ 分别表示样本的标准差和均值。离散系数能够比较不同总体或样本的离散程度，离散系数越大说明数据的离散程度越大；反之则数据的离散程度越小。

受到离散系数启示，为在一定程度上消除均值和不同量纲对离散程度的影响，将公式(3-1)中的类内类间方差分别除以各自的均值，得式(3-3)。

$$F_i'' = \frac{\frac{1}{(l-1)}\sum_{j=1}^{l}\frac{(\bar{x}_i^{(j)} - \bar{x}_i)^2}{\bar{x}_i}}{\sum_{j=1}^{l}\frac{1}{n_j-1}\sum_{k=1}^{n_j}\frac{(x_{k,i}^{(j)} - \bar{x}_i^{(j)})^2}{\bar{x}_i^{(j)}}} \tag{3-3}$$

由于计算每个特征的区分度时都要乘以 $\frac{1}{(l-1)}$，所以我们将式(3-3)同时乘以 $(l-1)$，得到式(3-4)，即 D-score 的定义。

$$D_i = F_i''' = \frac{\sum_{j=1}^{l} \frac{(\bar{x}_i^{(j)} - \bar{x}_i)^2}{\bar{x}_i}}{\sum_{j=1}^{l} \frac{1}{n_j - 1} \sum_{k=1}^{n_j} \frac{(x_{k,i}^{(j)} - \bar{x}_i^{(j)})^2}{\bar{x}_i^{(j)}}} \tag{3-4}$$

式(3-4)中的 $\bar{x}_i$，$\bar{x}_i^{(j)}$ 分别为第 i 个特征在整个数据集上、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 个特征的特征值。分析式（3-4）可知，分子表示各类别间的离散系数，其值越大，表示各类别间的分散程度越好；分母表示各类别内的变异系数之和，其值越小，表示每个类别越紧凑。因此式（3-4）的值越大，表示相应第 i 个特征的类别区分能力越强；反之，则弱。

3.2 基于 D-score 与 SVM 的顺序前向混合特征选择

该算法首先计算每个特征的 D-score 值，将特征根据其 D-score 值降序排序；利用 2.5.1 描述的 GSFS，每次从未被选取的特征中选择一个 D-score 值最大的特征添加到被选特征集合(被选特征集合初始为空集)；采用 SVM 对当前选取的特征子集进行评价，SVM 参数的选择依据 3.2.1 描述的网格搜索和五折交叉验证进行。迭代一直进行，直到所有特征都加入被选特征子集。选择分类效果最佳，即训练集分类正确率开始下降时的特征子集，作为最优特征子集，构建分类模型，得到具有最大分类间隔的分类超平面。

为了得到具有统计意义的实验结果，此处采用五折交叉验证的方法进行实验。将样本顺序随机打乱；然后对每一类的样本依次逐个加入到五个样本集合中（初始时样本集合为空），直到这一类的每一个样本都被加入。从而实现了样本均匀划分为五份的目的，以每一份作测试样本，其余四份作训练样本，实现五折交叉验证实验。

SVM 核函数参数选择、算法详细步骤，以及实验结果与分析分别描述如下。

3.2.1 SVM 最佳参数选择

实验中支持向量机的核函数采用径向基核函数 RBF[159]。为了得到具有较好推广性能的 SVM 分类模型，需要选择合适的 SVM 惩罚因子 C 和 RBF 的核函数参数 γ。为此，采用网格搜索和 5-折交叉验证相结合的方法来选择最优的参数对（C，γ）。其中，$C \in \{2^{-5},\cdots,2^{15}\}$，$\gamma \in \{2^{-15},\cdots,2^{5}\}$。实验中对每一组参数组合，训练集用 5-折交叉验证确定当前 C 和 γ 对应的分类正确率平均值，最后选取训练集平均分

类正确率最佳的一组 C 和 γ 做为 SVM 的最佳参数，利用训练及样本构建具有最大间隔分类面的 SVM 分类模型。以此模型对相应的、训练集、测试集进行测试，得到训练集、测试集的分类正确率。实验使用的 SVM 工具箱为台湾林智仁教授等开发的 LibSvm 工具箱[160]。

3.2.2 算法步骤描述

step 1：计算每个特征的 D-score 值，并初始化被选特征子集为空集。即，$selected\_feature\_subset = \Phi.$

step 2：对全部特征根据其 D-score 值降序排序，并构成备选特征子集。即，$candidate\_feature\_subset = full\_feature\_set.$

step 3：若备选特征子集不空，则从备选特征子集中选择当前第一个特征，也即 D-score 值最大的特征 f_i，添加到被选特征集合，并从备选特征子集删掉该特征。即，if $candidate\_feature\_subset \neq \Phi$ then

$f_i = \max\{F_i \mid D-score(f_i) = F_i, f_i \in candidate\_feature\_subset\}$，

$selected\_feature\_subset = selected\_feature\_subset + \{f_i\},$

$candidate\_feature\_subset = candidate\_feature\_subset - \{f_i\}.$

step 4：对只含有被选特征子集特征的训练集样本进行 5 折交叉验证训练，构造相应的最优 SVM 分类器，得到临时 SVM 分类模型；以只含有被选特征子集特征的训练集、测试集样本对该临时 SVM 分类模型进行测试，记录训练集的分类准确率，以及相应测试集的分类准确率，以该临时 SVM 分类器对训练集样本的分类准确率评价当前被选择特征子集的性能。

Step 5: 转 step 3，直到备选特征子集为空。即，$candidate\_feature\_subset = \Phi$。

最后选择分类效果最佳，即训练集分类正确率开始下降时的特征子集，作为最优特征子集，构建 SVM 分类模型，得到具有最大分类间隔的分类超平面。

3.2.3 实验结果与分析

为了证明 D-score 特征评价准则的有效性，我们对基于 D-score 与 GSFS 和 SVM 的混合特征选择方法，和基于 G-score 与 GSFS 和 SVM 的混合特征选择方法进行了实验比较。实验采用和 2.5.3 部分相同的 UCI 机器学习数据库[158]的 9 个标准数

据集进行测试。数据集描述如表 2.6 所示。实验所采用的五折交叉验证实验同 2.5.3 在同样划分的数据集上进行。

表 3.1 为基于 D-score 和 G-score 的顺序前向混合特征选择算法的训练集和测试集的 5-折交叉验证实验所选择的特征子集规模的比较。表 3.2 为该两种顺序前向混合特征选择算法的训练集和测试集的 5-折交叉验证实验的平均分类正确率比较。其中，表 3.1 特征数目中带下划线的表示的是特征数目较小，表 3.2 分类正确率中带下划线的表示分类正确率较高。

表 3.1 D-score+GSFS 与 G-score+GSFS 混合特征选择算法选择的特征子集规模
Table 3.1 the comparison of the size of selected feature subset of the hybrid feature selection algorithms of D-score + GSFS and G-score +GSFS

数据集	**原始特征数**	**G-score+SVM+GSFS**	**D-score+SVM+GSFS**
dermatology	34	24	24
glass	9	4	4
handwrite	255	13	18
Ionosphere	34	8	8
WDBC	30	11	11
WPBC	33	10	6
wine	13	7	4
thyroid-disease	5	4	4
Heart Disease	13	7	6

表 3.2 D-score+GSFS 与 G-score+GSFS 混合特征选择算法选择的分类正确率
Table 3.2 the comparison of the accuracy of the hybrid feature selection algorithms of D-score + GSFS and G-score +GSFS

数据集	训练集(%)		测试集(%)	
	D-score	G-score	D-score	G-score
dermatology	98.8821	98.8121	98.6023	98.3245
glass	98.3619	98.3619	92.5451	92.5451
handwrite	99.1473	99.3041	98.4566	98.2209
Ionosphere	95.7989	95.7989	92.0282	92.0282
wdbc	97.6271	97.6706	95.4259	95.2520
wpbc	84.4532	87.5068	73.7439	74.7699
wine	97.7503	99.1588	97.7292	97.1569
thyroid-disease	98.4884	98.8372	95.8140	95.8140
Heart Disease	85.0998	85.1016	80.8316	81.1525

从表 3.1 中 9 个数据集的实验结果可以看出，D-score 准则所选择的特征子集的规模只有在 handwrite 数据集上大于 G-score 准则，其他 8 个数据集上 D-score 准则所选择的特征数都不多于 G-socre 准则。

表 3.2 显示，以 D-score 特征重要性评价准则所选特征子集构造的分类器，其分类性能在 7 个数据集上超过或等于推广的 F-score，即 G-score 特征重要性衡量准则所选择特征构造的分类器；只有 WPBC 和 Heart Disease 两个数据集上的分类

性能略低于 G-score 准则。

以上分析说明，D-score 特征重要性评价准则优于 G-score 特征重要性评价准则，在特征选择中能选择出分类性能更强的特征子集。

3.3 基于 D-score 与 SVM 的顺序前向浮动混合特征选择

同样，为了克服顺序前向搜索 GSFS 的“子集嵌套包含”缺陷，本部分研究基于 D-score 特征区分能力衡量准则，以 SVM 为分类工具的顺序前向浮动混合特征选择算法，该算法以 D-score 度量特征的类间辨别能力大小，采用推广的顺序前向浮动搜索 GSFFS，以 SVM 为分类工具引导特征选择过程。算法详细步骤描述，以及实验结果与分析分别描述如下。

3.3.1 算法步骤描述

step 1：确定当前的训练集、测试集。在训练集上计算每个特征的 D-score 值，并依据 D-score 值将所有特征降序排序。

step 2：初始化被选特征子集为空集，备选特征子集为特征全集。即，使得

$$selected\_feature\_subset = \Phi,$$
$$candidate\_feature\_subset = full\_feature\_set.$$

step 3：从备选特征子集中选择当前第一个特征，也即当前 D-score 值最大的特征，添加到被选特征集合，并从备选特征子集删掉该特征。即，

if $candidate\_feature\_subset \neq \Phi$ then

$$f_i = \max\{F_i \mid D-score(f_i) = F_i, f_i \in candidate\_feature\_subset\},$$

$$candidate\_feature\_subset = candidate\_feature\_subset - \{f_i\}.$$

step 4：构造只含有被选特征子集特征的训练集和测试集。以训练集样本 5 折交叉验证训练 SVM， 得到最优的临时 SVM 分类模型；记录这个临时 SVM 分类模型对训练集的分类正确率，以此评价当前被选择特征子集的性能。

step 5：若备选特征子集不空，则从备选特征子集中选择当前第一个特征，也是区分度最高的特征试添加到被选特征集合，构成临时的被选择特征子集。即，

if $candidate\_feature\_subset \neq \Phi$ then

$$f_i = \max\{F_i \mid D-score(f_i) = F_i, f_i \in candidate\_feature\_subset\},$$

$$temp\_selected\_feature\_subset = selected\_feature\_subset + \{f_i\}.$$

Step 6：以只含有临时被选特征子集特征的训练集样本采用 5 折交叉训练 SVM 分类器，得到临时最优 SVM 分类模型；以训练集样本来测试这个临时 SVM 分类模型的分类性能。若这个临时 SVM 的分类准确率提高，则将试加入的特征加入被选特征子集；否则，舍弃该特征，不将其加入到被选特征子集。即，

if $current\_accuracy > pre\_accuracy$ then

$selected\_feature\_subset = selected\_feature\_subset + \{f_i\}$ 。

step 7：从备选特征子集删去该特征。即，

$candidate\_feature\_subset = candidate\_feature\_subset - \{f_i\}$ 。

step 8：转 step 5，直到 $candidate\_feature\_subset = \Phi$ 。

最后留在被选特征子集中的特征即为所选择的特征。以该特征子集利用训练集样本构造相应 SVM 分类模型，以其对测试集样本的分类准确率来评价该最终选择的特征子集。GSFFS 特征搜索策略能够在一定程度上避免 GSFS 的陷入局部最优解的缺陷，同时选取的特征数目通常会比 GSFS 策略的少。

3.3.2 实验结果与分析

为了说明 D-score 准则的有效性，我们对基于 D-score 准则与 SVM 的顺序前向浮动搜索特征选择算法的实验结果与基于 G-score 与 SVM 的顺序前向浮动搜索特征选择算法进行实验比较。实验和 2.5.3 部分的实验在同样划分的相同数据集上进行，也是 5 折交叉验证实验，实验中 SVM 的核函数参数采用同样的方法进行选择。实验结果比较如表 3.3 和表 3.4 所示。其中，表 3.3 特征数目中带下划线的表示的是特征数目较小，表 3.4 中带下划线的分类正确率表示相应的分类正确率较高。

表 3.3 D-score+GSFFS 与 G-score+GSFFS 混合特征选择算法的特征子集规模

Table 3.3 the comparison of the size of selected feature subset of the hybrid feature selection algorithms of D-score + GSFFS and G-score +GSFFS

数据集	原始特征数	G-score+SVM+GSFFS	D-score+SVM+GSFFS
dermatology	34	12.4(11, 18)	13.4(10, 15)
glass	9	4.4(3, 5)	4.4(3, 5)
handwrite	255	7.6(7, 9)	9.8(7, 11)
Ionosphere	34	11.8(10, 15)	12(10, 15)
WDBC	30	9(4, 13)	9.4(4, 12)
WPBC	33	7.4(5, 12)	7.6(4, 12)
wine	13	5.6(5, 7)	5.4(4, 7)

thyroid-disease	5	3.6(3, 4)	4.4(4, 5)
Heart Disease	13	5.4(4, 6)	6.2(4, 8)

表 3.4 D-score+GSFFS 与 G-score+GSFFS 混合特征选择算法的分类正确率
Table 3.4 the comparison of the accuracy of the hybrid feature selection algorithms of D-score + GSFFS and G-score +GSFFS

数据集	训练集(%)		测试集(%)	
	D-score	G-score	D-score	G-score
dermatology	98.8816	98.9515	97.1821	98.0463
glass	99.4166	99.4166	92.0477	92.0477
handwrite	99.9225	99.6908	98.7596	98.4471
Ionosphere	99.8576	99.8576	92.8853	92.8853
wdbc	99.1638	99.3842	94.2055	95.0873
wpbc	92.9113	93.1694	72.7456	70.1667
wine	99.7183	100	97.2522	99.4444
thyroid-disease	98.9535	99.0698	96.2791	95.8139
Heart Disease	91.2371	90.1489	78.4237	76.0734

表 3.3 的实验数据显示，D-score 准则 除了在 handwrite，thyroid-disease 和 Heart Disease 三个数据集上的 5 折交叉验证实验所选择的特征数的最多特征个数多于 G-score 准则外，其他数据集上 D-score 准则所选择的最多特征个数都不多于 G-score 准则。

表 3.4 的数据可以看出，D-score 特征重要性衡量准则结合 GSFFS 特征选择策略所选择的特征除在 dermatology, WDBC 和 wine 三个数据集的分能性能略低于 G-score 准则外，其他 6 个数据集上的分类性能均高于或等于 G-score 准则。

由此可见 D-score 特征重要性评价准则所选择的特征具有较好的分类性能，这说明 D-score 特征评价准则提供了有效的特征区分度衡量标准，以该准则作为特征区分度评价准则进行特征选择能选择到对分类起主要作用的特征。

3.4 基于 D-score 与 SVM 的顺序后向浮动混合特征选择

为了进一步说明 D-score 特征评价准则的有效性，此部分研究基于 D-score、SVM 和 GSBFS 的顺序后向浮动混合特征选择算法，并将该算法与基于 G-score、SVM 和 GSBFS 的后向顺序浮动特征选择方法进行实验比较。基于 D-score 与 SVM 的顺序后向浮动混合特征选择算法按 D-score 取值将对应的特征升序排序，采用 GSBFS 特征搜索策略进行搜索，以 SVM 为分类工具，指导特征选择过程。从特征全集开始，每次尝试删除当前第一个特征，如果删除该特征导致分类正确率上升，则将此特征从特征集中剔除；否则，保留该特征在特征子集里。直到所有的

特征都扫描一遍之后结束。算法实现步骤和实验结果与分析分别如 3.4.1 和 3.4.2 描述。

3.4.1 算法步骤描述

step 1：确定本折实验的训练集和测试集，并以训练样本集计算每个特征的 D-score 值，按照 D-score 值将特征升序排序，全部特征构成被选特征子集和备选特征子集。

step 2：对包含被选特征子集特征的训练集样本进行 5 折交叉验证训练，也即使用包含全部特征的训练样本训练 SVM 分类器，得到一个临时最优的 SVM 分类模型，测试该模型对训练集和测试样本集样本的分类正确率 。

step 3：试将当前备选特征子集的第一个特征，从被选特征子集中删除。

step 4：对只含有当前被选特征子集特征的训练集样本进行 5 折交叉验证训练，得到一个临时最优的 SVM 分类模型；以该模型对训练集和测试集样本进行测试，记录相应的分类正确率。

step 5：若训练集的分类正确率提高，则从被选特征子集删除该特征，否则不删去该特征。

step 6：将该特征从备选特征子集删除。

step 7：转 step 3，直到备选特征子集为空。

最后被选特征子集中的特征即为所选择的特征。该特征选择算法避免了顺序后向特征搜索策略的“子集嵌套”缺陷，部分考虑了特征之间的相关性，使得最终得到的特征子集避开了局部最优的缺憾。

3.4.2 实验结果与分析

为了与基于 G-score、SVM 和 GSBFS 的后向顺序浮动特征选择方法进行比较，本部分实验和 2.5.4 部分的实验在同样划分的相同数据集上进行，也是 5 折交叉验证实验，实验中 SVM 的核函数参数选择也是在训练集上采用网格搜索和 5 折交叉验证方法确定。实验结果比较如表 3.5 和表 3.6 所示。其中，所选特征向量的数目以平均特征数（最小值，最大值）的形式在表 3.5 给出，特征数目中带下划线的表示的是特征数目较小。表 3.6 是两个算法的分类正确率比较，其中带下划线的分类正确率表示相应的分类正确率较高。

表 3.5 D-score+GSBFS 与 G-score+GSBFS 两种混合特征选择算法选择的特征子集规模比较
Table 3.5 the comparison of the size of selected feature subset of the hybrid feature selection algorithms of D-score + GSBFS and G-score +GSBFS respectively

数据集	原始特征数	G-score+SVM+GSBFS	D-score+SVM+GSBFS
dermatology	34	14.2(9, 17)	15.6(10, 22)
glass	9	3.8(3, 5)	3.8(3, 5)
handwrite	255	13.6(8, 23)	11.6(6, 24)
Ionosphere	34	14.6(10, 21)	15.4(10, 21)
WDBC	30	20(15, 23)	19.2(15, 22)
WPBC	33	25(19, 29)	24.8(19, 29)
wine	13	5.8(4, 7)	5.6(5, 8)
thyroid-disease	5	3.6(3, 4)	3.6(3, 4)
Heart Disease	13	9.6(6, 12)	9.8(7, 12)

表 3.6 D-score+GSBFS 与 G-score+GSBFS 混合特征选择算法的分类正确率比较
Table 3.6 the comparison of the accuracy of the hybrid feature selection algorithms of D-score + GSBFS and G-score +GSBFS

数据集	训练集(%)		测试集(%)	
	D-score	G-score	D-score	G-score
dermatology	99.5115	99.5814	95.7999	94.1169
glass	99.5335	99.5335	93.0096	89.7206
handwrite	100	100	97.8412	98.1489
Ionosphere	99.8574	99.8574	92.8732	92.8773
wdbc	99.3411	99.1211	97.1958	96.6556
wpbc	94.8369	94.8369	74.7031	74.7031
wine	100	99.8612	88.1868	89.3800
thyroid-disease	99.0698	99.0698	87.4418	89.3023
Heart Disease	91.069	91.069	77.4181	76.4181

从表 3.5 所示的 5 折交叉验证实验所选择的特征数分析，除了 dermatology，wine 和 Heart Disease 三个数据集 G-score 所选择的最少和最多特征数略低于 D-score 准则；handwrite 数据集的最多特征数略低于 D-score 外，其他五个数据集上 D-score 准则的 5 折交叉验证实验所选择的特征数都不多于 G-score 准则所选择的特征数。

由表 3.6 可以看出 D-score 特征重要性评价准则所选择特征子集的分类正确率只有在 handwrite，Ionosphere，wine 和 thyroid-disease 四个数据集上略低于 G-score 准则，在其他 5 个数据集上由 D-score 准则所选择特征子集构造的 SVM 分类模型的分类正确率都不低于依据 G-score 准则所选择的特征子集构造的 SVM 分类模型的分类准确率。

以上基于 D-score 特征特征区分度准则的 3 种混合特征选择方法的 5 折交叉验证实验结果的分析比较可见，本章提出的 D-score 特征评价准则是一种有效的特征重要性衡量标准。以该准则作为特征在类别间辨别能力大小的度量，以 SVM 为分

类工具进行特征选择，可以实现更优特征子集的选择，构造具有更大分类间隔的分类模型。

3.5 小结

本章提出了基于 D-score 与 SVM 的特征选择算法，解决了基于 G-score 与 SVM 的特征选择算法在衡量特征的类间辨别能力大小时，没有考虑不同特征的测量量纲对于特征区分能力大小影响的缺陷。作为一种新的特征重要性，即特征区分度，评价准则 D-score，不但保持了第二章提出的 G-score 准则可以评价两类或多类分类问题中特征分类辨识能力大小的性能；而且克服了 G-score 准则在判断特征辨识能力大小时没有考虑特征测量量纲对特征辨识能力大小影响的缺陷。将 D-score 特征评价准则结合我们在第二章提出的 GSFS、GSFFS 以及 GSBFS 三种推广特征子集搜索策略，以 SVM 为分类工具对所选择特征子集进行评价，指导特征选择过程，实现三种混合的特征选择算法。UCI 机器学习数据库中 9 个特征选择的常用数据集实验测试，以及与基于 G-score 与 SVM 的相应混合特征选择方法的 5 折交叉验证实验的实验结果比较得出：本章提出的 D-score 特征重要性评价准则是一种有效的特征辨识能力评价准则，基于该准则与 SVM 的混合特征选择方法所选择的特征具有较好的分类效果，其分类性能优于基于 G-score 与 SVM 的混合特征选择方法，达到了保持数据集辨识能力不变情况下进行维数压缩的目的。

基于 D-score 与 SVM 的三种混合特征选择算法的比较显示：就所选择的特征子集规模来看，基于 D-score 与 SVM 的前向顺序浮动特征选择算法最优；但就分类器的泛化性能来看，基于 D-score 与 SVM 的前向顺序混合特征选择算法具有最优的泛化性能。

第四章　基于 DFS 与 SVM 的特征选择

第 2～3 章分别介绍了基于 G-score 与 SVM 的特征选择算法，以及基于 D-score 与 SVM 的特征选择算法。其中，G-score 是特征重要性评价准则 F-score 的推广；D-score 是 G-score 准则的改进。并以 UCI 机器学习数据库的数据集对基于 G-score 与 SVM，和基于 D-score 与 SVM 的混合特征选择算法进行了实验比较，证明了算法的有效性，同时也证明了 G-score 与 D-score 该两个特征区分能力大小衡量准则的有效性。

然而，无论 G-score 准则，还是 D-score 准则，准则本身并没有考虑特征之间的相关性，即就是这两个准则在衡量单个特征的类间辨别能力大小时候，没有考虑特征之间的相关性对于单个特征辨别能力大小的影响。为此，本章基于特征间相关性的考虑，提出 DFS（Discernibility of Feature Subsets，DFS）特征子集区分度衡量准则，计算特征子集中特征的联合 G-score 值，考虑特征子集中特征的联合作用，以判断特征子集在类别间区分度的大小，并以此作为特征子集对分类贡献大小的度量，进行特征选择。提出基于 DFS 与 SVM 的特征选择方法，该算法将克服基于 G-score 与 SVM，以及基于 D-score 与 SVM 的特征选择方法在特征选择时，没有考虑特征之间的相关性对于特征的类间区分度大小影响之缺憾。

本章内容组织如下，首先介绍 DFS 特征区分度评价准则；然后论述以该准则为特征子集区分能力度量准则，分别结合经典的顺序前向搜索 SFS、顺序后向搜索 SBS、顺序前向浮动搜索 SFFS、顺序后向浮动搜索 SBFS 四种特征搜索策略，以 SVM 为分类工具的四种特征选择算法。

4.1　DFS 特征重要性评价准则

G-score 只考虑了单个特征对于分类的贡献，没有考虑特征之间的相关性。DFS 在度量特征的类间区分能力大小时，考虑特征之间的相关性，提出了特征子集的区分度概念，将 G-score 由衡量单个特征对分类的贡献，推广到衡量特征子集对于分类的贡献。DFS 的定义如下。

对于两类分类问题，设有规模为 n 的训练样本集 $\{(x_k,y_k)\mid x_k\in R^m,\ m\succ 0,\ y_k\in\{1,-1\},\ k=1,2,\cdots,n\}$。其中，$m$ 表示样本空间维数，即每个样本的特征数，$\left\|\{y_k\,|y_k=+1,k=1,2,\cdots,n\}\right\|=n_+$，$\left\|\{y_k\,|y_k=-1,k=1,2,\cdots,n\}\right\|=n_-$，即正类和负类的样本数分别为 n_+ 和 n_-。则包含有 $i(i=1,2,\cdots m)$ 个特征的特征子集的区分度

DFS_i 定义为：

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

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 个特征的特征值；$x_{k,j}^{(-)}$ 为第 k 个负类样本的第 j 个特征的特征值。以上式（4-1）可以简记为式（4-2）。

$$DFS_i = \frac{\left\|\bar{x}^{(+)} - \bar{x}\right\| + \left\|\bar{x}^{(-)} - \bar{x}\right\|}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left\|x_k^{(+)} - \bar{x}^{(+)}\right\| + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left\|x_k^{(-)} - \bar{x}^{(-)}\right\|} \tag{4-2}$$

其中，$\|X - Y\|$ 表示向量 X, Y 之间的平方距离；$\bar{x}$， $\bar{x}^{(+)}$， $\bar{x}^{(-)}$ 分别为包含 i 个特征的特征子集在整个数据集上的均值，在正类数据集上的均值和在负类数据集上的均值；$x_k^{(+)}$ 为正类第 k 个样本点的对应当前 i 个特征构成的特征子集的特征值构成的向量；$x_k^{(-)}$ 为第 k 个负类样本的对应当前 i 个特征构成的特征子集的特征值构成的向量。

分析式（4-1）和式（4-2）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到整个样本集对应当前特征子集的均值向量的平方距离；分母表示正类、负类对应当前含有 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特征子集的类间越疏；分母值越小表示对应当前特征子集的类内越聚。因此 DFS_i 值越大，表明当前含有 i 个特征的特征子集的类间区分能力越强，也即类别辨识力越强。

对于 $l(l \geq 2)$ 类的多类分类问题，假设训练样本集规模为 n，样本空间维数为 m。即 $\{(x_k, y_k) \mid x_k \in R^m, m \succ 0, y_k \in \{1, 2, \cdots, l\}, l \geq 2, k = 1, 2, \cdots, n\}$。其中，第 j 类的样本数为 n_j，即 $\left\|\{y_k \mid y_k = j, k = 1, 2, \cdots, n\}\right\| = n_j$，$j = 1, 2, \cdots, l$，则样本的含有 $i\,(i = 1, 2, \cdots, m)$ 个特征的特征子集的区分度 DFS_i 定义为式（4-3）。

$$DFS_i = \frac{\sum_{j=1}^{l}\left\|\bar{x}^{(j)} - \bar{x}\right\|}{\sum_{j=1}^{l}\frac{1}{n_j - 1}\sum_{k=1}^{n_j}\left\|x_k^{(j)} - \bar{x}^{(j)}\right\|} \tag{4-3}$$

其中 $\bar{x}$, $\bar{x}^{(j)}$ 分别为当前 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；$x_k^{(j)}$ 为第 j 类中第 k 个样本对应当前 i 个特征的特征子集的特征值向量。式（4-3)的分子表示 l 个类别中各类别的对应当前 i 个特征的特征子集的样本中心向量到整个样本集的中心向量的距离平方和，其值越大，类间越疏。式(4-3)的分母表示各个类别对应当前 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（4-3）定义的 DFS_i 近似表示了当前 i 个特征的特征子集的类间、类内方差之比，其值越大表明当前 i 个特征对应的特征子集的辨识力越强。

为了比较本章提出的 DFS 特征子集重要性评价准则的有效性，我们将 DFS 特征子集评价准则与 CFS（Correlation based Feature Selector，CFS）[162]特征子集评价准则进行实验比较，CFS 准则如式（4-4）定义。

$$Ms = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{4-4}$$

式（4-4）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，$\overline{r_{ff}}$ 是特征之间相关系数的均值。公式（4-4）的分子可看成特征子集 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，就是选择一组最优的特征子集，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[162]。由此可见 Ms 的值越大，说明了当前特征子集 S 对于分类的贡献越大，是优良的特征子集。其中，相关系数一般使用的是 Pearson 相关系数，计算公式如式（4-5）所示。

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}} \tag{4-5}$$

其中，X, Y 表示待求相关系数样本的两个向量，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。

由于 CFS 准则中相关系数计算采用的 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。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，相关度越弱。因此，我们将计算相关系数使用的 Pearson 相关系数进行改进，对 Pearson 相关系数取绝对值后再计算 CFS。改进后的 CFS 记为 CFSPabs（Correlation based Feature Selector based on the absolute of Pearson's correlation coefficient，CFSPabs）。CFSPabs 中计算相关系数的公式如下式（4-6）所示。

$$r' = \frac{\left| \sum XY - \frac{\sum X \sum Y}{N} \right|}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X \right)^2}{N} \right) \left(\sum Y^2 - \frac{\left(\sum Y \right)^2}{N} \right)}} \tag{4-6}$$

4.2 基于 DFS 与 SVM 的顺序前向混合特征选择

式（4-3）定义了特征子集的类间区分能力大小，根据此公式，可以计算包含任意多个特征的特征子集在分类中的重要性，即对分类的贡献大小。为此，我们将 DFS_i 引入到特征选择过程，以此代替前两章的 G-score 和 D-score 准则，克服 G-score 和 D-score 准则只考虑单个特征的分类贡献，没有考虑特征间联合作用的缺陷。

我们首先提出将 DFS_i 与顺序前向搜索策略 SFS 结合的特征选择算法。算法以 SVM 为分类工具，算法思想和实验结果分别见 4.2.1 和 4.2.2 描述。

4.2.1 算法思想描述

设 S 为全部特征构成的集合，C 为被选择特征构成的子集，C 初始为空集。确定当前的训练集和测试集。

Step 1：在训练集上计算每个特征的 G-score 值，选择 G-score 值最大的特征 k 加入被选特征子集 C，并从 S 集合中删去特征 k，即置 C=C+k，S=S-k；

Step 2：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；

Step 3：判断 S 是否为空，若 S 不为空集，将 S 中每一个特征与特征子集 C 组合，构成特征数增 1 的特征子集，计算相应特征子集的 DFS_i 值，选择 DFS_i 最大的特征子集对应的特征 k 加入特征子集 C，即置 C=C+k，S=S-k，go to Step 2；否则，若 S 为空，

则算法结束。

最后选择训练集分类正确率不再上升时对应的最小特征子集为被选择特征子集。

4.2.2 实验结果与分析

本部分实验采用 UCI 机器学习数据库[158]的 10 个数据集 iris，dermatology，glass，handwrite，Ionosphere，WDBC (Wisconsin Diagnostic Breast Cancer，WDBC)，WPBC (Wisconsin Prognostic Breast Cancer, WPBC)，wine，thyroid-disease，和 heart disease。实验所用数据集描述如表 4.1 所示。其中，dermatology 数据集，删去了 8 个含有缺失数据的样本，因此总的数据集规模由原来的 366 变成了 358；glass 数据集分成了 window glass 和 non-window glass 两类；handwrite，即 Semeion Handwritten Digit 数据集，只选择了前两类进行实验；WPBC 数据集删去了 4 个含有缺失数据的样本，因此数据集规模为 194；关于 thyroid-disease 数据集，实验使用了其中的 new-thyroid，也即 Thyroid gland data 数据集；Heart Disease 数据集使用的是其中的 processed cleveland 数据集，实验中删去了 6 个含有缺失数据的样本，因此数据集规模为 297。

表 4.1 实验所用 UCI 数据集描述

Table 4.1 the description of data sets from UCI machine learning repository

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

为了得到具有统计意义的实验结果，首先将样本顺序随机打乱，然后对每一类的样本依次逐个加入到五个样本集合中（初始时样本集合为空），直到这一类的每一个样本都被加入。这样将样本均匀划分为五份，以每一份作测试样本，其余四份作训练样本，最后以每一份都做过测试集后结束，实现五折交叉验证实验。其中，样本随机打乱的方法是：生成一个 5000 行 2 列的 2 维数组，数组的每一个元素的值在 1～数据集规模之间；交换 2 维数组每一行的两个元素值对应的数据集的两个样本。 实验采用和前两章同样的 SVM 工具箱[160]，SVM 的核函数参数采用默认的参数。

同时为了比较本章提出的 DFS 准则的有效性，将算法 4.2.1 的“Step 3 的计算相应特

征子集的 DFS_i 值，选择 DFS_i 最大的特征子集对应的特征 k 加入特征子集 C，即置 C=C+k，S=S-k”，修改为根据式（4-4）和式（4-5），以及根据式（4-4）和（4-6）计算相应特征子集的 Ms 值，即计算 CFS 特征子集评价准则与 CFSPabs 特征子集评价准则的相应值，并据此来评估相应的特征子集的分类性能。并将基于 DFS、CFS 和 CFSPabs 三种不同特征子集重要性评价准则与 SVM 的特征选择算法进行比较。实验结果如表 4.2 到表 4.12 所示。

表 4.2 到表 4.11 展示了分别基于 DFS、CFS，以及 CFSPabs 特征子集评价准则与 SVM 的顺序前向特征选择算法在各个数据集上的 5 折交叉验证实验的详细结果。表中加重的测试集分类正确率值表示最优的测试集分类正确率。表 4.12 给出了分别基于 DFS，CFS，CFSPabs 与 SVM 的特征选择算法的 5 折交叉验证实验结果的平均值比较。表中加重和加下划线的被选择特征子集规模值表示最小的特征子集规模，加重和加下划线的测试集分类正确率值表示最优的分类正确率。

表 4.2 Iris 数据集前向顺序特征选择算法 5 折交叉验证的实验结果

Table 4.2 the experimental results of SFS on Iris dataset

Iris	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	4	4	4	99.16667	99.16667	99.16667	96.66667	96.66667	96.66667
Fold 2	4	4	4	98.33333	98.33333	98.33333	93.33333	93.33333	93.33333
Fold 3	4	4	4	99.16667	99.16667	99.16667	96.66667	96.66667	96.66667
Fold 4	4	4	4	99.16667	99.16667	99.16667	96.66667	96.66667	96.66667
Fold 5	3	3	3	97.5	97.5	97.5	100	96.66667	96.66667
Average	3.8	3.8	3.8	98.66667	98.66667	98.66667	**96.66667**	96	96

表 4.3 Dermatology 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.3 the experimental results of SFS on Dermatology dataset

Dermatology	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	33	34	20	98.94366	98.94366	99.29577	97.2973	89.18919	86.48649
Fold 2	33	30	25	99.3007	97.9021	98.6014	95.83333	93.05556	90.27778
Fold 3	34	29	28	98.95105	98.95105	98.95105	93.05556	88.88889	93.05556
Fold 4	32	30	23	99.30556	98.61111	98.61111	95.71429	95.71429	95.71429
Fold 5	34	34	25	98.95833	98.95833	98.95833	95.71429	95.71429	92.85714
Average	33.2	31.4	24.2	99.09186	98.67325	98.88353	**95.52295**	92.51244	91.67825

表 4.4 glass 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.4 the experimental results of SFS on glass dataset

Glass	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	8	7	7	96.47059	96.47059	96.47059	93.18182	95.45455	93.18182
Fold 2	9	9	7	95.90643	95.90643	95.90643	90.69767	90.69767	93.02326
Fold 3	9	6	9	95.32164	95.90643	95.32164	97.67442	97.67442	97.67442
Fold 4	9	8	9	97.09302	97.09302	97.09302	95.2381	95.2381	95.2381
Fold 5	9	9	4	97.67442	97.67442	97.67442	90.47619	90.47619	88.09524
Average	8.8	7.8	7.2	96.49322	96.61018	96.49322	93.45364	**93.90818**	93.44257

表 4.5 handwrite 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.5 the experimental results of SFS on handwrite dataset

Handwrite	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	30	143	9	100	100	100	95.45455	98.48485	98.48485
Fold 2	40	145	27	100	100	100	100	100	98.46154
Fold 3	41	146	10	100	100	100	100	98.4375	100
Fold 4	34	132	14	100	100	100	100	96.875	98.4375
Fold 5	41	131	8	100	100	100	98.4375	100	96.875
Average	37.2	139.4	13.6	100	100	100	**98.77841**	98.75947	98.45178

表 4.6 Ionosphere 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.6 the experimental results of SFS on Ionosphere dataset

Ionosphere	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	28	22	27	94.64285714	94.28571429	94.64286	97.18309859	97.18309859	97.1831
Fold 2	15	19	8	96.08540925	95.72953737	96.08541	85.71428571	85.71428571	88.57143
Fold 3	18	20	19	95.72953737	95.72953737	95.72954	87.14285714	88.57142857	90
Fold 4	17	20	24	94.30604982	94.30604982	94.66192	97.14285714	97.14285714	97.14286
Fold 5	26	13	16	94.66192171	94.66192171	94.66192	94.28571429	91.42857143	92.85714
Average	20.8	18.8	18.8	95.08515506	94.94255211	95.15633	92.29376258	92.00804829	**93.15091**

表 4.7 WDBC 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.7 the experimental results of SFS on WDBC dataset

WDBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	28	13	14	100	100	100	66.08696	62.6087	62.6087
Fold 2	28	14	14	100	100	100	61.73913	64.34783	64.34783
Fold 3	28	15	15	100	100	100	62.83186	62.83186	62.83186
Fold 4	28	15	16	100	100	100	67.25664	62.83186	62.83186
Fold 5	28	13	16	100	100	100	65.48673	62.83186	62.83186
Average	28	14	15	100	100	100	**64.68026**	63.09042	63.09042

表 4.8 WPBC 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.8 the experimental results of SFS on WPBC dataset

WPBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	29	4	7	100	100	100	75	77.5	75
Fold 2	29	3	2	100	100	100	76.92308	74.35897	74.35897
Fold 3	29	9	8	100	100	100	76.92308	76.92308	76.92308
Fold 4	30	8	8	100	100	100	76.31579	76.31579	76.31579
Fold 5	30	4	6	100	100	100	76.31579	81.57895	76.31579
Average	29.4	5.6	6.2	100	100	100	76.29555	**77.33536**	75.78273

表 4.9 Wine 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.9 the experimental results of SFS on Wine dataset

Wine	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	10	13	6	100	100	100	56.75676	43.24324	43.24324
Fold 2	13	8	6	100	100	100	36.11111	86.11111	36.11111
Fold 3	13	13	11	100	100	100	41.66667	41.66667	38.88889
Fold 4	10	13	6	100	100	100	54.28571	48.57143	45.71429
Fold 5	13	13	6	100	100	100	50	50	44.11765
Average	11.8	12	7	100	100	100	47.76405	**53.91849**	41.61504

表 4.10 thyroid-disease 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.10 the experimental results of SFS on thyroid-disease dataset

thyroid-disease	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	5	4	4	100	100	100	74.4186	79.06977	79.06977

Fold 2	5	4	4	100	100	100	76.74419	76.74419	76.74419
Fold 3	5	5	5	100	100	100	76.74419	76.74419	76.74419
Fold 4	5	4	4	100	100	100	74.4186	74.4186	74.4186
Fold 5	4	4	4	100	100	100	86.04651	74.4186	74.4186
Average	4.8	4.2	4.2	100	100	100	**77.67442**	76.27907	76.27907

表 4.11 Heart Disease 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.11 the experimental results of SFS on Heart Disease dataset

Heart Disease	**被选择特征数**			**训练集分类正确率**			**测试集分类正确率**		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	13	13	13	100	100	100	53.33333	53.33333	53.33333
Fold 2	13	12	12	100	100	100	53.33333	53.33333	53.33333
Fold 3	13	12	13	100	100	100	53.33333	53.33333	53.33333
Fold 4	13	9	11	100	100	100	54.23729	54.23729	54.23729
Fold 5	13	13	13	100	100	100	55.17241	55.17241	55.17241
Average	13	11.8	12.4	100	100	100	53.88194	53.88194	53.88194

表 4.12 UCI 数据集前向顺序特征选择算法 5 折交叉验证实验结果

Table 4.12 the experimental results of SFS on UCI datasets

Data sets	**原始特征数**	**被选择特征数**			**测试集分类正确率**		
		DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Iris	4	3.8	3.8	3.8	**96.66667**	96	96
dermatology	34	33.2	31.4	**24.2**	**95.52295**	92.51244	91.67825
glass	9	8.8	7.8	**7.2**	93.45364	**93.90818**	93.44257
handwrite	255	37.2	139.4	**13.6**	**98.77841**	98.75947	98.45178
Ionosphere	34	20.8	18.8	18.8	92.29376258	92.00804829	**93.15091**
WDBC	30	28	**14**	15	**64.68026**	63.09042	63.09042
WPBC	33	29.4	**5.6**	6.2	76.29555	**77.33536**	75.78273
wine	13	11.8	12	**7**	47.76405	**53.91849**	41.61504
thyroid-disease	5	4.8	4.2	4.2	**77.67442**	76.27907	76.27907
Heart Disease	13	13	**11.8**	12.4	**53.88194**	53.88194	53.88194

从以上表中的实验结果可见，我们提出的 DFS 特征评价准则在 Iris，Dermatology，handwrite，WDBC，thyroid-disease，和 Heart Disease 6 个数据集上进行特征选择后，所得最优特征子集构造的 SVM 分类模型的分类正确率不低于 CFS 准则和我们改进的 CFS 准则——CFSPabs 准则。CFS 准则在 glass，WPBC 和 wine 3 个数据集上的分类正确率超过 DFS 准则和 CFSPabs 准则。我们改进的 CFSPabs 准则只在 Ionosphere 一个数据集上的分类正确率超过了 DFS 准则和 CFS 准则。就选择的特征子集的规模分析，我们改进的 CFSPabs 准则最优，在 Iris，Dermatology，glass，handwrite，Ionosphere，wine，thyroid-disease 共 7 个数聚集上所选择的最优特征子集的规模不超过本章提出的 DFS 准则和 Hall[162]在其博士论文中提出的 CFS 准则。本章提出的 DFS 准则只有在 Iris 数据集上选择的特征子集的规模与其他两个准则选择的特征子集规模持平，在其他 9 个数据集上都不如其他两个准则。CFS 准则在 WDBC，WPBC，Heart Disease 三个数据集上选择的特征子集的规模优于其他两个准则。以上分析显示，就选择的特征子集的规模而言，CFSPabs 准则最优，CFS 准则次之，本章提出的 DFS 准则最差，排第三。但是，就分类正确率来看，本章提出的 DFS 准则第一，最优，CFS 准则第二，CFSPabs 准则第三，

最差。由此可见，基于 DFS 与 SVM 的特征选择算法具有最好的泛化性能，但该算法所得的特征子集的规模不是最小的；基于 CFSPabs 与 SVM 的特征选择算法所选择的特征子集的规模最小（最优），但是该特征选择算法所得分类器的泛化性能不是最优的。

4.3 基于 DFS 与 SVM 的顺序后向混合特征选择

此部分我们提出将 DFS 特征评价准则和经典顺序后向特征搜索策略 SBS 结合，以 SVM 为分类工具的顺序后向特征选择算法——基于 DFS 与 SVM 的顺序后向混合特征选择算法。算法实现在和 4.2 同样划分的数据集上进行，采用相同 SVM 工具箱[160]，SVM 的核函数参数采用默认参数，实验也是 5 折交叉验证实验。算法详细步骤和实验结果分别描述如下。

4.3.1 算法思想描述

确定当前的训练集和测试集。

Step 1：设 S 为全部特征构成的集合，记集合 S 的规模为 n。

Step 2：判断 S 是否为空，若 S 不为空集，则以 S 中特征在训练集上训练 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；否则，若 S 为空，则算法结束。

Step 3：尝试删除 S 中每一个特征，计算 n 个特征数减一的特征子集的 DFS_i 值（或者 Ms 值）。即，计算 n 个规模为 $n-1$ 的特征子集的 DFS_i 值（或者 Ms 值）。

Step 4：删除 DFS_i 或者 Ms 值最大的特征子集对应的特征 k，即置 S=S-k，记新集合 S 的规模为 $n-1$，go to Step 2 。

最后选择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。

4.3.2 实验结果与分析

表 4.13 到表 4.22 展示了基于三种特征子集评价准则——DFS 准则、CFS 准则、CFSPabs 准则与 SVM 的顺序后向特征选择算法在各个数据集上的 5 折交叉验证实验的详细结果。表中加粗的测试集分类正确率值表示最优的测试集分类正确率。表 4.23 给出了分别基于三个不同特征子集评价准则与 SVM 的顺序后向特征选择算法的 5 折交叉验证实验的实验结果平均值比较。表中加粗和加下划线的被选择特征子集规模值表示最优（最小）的特征子集；加粗和加下划线的测试集分类正确率值表示最佳（最高）的分类

正确率，也即最优的分类器泛化能力。

表 4.13 Iris 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.13 the experimental results of SBS on Iris dataset

Iris	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	3	3	3	99.16667	99.16667	99.16667	96.66667	96.66667	96.66667
Fold 2	3	3	3	98.33333	98.33333	98.33333	93.33333	93.33333	93.33333
Fold 3	3	3	3	99.16667	99.16667	99.16667	96.66667	96.66667	96.66667
Fold 4	3	3	3	99.16667	99.16667	99.16667	96.66667	96.66667	96.66667
Fold 5	2	2	2	97.5	97.5	97.5	100	96.66667	96.66667
Average	2.8	2.8	2.8	98.66667	98.66667	98.66667	**96.66667**	96	96

表 4.14 Dermatology 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.14 the experimental results of SBS on Dermatology dataset

Dermatology	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	32	30	28	98.94366	99.29577	99.29577	97.2973	90.54054	87.83784
Fold 2	32	31	28	99.3007	97.9021	98.25175	95.83333	93.05556	94.44444
Fold 3	33	32	28	98.95105	98.95105	98.95105	93.05556	93.05556	93.05556
Fold 4	31	32	28	99.30556	98.26389	98.61111	95.71429	97.14286	97.14286
Fold 5	33	31	27	98.95833	98.95833	98.95833	95.71429	94.28571	94.28571
Average	32.2	31.2	27.8	99.09186	98.67423	98.8136	**95.52295**	93.61604	93.35328

表 4.15 glass 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.15 the experimental results of SBS on glass dataset

Glass	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	7	3	6	96.47059	98.82353	96.47059	93.18182	95.45455	93.18182
Fold 2	8	5	6	95.90643	96.49123	95.90643	90.69767	93.02326	93.02326
Fold 3	8	4	8	95.32164	97.66082	95.32164	97.67442	97.67442	97.67442
Fold 4	8	3	8	97.09302	97.67442	97.09302	95.2381	92.85714	95.2381
Fold 5	8	3	8	97.67442	98.83721	97.67442	90.47619	92.85714	90.47619
Average	7.8	3.6	7.2	96.49322	97.89744	96.49322	93.45364	**94.3733**	93.91876

表 4.16 handwrite 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.16 the experimental results of SBS on handwrite dataset

Handwrite	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	29	127	13	100	100	100	95.45455	98.48485	93.93939
Fold 2	38	123	13	100	100	100	100	96.92308	96.92308
Fold 3	40	126	16	100	100	100	100	98.4375	98.4375
Fold 4	33	128	13	100	100	100	100	100	96.875
Fold 5	40	135	23	100	100	100	98.4375	98.4375	100
Average	36	127.8	15.6	100	100	100	**98.77841**	98.45659	97.23499

表 4.17 Ionosphere 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.17 the experimental results of SBS on Ionosphere dataset

Ionosphere	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	27	22	26	94.64286	93.92857	94.64286	97.1831	98.59155	97.1831
Fold 2	14	17	8	96.08541	95.72954	96.08541	85.71429	87.14286	85.71429
Fold 3	17	18	7	95.72954	95.72954	95.72954	87.14286	90	91.42857
Fold 4	16	23	7	94.30605	94.30605	94.30605	97.14286	97.14286	94.28571
Fold 5	25	16	28	94.66192	94.66192	95.01779	94.28571	92.85714	94.28571
Average	19.8	19.2	15.2	95.08516	94.87112	95.15633	92.29376	**93.14688**	92.57948

表 4.18 WDBC 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.18 the experimental results on WDBC dataset

WDBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	27	7	10	100	100	100	66.08696	66.08696	67.82609
Fold 2	27	11	12	100	100	100	61.73913	63.47826	64.34783
Fold 3	27	12	13	100	100	100	62.83186	62.83186	62.83186
Fold 4	27	14	14	100	100	100	67.25664	62.83186	62.83186
Fold 5	27	13	14	100	100	100	65.48673	62.83186	62.83186
Average	27	11.4	12.6	100	100	100	**64.68026**	63.61216	64.1339

表 4.19 WPBC 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.19 the experimental results of SBS on WPBC dataset

WPBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	28	9	2	100	100	100	75	75	75
Fold 2	28	7	1	100	100	100	76.92308	76.92308	74.35897
Fold 3	28	10	8	100	100	100	76.92308	76.92308	76.92308
Fold 4	29	7	5	100	100	100	76.31579	76.31579	76.31579
Fold 5	29	8	5	100	100	100	76.31579	76.31579	76.31579
Average	28.4	8.2	4.2	100	100	100	**76.29555**	76.29555	75.78273

表 4.20 Wine 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.20 the experimental results of SBS on Wine dataset

Wine	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	9	8	5	100	100	100	56.75676	43.24324	43.24324
Fold 2	12	8	5	100	100	100	36.11111	36.11111	36.11111
Fold 3	12	8	10	100	100	100	41.66667	38.88889	38.88889
Fold 4	9	6	7	100	100	100	54.28571	54.28571	57.14286
Fold 5	12	8	5	100	100	100	50	47.05882	44.11765
Average	10.8	7.6	6.4	100	100	100	**47.76405**	43.91756	43.90075

表 4.21 thyroid-disease 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.21 the experimental results of SBS on thyroid-disease dataset

thyroid-disease	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	4	4	4	100	100	100	74.4186	74.4186	74.4186
Fold 2	4	4	3	100	100	100	76.74419	76.74419	76.74419
Fold 3	4	4	4	100	100	100	76.74419	76.74419	76.74419
Fold 4	4	4	4	100	100	100	74.4186	74.4186	74.4186
Fold 5	3	4	4	100	100	100	86.04651	74.4186	74.4186
Average	3.8	4	3.8	100	100	100	**77.67442**	75.34884	75.34884

表 4.22 Heart Disease 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.22 the experimental results of SBS on Heart Disease dataset

Heart Disease	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	12	9	12	100	100	100	53.33333	53.33333	53.33333
Fold 2	12	11	10	100	100	100	53.33333	53.33333	53.33333
Fold 3	12	9	12	100	100	100	53.33333	53.33333	53.33333
Fold 4	12	12	10	100	100	100	54.23729	54.23729	54.23729
Fold 5	12	12	12	100	100	100	55.17241	55.17241	55.17241
Average	12	10.6	11.2	100	100	100	53.88194	53.88194	53.88194

表 4.23 UCI 数据集顺序后向特征选择算法 5 折交叉验证实验结果

Table 4.23 the experimental results of SBS on UCI datasets

Data sets	原始特征数	被选择特征数			测试集分类正确率		
		DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**

Iris	4	2.8	2.8	2.8	**96.66667**	96	96
dermatology	34	32.2	31.2	**27.8**	**95.52295**	93.61604	93.35328
glass	9	7.8	**3.6**	7.2	93.45364	**94.3733**	93.91876
handwrite	255	36	127.8	**15.6**	**98.77841**	98.45659	97.23499
Ionosphere	34	19.8	19.2	**15.2**	92.29376	**93.14688**	92.57948
WDBC	30	27	**11.4**	12.6	**64.68026**	63.61216	64.1339
WPBC	33	28.4	8.2	**4.2**	**76.29555**	76.29555	75.78273
wine	13	10.8	7.6	**6.4**	**47.76405**	43.91756	43.90075
thyroid-disease	5	4	3.8	3.8	**77.67442**	75.34884	75.34884
Heart Disease	13	12	**10.6**	11.2	**53.88194**	53.88194	53.88194

从以上表中的实验结果可见，我们提出的 DFS 特征评价准则在 Iris，Dermatology，handwrite，WDBC，WPBC，wine，thyroid-disease 和 Heart Disease 共 8 个数据集上进行特征选择所得的最优特征子集对应的 SVM 分类模型具有最优的分类正确率；HALL 的 CFS 准则在 glass 和 Ionosphere 两个数据集上的最优特征子集具有最高的分类正确率；我们改进的 CFS 准则——CFSPabs 准则在分类正确率上不占优势。最优特征子集的规模分析可见，我们改进的 CFSPabs 准则最优，在 Iris，Dermatology，handwrite，Ionosphere，WPBC，wine，thyroid-disease 共 7 个数聚集上所选择的最优特征子集的规模不超过本章提出的 DFS 准则和 Hall 提出的 CFS 准则；Hall 的 CFS 准则只在 glass，WDBC 和 Heart disease 3 个数据集上最优；本章提出的 DFS 准则只有在 Iris 数据集上选择的特征子集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。以上分析显示，就选择的特征子集规模而言，我们改进的 CFSPabs 准则最优，CFS 准则次之，本章提出的 DFS 准则最差，排第三。但是，就分类正确率来看，本章提出的 DFS 准则第一，最优，CFS 准则第二，CFSPabs 准则第三，最差。

由此可见，由基于 DFS 与 SVM 的顺序后向特征选择算法所选择的特征子集构造所得的分类器，具有最优的泛化性能，但是所选择的特征子集的规模不是最小的；而基于 CFSPabs 与 SVM 的特征选择算法可以得到最优规模的特征子集，但是以该特征子集构造的分类器不具有最好的泛化性能。这一结论与基于 DFS 与 SVM 的顺序前向特征选择算法的实验结果一致。

4.4 基于 DFS 与 SVM 的顺序前向浮动混合特征选择

本部分将 DFS 特征评价准则与经典的顺序前向浮动搜索策略 SFFS 相结合，提出基于 DFS 与 SVM 的顺序前向浮动混合特征选择算法。该算法以 SVM 为分类工具，以相应特征子集的分类正确率来判定刚加入的特征是否保留；特征的加入原则是依据 DFS 准则加入，即依据相应特征子集的 DFS 区分度判定相应特征加入与否。

实验采用与 4.2 节和 4.3 节相同的数据集，且实验数据集的划分同 4.2 和 4.3 节部分的实验数据集划分相同，即本部分实验与前两部分的实验在同样划分的数据集上进行，实验采用的SVM工具箱也与4.2节和4.3节相同，SVM 的核函数参数也采用默认参数，实验也是 5 折交叉验证实验。算法的详细步骤描述，以及相应的实验结果分别见下面 4.4.1 和 4.4.2 部分。

4.4.1 算法思想描述

设 S 为包含全部特征的集合，C 为被选择特征构成的子集，C 初始为空集。确定当前的训练集和测试集。

Step 1：在训练集上计算每个特征的 G-score 值，选择 G-score 值最大的特征 k 加入被选特征子集 C，并从 S 集合中删去特征 k，即置 C=C+k，S=S-k。

Step 2：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。

Step 3：判断 S 是否为空，若 S 不为空集，尝试将被选特征子集 C 与 S 中每一个特征组合，构成特征数增 1 的临时特征子集 C，计算相应临时特征子集 C 的 DFS_i 值（或者 Ms 值），选择具有最大 DFS_i 值的临时特征子集 C 对应的特征 k 加入特征子集 C，即置 C=C+k，S=S-k；否则，若 S 为空，则算法结束。

Step 4：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。

Step 5：若训练集的分类正确率提高，则转 Step 3；否则，从特征子集 C 中去掉刚加入的那个特征 k，即置 C=C－k，然后转 Step 3。

最后留在被选择特征子集 C 中的特征构成最优被选择特征子集。

4.4.2 实验结果与分析

表 4.24 到表 4.33 展示了三种顺序前向混合特征选择算法的 5 折交叉验证实验的详细结果。这三个混合特征选择算法的特点是：特征加入过程分别以特征子集区分度评价准则——DFS 准则、CFS 准则、CFSPabs 准则为依据判断加入相应的特征，而在浮动删除相应特征的时候，则是以训练集的分类正确率为依据，删除那些对提高分类正确率没有贡献的特征。表 4.34 比较了这三个基于不同特征子集评价准则与 SVM 的顺序前向混合特征选择算法的实验结果的平均值。表 4.24 到表 4.33 中加粗的测试集分类正确率表示

最高的测试集分类正确率，表 4.34 中加粗和加下划线的测试集分类正确率表示最佳的分类正确率；加粗和加下划线的被选择特征数值是最小（最佳）的被选择特征子集规模。

表 4.24 Iris 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.24 the experimental results of SFFS on Iris dataset

Iris	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	3	4	4	98.33333	99.16667	99.16667	96.66666667	96.66666667	96.66667
Fold 2	2	2	2	97.5	97.5	97.5	96.66666667	96.66666667	96.66667
Fold 3	4	4	4	99.16667	99.16667	99.16667	96.66666667	96.66666667	96.66667
Fold 4	4	4	4	99.16667	99.16667	99.16667	96.66666667	96.66666667	96.66667
Fold 5	3	3	3	97.5	97.5	97.5	100	96.66666667	96.66667
Average	3.2	3.4	3.4	98.33333	98.5	98.5	**97.33333333**	96.66666667	96.66667

表 4.25 Dermatology 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.25 the experimental results of SFFS on Dermatology dataset

Dermatology	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	15	12	9	98.94366	98.94366	97.53521	93.24324324	82.43243243	95.94595
Fold 2	14	15	12	98.25175	99.65035	97.55245	95.83333333	87.5	93.05556
Fold 3	14	10	12	98.25175	98.6014	98.95105	88.88888889	95.83333333	98.61111
Fold 4	14	13	10	98.95833	97.91667	99.30556	95.71428571	94.28571429	95.71429
Fold 5	14	10	12	97.91667	98.61111	98.26389	100	100	100
Average	14.2	12	11	98.46443	98.74464	98.32163	94.73595024	92.01029601	**96.66538**

表 4.26 Glass 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.26 the experimental results of SFFS on Glass dataset

Glass	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	6	6	5	97.64706	97.64706	97.64706	93.18181818	93.18181818	95.45455
Fold 2	3	7	3	95.32164	96.49123	95.32164	93.02325581	93.02325581	93.02326
Fold 3	5	4	4	95.90643	96.49123	95.32164	93.02325581	97.6744186	90.69767
Fold 4	4	3	3	97.09302	95.93023	95.93023	92.85714286	90.47619048	90.47619
Fold 5	5	5	4	97.67442	97.67442	97.67442	88.0952381	90.47619048	88.09524
Average	4.6	5	3.8	96.72851	96.84683	96.379	92.03614215	**92.96637471**	91.54938

表 4.27 Handwrite 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.27 the experimental results of SFFS on Handwrite dataset

Handwrite	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	8	9	6	100	98.83268	100	96.96969697	92.42424242	93.93939
Fold 2	7	11	6	98.83721	99.6124	100	98.46153846	98.46153846	96.92308
Fold 3	9	8	2	99.6139	98.8417	98.0695	98.4375	93.75	93.75
Fold 4	10	13	8	99.2278	99.6139	100	100	96.875	98.4375
Fold 5	11	7	5	99.6139	96.5251	100	100	96.875	93.75
Average	9	9.6	5.4	99.45856	98.68516	99.6139	**98.77374709**	95.67715618	95.35999

表 4.28 Ionosphere 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.28 the experimental results of SFFS on Ionosphere dataset

Ionosphere	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	11	9	8	95.35714	94.64286	94.64286	95.77464789	95.77464789	95.77465
Fold 2	9	12	8	95.37367	95.72954	96.08541	88.57142857	87.14285714	88.57143
Fold 3	8	8	8	95.72954	95.01779	95.72954	91.42857143	87.14285714	91.42857
Fold 4	10	10	10	94.66192	95.01779	95.01779	95.71428571	97.14285714	97.14286
Fold 5	10	8	7	94.66192	94.66192	94.30605	91.42857143	90	94.28571
Average	9.6	9.4	8.2	95.15684	95.01398	95.15633	92.58350101	91.44064386	**93.44064**

表 4.29 WDBC 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.29 the experimental results of SFFS on WDBC dataset

WDBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	8	5	5	100	100	100	93.04347826	62.60869565	62.6087

Fold 2	7	3	3	100	100	100	63.47826087	60.86956522	60.86957
Fold 3	9	6	6	100	100	100	62.83185841	92.92035398	92.92035
Fold 4	11	8	8	100	100	100	63.71681416	95.57522124	95.57522
Fold 5	9	9	9	100	100	100	62.83185841	62.83185841	62.83186
Average	8.8	6.2	6.2	100	100	100	69.18045402	74.9611389	**74.96114**

表 4.30 WPBC 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.30 the experimental results of SFFS on WPBC dataset

WPBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	9	4	5	100	100	100	75	77.5	75
Fold 2	10	3	2	100	100	100	76.92307692	74.35897436	74.35897
Fold 3	8	7	6	100	100	100	76.92307692	76.92307692	76.92308
Fold 4	9	4	3	100	100	100	76.31578947	76.31578947	76.31579
Fold 5	8	4	3	100	100	100	76.31578947	81.57894737	76.31579
Average	8.8	4.4	3.8	100	100	100	76.29554656	**77.33535762**	75.78273

表 4.31 Wine 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.31 the experimental results of SFFS on Wine dataset

Wine	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	7	7	5	100	100	100	51.35135135	43.24324324	51.35135
Fold 2	8	6	6	100	100	100	38.88888889	80.55555556	38.88889
Fold 3	8	7	5	100	100	100	38.88888889	38.88888889	38.88889
Fold 4	7	7	5	100	100	100	51.42857143	45.71428571	57.14286
Fold 5	7	7	5	100	100	100	58.82352941	44.11764706	44.11765
Average	7.4	6.8	5.2	100	100	100	47.87624599	**50.50392409**	46.07793

表 4.32 thyroid-disease 数据集顺序前向浮动特征选择 5 折交叉验证实验结果

Table 4.32 the experimental results of SFFS on thyroid-disease dataset

thyroid-disease	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	4	4	4	100	100	100	74.41860465	79.06976744	79.06977
Fold 2	5	4	4	100	100	100	76.74418605	76.74418605	76.74419
Fold 3	5	5	5	100	100	100	76.74418605	76.74418605	76.74419
Fold 4	4	4	4	100	100	100	72.09302326	74.41860465	74.4186
Fold 5	4	4	4	100	100	100	86.04651163	74.41860465	74.4186
Average	4.4	4.2	4.2	100	100	100	**77.20930233**	76.27906977	76.27907

表 4.33 Heart disease 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.33 the experimental results of SFFS on Heart disease dataset

Heart Disease	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	7	9	7	100	100	100	53.33333333	53.33333333	53.33333
Fold 2	12	9	10	100	100	100	53.33333333	53.33333333	53.33333
Fold 3	8	9	9	100	100	100	53.33333333	53.33333333	53.33333
Fold 4	11	7	8	100	100	100	54.23728814	54.23728814	54.23729
Fold 5	12	10	11	100	100	100	55.17241379	55.17241379	55.17241
Average	10	8.8	9	100	100	100	53.88194039	53.88194039	53.88194

表 4.34 UCI 数据集顺序前向浮动特征选择算法 5 折交叉验证实验结果

Table 4.34 the experimental results of SFFS on UCI datasets

Data sets	原始特征数	被选择特征数			测试集		
		DFS	CFS	CFSPabs	DFS	CFS	CFSPabs
Iris	4	**3.2**	3.4	3.4	**97.33333333**	96.66666667	96.66667
dermatology	34	14.2	12	**11**	94.73595024	92.01029601	**96.66538**
glass	9	4.6	5	**3.8**	92.03614215	**92.96637471**	91.54938
handwrite	255	9	9.6	**5.4**	**98.77374709**	95.67715618	95.35999
Ionosphere	34	9.6	9.4	**8.2**	92.58350101	91.44064386	**93.44064**
WDBC	30	8.8	6.2	6.2	69.18045402	74.9611389	**74.96114**

WPBC	33	8.8	4.4	**3.8**	76.29554656	**77.33535762**	75.78273
wine	13	7.4	6.8	**5.2**	47.87624599	**50.50392409**	46.07793
thyroid-disease	5	4.4	4.2	4.2	**77.20930233**	76.27906977	76.27907
Heart Disease	13	10	**8.8**	9	53.88194039	53.88194039	53.88194

从以上表中的实验结果可见，本章提出的基于 DFS 特征评价准则与 SVM 的顺序前向浮动特征选择算法只在 Iris，handwrite，和 thyroid-disease 共 3 个数据集上的最优特征子集对应的 SVM 分类模型的分类正确率最高；基于 HALL 提出的 CFS 准则与 SVM 的顺序前向浮动特征选择算法在 glass，WPBC 和 wine 三个数据集上的最优特征子集具有最高的分类正确率；基于我们改进的 CFS 准则——CFSPabs 准则与 SVM 的顺序前向浮动特征选择算法在 Dermatology，Ionosphere 和 WDBC 共 3 个数聚集上的最优特征子集的分类正确率最优;对于 Heart disease 数据集分别基于三个特征子集评价准则与 SVM 的顺序前向浮动特征选择算法的分类正确率相同。

就最优特征子集规模分析可见，基于我们改进的 CFSPabs 准则与 SVM 的顺序前向浮动特征选择最优，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的规模不超过本章提出的基于 DFS 与 SVM 的顺序前向浮动特征选择算法，也不超过基于 CFS 与 SVM 的顺序前向浮动特征选择算法所选择的特征子集的规模，其中在 WDBC 和 thyroid-disease 两个数据集上的特征子集规模与基于 Hall 提出的 CFS 准则与 SVM 的顺序前向浮动特征选择算法相同；基于 Hall 的 CFS 准则与 SVM 的顺序前向浮动特征选择算法只 Heart disease 数据集上最优，在 Iris，WDBC 和 thyroid-disease 三个数据集上的最优特征子集的规模与基于 CFSPabs 准则与 SVM 的顺序前向浮动特征选择算法持平，在其他 6 个数据集的最优特征子集的规模不如基于其他两个准则与 SVM 的顺序前向浮动特征选择算法好；本章提出的基于 DFS 与 SVM 的顺序前向浮动特征选择算法只有在 Iris 数据集上选择的特征子集最优，在 glass 和 handwrite 两个数据集上选择的特征子集的规模介于分别基于其他两个准则与 SVM 的顺序前向浮动特征选择算法之间，其他 7 个数据集上都不如分别基于其他两个准则与 SVM 的顺序前向浮动特征选择算法。

以上分析显示，就特征子集的规模来看，基于我们改进的 CFSPabs 准则与 SVM 的顺序前向浮动混合特征选择算法所选择的特征子集最优；基于 CFS 准则与 SVM 的顺序前向浮动特征选择算法次之；基于 DFS 与 SVM 的顺序前向浮动混合特征选择算法所选择的特征子集最差。但是，就分类正确率来看，基于 DFS 与 SVM 的顺序前向浮动特征选择算法、基于 CFS 与 SVM 的顺序前向浮动特征选择算法，以及基于 CFSPabs 与 SVM

的顺序前向浮动特征选择算法的性能基本持平。

4.5 基于 DFS 与 SVM 的顺序后向浮动混合特征选择

本章前边部分分别介绍了基于 DFS 与 SVM 的顺序前向、顺序后向特征选择算法，以及顺序前向浮动特征选择算法，并对实验结果进行了详细分析。本小结将 DFS 特征子集区分度评价准则与 SVM 结合，并以顺序后向浮动搜索策略进行特征搜索，提出基于 DFS 特征子集区分度准则与 SVM 的顺序后向浮动混合特征选择算法。该算法的详细步骤描述和实验结果与分析分别见 4.5.1 和 4.5.2 小节部分。

4.5.1 算法思想描述

设 S 为包含全部特征的集合，C 为被选择特征构成的子集，C 初始化为包含所有特征的特征全集。确定当前的训练集和测试集。

Step 1：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。

Step 2：判断 S 是否为空，若 S 不为空集，尝试删除 S 中的每一个特征，得到特征数少 1 的临时特征子集，计算每一个临时特征子集的 DFS_i 值（或者 Ms 值），选择具有最大的 DFS_i 值的临时特征子集对应的特征 k 从 S 中删除，即置 S=S-k；否则，若 S 为空，则算法终止。

Step 3：使用 S 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。

Step 4：若训练集的分类正确率提高或者不变，则置 C=C－k。

Step 5：go to Step 2。

最后留在被选择特征子集 C 中的特征构成该算法的所选择的最优被选择特征子集。

4.5.2 实验结果与分析

本部分实验的数据集同本章第 2、3、4 部分，实验同样采用 5 折交叉验证实验，并在相同划分的数据集上进行实验，实验使用相同的 SVM 工具箱，默认的 SVM 核函数参数。表 4.35 到表 4.44 展示了三种顺序后向浮动混合特征选择算法的 5 折交叉验证实验的详细结果。表中加粗的测试集分类正确率为最高的分类正确率值。这三个顺序后向浮动混合特征选择算法的特点是：特征试剔除过程分别以特征子集区分度评价准则——DFS 准则、CFS 准则、CFSPabs 准则为依据进行判断，而浮动加入过程则是以相应

的训练集分类正确率为依据，召回那些导致分类正确率降低的特征，而真正删除那些对提高分类正确率没有贡献的特征。表 4.45 对这三个基于不同特征子集区分度评价准则与 SVM 的顺序后向浮动混合特征选择算法的实验结果平均值进行了比较。表中加黑加下划线的被选择特征数表示最小（优）的被选择特征子集规模，表中加黑加下划线的测试集分类正确率表示最高（优）的分类正确率。

表 4.35 Iris 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.35 the experimental results of SBFS on Iris dataset

Iris	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	3	3	3	99.16667	99.16667	99.16667	96.66666667	96.66666667	96.66667
Fold 2	3	3	3	98.33333	98.33333	98.33333	93.33333333	93.33333333	93.33333
Fold 3	3	3	3	99.16667	99.16667	99.16667	96.66666667	96.66666667	96.66667
Fold 4	3	3	3	99.16667	99.16667	99.16667	96.66666667	96.66666667	96.66667
Fold 5	2	1	1	97.5	97.5	97.5	100	93.33333333	93.33333
Average	2.8	2.6	2.6	98.66667	98.66667	98.66667	**96.66666667**	95.33333333	95.33333

表 4.36 Dermatology 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.36 the experimental results of SBFS on Dermatology dataset

Dermatology	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	16	15	15	98.94366	99.29577	99.29577	95.94594595	89.18918919	89.18919
Fold 2	12	12	14	99.65035	98.6014	99.3007	94.44444444	87.5	90.27778
Fold 3	14	15	12	99.3007	99.3007	98.95105	88.88888889	83.33333333	88.88889
Fold 4	12	24	12	99.30556	98.61111	99.30556	97.14285714	95.71428571	97.14286
Fold 5	15	16	13	99.30556	99.65278	98.95833	87.14285714	85.71428571	84.28571
Average	13.8	16.4	13.2	99.30116	99.09235	99.16228	**92.71299871**	88.29021879	89.95689

表 4.37 Glass 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.37 the experimental results of SBFS on Glass dataset

Glass	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	5	3	4	97.05882	98.82353	96.47059	88.63636364	95.45454545	93.18182
Fold 2	3	4	4	95.90643	96.49123	96.49123	90.69767442	90.69767442	93.02326
Fold 3	2	3	4	96.49123	97.66082	96.49123	97.6744186	95.34883721	97.67442
Fold 4	5	5	3	97.67442	97.67442	97.09302	95.23809524	95.23809524	95.2381
Fold 5	2	3	2	97.67442	98.83721	97.67442	92.85714286	92.85714286	92.85714
Average	3.4	3.6	3.4	96.96106	97.89744	96.8441	93.02073895	93.91925904	**94.39495**

表 4.38 Handwrite 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.38 the experimental results of SBFS on Handwrite dataset

Handwrite	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	15	39	10	100	100	100	96.96969697	96.96969697	90.90909
Fold 2	9	33	10	100	100	100	98.46153846	98.46153846	96.92308
Fold 3	8	24	12	100	100	100	100	100	93.75
Fold 4	8	27	14	100	100	100	98.4375	100	98.4375
Fold 5	8	30	10	100	100	100	98.4375	93.75	98.4375
Average	9.6	30.6	11.2	100	100	100	**98.46124709**	97.83624709	95.69143

表 4.39 Ionosphere 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.39 the experimental results of SBFS on Ionosphere dataset

Ionosphere	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	13	15	15	95.71429	94.28571	95	95.77464789	97.18309859	94.3662
Fold 2	15	10	11	95.72954	95.72954	95.72954	87.14285714	90	88.57143
Fold 3	16	19	17	95.72954	96.79715	96.79715	87.14285714	91.42857143	91.42857
Fold 4	20	19	16	94.66192	94.66192	94.30605	95.71428571	97.14285714	95.71429

Fold 5	12	14	13	95.01779	95.01779	95.01779	94.28571429	94.28571429	92.85714
Average	15.2	15.4	14.4	95.37062	95.29842	95.37011	92.01207243	**94.00804829**	92.58753

表 4.40 WDBC 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.40 the experimental results of SBFS on WDBC dataset

WDBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	4	3	4	100	100	100	62.60869565	62.60869565	64.34783
Fold 2	3	3	3	100	100	100	62.60869565	62.60869565	62.6087
Fold 3	5	3	4	100	100	100	63.71681416	62.83185841	62.83186
Fold 4	4	3	3	100	100	100	62.83185841	63.71681416	63.71681
Fold 5	4	3	3	100	100	100	62.83185841	61.94690265	62.83186
Average	4	3	3.4	100	100	100	62.91958446	62.74259331	**63.26741**

表 4.41 WPBC 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.41 the experimental results of SBFS on WPBC dataset

WPBC	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	4	2	1	100	100	100	75	75	75
Fold 2	4	2	1	100	100	100	76.92307692	76.92307692	74.35897
Fold 3	5	2	2	100	100	100	76.92307692	76.92307692	76.92308
Fold 4	4	1	2	100	100	100	76.31578947	76.31578947	76.31579
Fold 5	4	2	2	100	100	100	76.31578947	76.31578947	76.31579
Average	4.2	1.8	1.6	100	100	100	**76.29554656**	**76.29554656**	75.78273

表 4-42 Wine 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4-42 the experimental results of SBFS on Wine dataset

Wine	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	2	1	2	100	100	100	54.05405405	43.24324324	43.24324
Fold 2	2	2	3	100	100	100	38.88888889	36.11111111	38.88889
Fold 3	2	2	3	100	100	100	38.88888889	38.88888889	38.88889
Fold 4	2	2	3	100	100	100	54.28571429	51.42857143	57.14286
Fold 5	2	2	3	100	100	100	44.11764706	44.11764706	44.11765
Average	2	1.8	2.8	100	100	100	**46.04703864**	42.75789235	44.45631

表 4.43 thyroid-disease 数据集顺序后向浮动特征选择 5 折交叉验证实验结果

Table 4.43 the experimental results of SBFS on thyroid-disease dataset

thyroid-disease	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	2	3	3	100	100	100	76.74418605	79.06976744	79.06977
Fold 2	2	3	3	100	100	100	76.74418605	76.74418605	76.74419
Fold 3	3	4	4	100	100	100	74.41860465	76.74418605	76.74419
Fold 4	2	3	3	100	100	100	74.41860465	74.41860465	74.4186
Fold 5	2	3	3	100	100	100	90.69767442	86.04651163	74.4186
Average	2.2	3.2	3.2	100	100	100	**78.60465116**	78.60465116	76.27907

表 4.44 Heart disease 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.44 the experimental results of SBFS on Heart disease dataset

Heart Disease	被选择特征数			训练集分类正确率			测试集分类正确率		
	DFS	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**	**DFS**	**CFS**	**CFSPabs**
Fold 1	3	4	4	100	100	100	53.33333333	53.33333333	53.33333
Fold 2	3	4	3	100	100	100	53.33333333	53.33333333	53.33333
Fold 3	5	5	4	100	100	100	51.66666667	53.33333333	53.33333
Fold 4	3	4	3	100	100	100	54.23728814	54.23728814	52.54237
Fold 5	3	6	4	100	100	100	55.17241379	55.17241379	55.17241
Average	3.4	4.6	3.6	100	100	100	53.54860705	**53.88194039**	53.54296

表 4.45 UCI 数据集顺序后向浮动特征选择算法 5 折交叉验证实验结果

Table 4.45 the experimental results of SBFS on UCI datasets

Data sets	原始特征数	被选择特征数			测试集分类正确率		
		DFS	CFS	CFSPabs	DFS	CFS	CFSPabs
Iris	4	2.8	**2.6**	**2.6**	**96.66666667**	95.33333333	95.33333
dermatology	34	13.8	16.4	**13.2**	**92.71299871**	88.29021879	89.95689
glass	9	**3.4**	3.6	**3.4**	93.02073895	93.91925904	**94.39495**
handwrite	255	**9.6**	30.6	11.2	**98.46124709**	97.83624709	95.69143
Ionosphere	34	15.2	15.4	**14.4**	92.01207243	**94.00804829**	92.58753
WDBC	30	4	**3**	3.4	62.91958446	62.74259331	**63.26741**
WPBC	33	4.2	1.8	**1.6**	**76.29554656**	**76.29554656**	75.78273
wine	13	2	**1.8**	2.8	**46.04703864**	42.75789235	44.45631
thyroid-disease	5	**2.2**	3.2	3.2	**78.60465116**	**78.60465116**	76.27907
Heart Disease	13	**3.4**	4.6	3.6	53.54860705	**53.88194039**	53.54296

以上表中实验结果显示，本部分提出的基于 DFS 特征子集区分度评价准则与 SVM 的顺序后向浮动特征选择算法在 Iris，dermatology，handwrite，WPBC，wine 和 thyroid-disease 共 6 个数据集上所得的最优特征子集对应的 SVM 分类模型的分类正确率高于等于基于 Hall 的 CFS 特征子集评价准则与 SVM 的顺序后向浮动特征选择算法，并高于等于基于我们在本章改进的 CFSPabs 特征子集评价准则与 SVM 的顺序后向浮动特征选择算法所选择的最优特征子集对应的 SVM 分类模型的分类正确率，其中在 WPBC 和 thyroid-disease 两个数据集上的分类正确率相比，基于 DFS 与 SVM 的顺序后向浮动特征选择算法和基于 CFS 与 SVM 的顺序后向浮动特征选择算法持平；基于 HALL 的 CFS 准则与 SVM 的顺序后向浮动特征选择算法在 Ionosphere 和 Heart Disease 两个数据集上最高，在 WPBC 和 thyroid-disease 数据集上同本章提出的基于 DFS 准则与 SVM 的后向浮动特征选择算法相同；基于我们改进的 CFS 准则——CFSPabs 准则与 SVM 的顺序后向浮动特征选择算法在 glass，WDBC 两个数据集上的最优特征子集的分类正确率最高。

分析最优特征子集规模可见，基于 DFS 准则与 SVM 的顺序后向浮动特征选择算法在 glass，handwrite，thyoid-disease 和 heart Disease 这 4 个数据集上的最优特征子集规模不高于分别基于 CFS 和 CFSPabs 两个准则与 SVM 顺序后向浮动特征选择算法，其中在 glass 数据集上与基于 CFSPabs 准则与 SVM 的顺序后向浮动特征选择算法持平，在 handwrite，thyoid-disease 和 heart Disease 这 3 个数据集上优于分别基于 CFS 与 CFSPabs 两个准则与 SVM 的顺序后向浮动特征选择算法；基于 Hall 的 CFS 特征子集评价准则与 SVM 的后向浮动特征选择算法在 Iris，WDBC 和 wine 这 3 个数据集上不高于分别基于 DFS 和 CFSPabs 两个准则与 SVM 的顺序后向浮动特征选择算法，其中在在 Iris 数据集上与基于 CFSPabs 准则与 SVM 的顺序后向浮动特征选择算法持平；基于 CFSPabs 与 SVM 的顺序后向浮动特征选择算法在 Iris，dermatology，galss，Ionosphere 和 WPBC 这

5 个数聚集上的最优特征子集规模不高于分别基于 DFS 和 CFS 与 SVM 的顺序后向浮动特征选择算法，其中在 Iris 数据集上与基于 Hall 的 CFS 准则与 SVM 的顺序后向浮动特征选择算法相同，在 glass 数据集上与基于 DFS 与 SVM 的顺序后向浮动特征选择算法的特征子集规模相同。

以上分析显示，就特征子集的规模而言，基于 CFSPabs 准则与 SVM 的顺序后向浮动特征选择算法最优，基于 DSF 特征子集评价准则与 SVM 的顺序后向浮动混合特征选择算法次之，基于 CFS 准则与 SVM 的顺序后向浮动特征选择算法最差。但是，就分类正确率来看，基于 DFS 与 SVM 的顺序后向浮动特征选择算法最优，基于 Hall 的 CFS 准则与 SVM 的顺序后向浮动特征选择算法次之，基于 CFSPabs 准则与 SVM 的顺序后向浮动特征选择算法最差。

4.6 小结

本章提出了基于 DFS 与 SVM 的 4 种混合特征选择算法，这些算法充分考虑了特征之间的相关性，以特征子集对于分类的联合贡献度量整个特征子集的类间区分度，解决了第 2 章基于 G-score 与 SVM 的特征选择算法，与第 3 章基于 D-score 与 SVM 的特征选择算法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。其中的 DFS 是一种特征子集区分度衡量准则，克服了第二章的 G-score 特证重要性评价准则和第三章的 D-score 特征重要性评价准则没有考虑特征之间相关性的缺陷。该准则的实值是：在 G-score 准则的基础上，考虑特征之间的相互作用，计算多个特征构成的特征子集的联合 G-score 值，以判断特征子集中所有特征对于分类的联合贡献，以此作为特征子集在类别间区分度的大小，即作为特征子集对分类贡献大小的度量，提出了特征子集的区分度概念，并以此为特征选择的依据。

本章首先提出了适用于两类和多类问题的 DFS 特征子集区分度概念，然后结合经典的顺序前向搜索 SFS、顺序后向搜索 SBS、顺序前向浮动搜索 SFFS、顺序后向浮动搜索 SBFS 四种特征搜索策略，以 SVM 为分类工具，得到 4 种基于 DFS 与 SVM 的混合特征选择算法。

为了证明本章提出的基于 DFS 与 SVM 的特征选择算法的有效性，同时也验证 DFS 特征子集评价准则的有效性，我们首先根据 Hall 提出的 CFS 准则的意义提出改进的 CFS 准则——CFSPabs，然后将分别基于 DFS、CFS，以及 CFSPabs 与 SVM 的特征选择算法进行实验比较。

UCI 机器学习数据库的 10 个经典数据集的 5 折交叉验证实验证明：本章提出的 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评价准则，基于该准则与 SVM 的混合特征选择方法所选择的特征具有较好的分类效果，其分类性能优于分别基于 CFS 准则与 SVM 和 CFSPabs 准则与 SVM 的混合特征选择方法所选择的特征子集的分类性能，达到了保持数据集辨识能力不变情况下进行维数压缩的目的。实验比较结果同时显示：基于 DFS 与 SVM 的特征选择算法所选择的特征子集的规模可能不是最优的；就选择的特征子集规模来看，基于 CFSPabs 与 SVM 的特征选择算法最优。

第五章　基于 SVM 分类模型的特征选择

第 2～4 章分别介绍了推广的 F-score 特征重要性评价准则——G-score 准则、G-score 准则的改进准则——D-score 准则，以及联合 F-score 特征评价准则——DFS 特征子集区分度评价准则。其中前两个准则只衡量单个特征对于分类的贡献，联合 F-score 准则是一个特征子集评价准则，考虑了特征子集中特征的联合作用，所以我们称其为 DFS 特征子集区分度评价准则。DFS 特征子集评价准则克服了 G-score 和 D-score 准则没有考虑特征之间相互作用的缺陷。

基于 G-score 准则、D-score 准则、DFS 准则，分别结合不同的特征搜索策略，以 SVM 为分类工具，我们提出了多个混合的特征选择算法，并用 UCI 机器学习数据库的数据集对算法进行了实验测试和比较，证明了基于以上准则与 SVM 的特征选择算法在特征选择中的有效性。

然而，无论以 G-score 准则、D-score 准则，还是联合 F-score 特征评价准则——DFS 特征子集区分度评价准则度量特征或特征子集对于分类的贡献，基于这些准则与 SVM 的特征选择算法，SVM 都只是作为分类工具，以其分类性能引导特征搜索过程。特征重要性评价，或者说特征对于分类的贡献大小没有直接使用 SVM 分类器来度量，而是以独立于 SVM 分类器的 G-score、D-score 和 DFS 准则来度量。这些度量方法均基于类内、类间距离，对于非线性可分的分类问题，有可能造成重要区分特征的区分能力度量有误，从而被错误剔除的潜在危险。因此，本章提出基于 SVM 分类模型的特征选择算法，以 SVM 分类器的权重来度量相应特征的区分度，进行特征选择，充分利用 SVM 对于非线性可分问题的良好泛化性能，解决第 2～4 章的特征选择算法在处理非线性可分的分类问题时的潜在缺陷。

本章内容组织如下，首先提出基于 SVM 分类模型的特征重要性评价方法；然后提出基于此模型的两种特征选择算法——SVM RFA（SVM Recursive Feature Addition）特征选择算法，和推广的 SVM RFE（SVM Recursive Feature Elimination）特征选择算法；并用 UCI 机器学习数据库的数据集对这两种特征选择算法进行了实验测试和比较。

5.1　基于 SVM 分类模型的特征重要性评价方法

SVM 分类模型为 $g(\boldsymbol{x})=\langle \boldsymbol{w}\cdot\boldsymbol{x}\rangle+b$，其中 $\boldsymbol{w}=\sum_{i=1}^{m}y_i\alpha_i\boldsymbol{x}_i$ 是权向量，$\boldsymbol{x}$ 是样本，α_i

对应相应样本的拉格朗日乘子，α_i 不为 0 的向量 x_i 是支持向量。这样，根据 SVM 分类模型可以得到样本每个分量（特征）的权 $w_j, j=1,2,\cdots,n$，以该权作为样本第 j 个特征的重要性度量。

传统 SVM 算法只能解决两分类问题，我们使用的 SVM 工具箱中的 libsvm 使用的是一对一(one-against-one)方式来解决多类问题。因此，对于多类问题 libsvm 将要进行多个两类分类，对应的将会产生多个两类分类问题的分类器，w 也就有多组，用 $w^k=\left[w_1^k, w_2^k, \cdots, w_n^k\right], k=1,2,\cdots, l(l-1)/2$ 来表示，其中 l 表示数据集的类别数，此时一对一方法分类器的个数为 $l(l-1)/2$，n 表示特征个数。则我们采用式(5-1)来计算各个特征的权值 w，以此权值作为相应特征的重要性评价准则。

$$w=\left[\sum_{i=1}^{l(l-1)/2}\left|w_1^i\right|, \sum_{i=1}^{l(l-1)/2}\left|w_2^i\right|, \cdots, \sum_{i=1}^{l(l-1)/2}\left|w_n^i\right|\right]=\left[w_1, w_2, \cdots, w_n\right] \tag{5-1}$$

5.2 基于 SVM 分类模型的特征选择算法

基于 SVM 分类模型的特征选择算法以 Guyon 等人[33]的 SVM-RFE（SVM Recursive Feature Elimination, SVM-RFE）方法最为有名。该方法在每一次递归迭代时剔除掉排序在最后的那个特征。特征排序的原则是：训练 SVM 学习机，得到当前的最优分类超平面，计算权向量 $w=\sum_k \alpha_k y_k x_k$，则第 i 个特征的重要性为 $c_i=(w_i)^2$。Guyon 发现：在相同的学习机下，分类器的性能取决于所采用的特征选择方法；基于 SVM 特征选择算法的分类器的分类性能优于使用别的特征选择算法的分类器。 但是，Guyou 的 SVM-RFE 方法主要针对两类分类问题。本文将 Guyon 的 SVM-RFE 进行推广，提出适用于任意类分类问题的 SVM RFE（SVM Recursive Feature Elimination, SVM RFE）特征选择算法；并受此启发，相应地提出 SVM RFA（SVM Recursive Feature Addition, SVM RFA）算法；最后通过 UCI 机器学习数据库的数据集对这两个算法的性能进行了测试。下面分别是两个算法的描述。

5.2.1 适用于多类的 SVM RFE 特征选择算法

1st 先用全部特征训练 SVM，根据一对一的方法得到多个两类分类的最优 SVM 分类模型;

2nd 依据 5.1 描述的特征重要性计算方法，根据刚得到的多个最优 SVM 分类器，利用公式（5-1）计算各特征的权重，对特征进行降序排序；

3rd 删除最后一个权重最小的特征；

4th 用剩余的特征训练 SVM，得到多个二分问题的新的最优 SVM 分类模型；

5th 转 2nd ，直到剩余特征为空。

5.2.2 适用于多类的 SVM RFA 特征选择算法

step 1: 初始化被选特征子集为空集，备选特征子集为全集，包含所有特征；

step 2: 使用包含备选特征子集全部特征的训练集样本训练 SVM，根据一对一的方法得到多个两类分类的最优 SVM 分类模型;

step 3: 依据 5.1 描述的特征重要性计算方法，根据 step 2 得到的多个最优 SVM 分类器，利用式（5-1）计算各特征的权重，对特征进行降序排序；

step 4: 加入第一个最重要的特征，即首个权值最大的特征到被选特征子集，并从备选特征子集将该特征删除；

step 5: 转 step 2 ，直到备选特征子集为空。

备注：对于两类分类问题，在上述 SVM RFA 的 step 2 步，只要计算一个最优的 SVM 分类模型，式（5-1）此时依然成立。

5.3 实验结果与分析

为了比较 SVM RFA 和 SVM RFE 的性能，我们以 UCI 机器学习数据库[158] 的 dermatology，glass，handwrite，Ionosphere，WDBC (Wisconsin Diagnostic Breast Cancer，WDBC)，WPBC (Wisconsin Prognostic Breast Cancer, WPBC)，wine，thyroid-disease，和 heart disease 九个数据集分别进行实验。实验所用数据集描述见表 5.1。其中，dermatology 数据集，删去了 8 个含有缺失数据的样本，因此总的数据集规模由原来的 366 变成了 358；glass 数据集分成了 window glass 和 non-window glass 两类；handwrite，即 Semeion Handwritten Digit 数据集，只选择了前两类进行实验；WPBC 数据集删去了 4 个含有缺失数据的样本，因此数据集规模为 194；关于 thyroid-disease 数据集，实验使用了其中的 new-thyroid，也即 Thyroid gland data 数据集；Heart Disease 数据集使用的是其中的 processed cleveland 数据集，实验中删去了 6 个含有缺失数据的样本，因此数据集规模为 297。

表 5.1 UCI 数据集描述
Table 5.1 the description of data sets from UCI machine learning repository

数据集	样本个数	特征数	类别数
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

为了使实验结果具有统计意义，我们将样本顺序随机打乱，对每一类的样本依次逐个加入到五个初始为空的样本集合中，直到这一类的样本被处理完毕，得到样本均匀划分的五折交叉验证实验的数据集划分。以每一份作测试样本，其余四份作训练样本，以每一份都做过测试集后结束，实现五折交叉验证实验。样本随机打乱的方法同前面章节：生成一个 5000 行 2 列的 2 维数组，数组的每一个元素的值在 1～数据集规模；交换 2 维数组每一行的两个元素值对应的两个样本，以此方法实现样本随机打乱。实验采用的 SVM 工具箱为台湾林智仁教授等开发的 LibSvm 工具箱[160]的线性核函数，核函数参数采用默认参数。

对 WDBC，WPBC 和 wine 三个数据集，我们在原始数据集和标准化后的数据集上分别进行了实验。标准化的方法我们采用了最大最小标准化方法[163]。

表 5.1~表 5.12 分别给出了 SVM RFA 和 SVM RFA 在各数据集上进行 5 折交叉验证实验分别得到的特征排序结果。图 5.1~图 5.24 分别展示了 SVM RFA 和 SVM RFE 的 5 折交叉验证实验的训练集正确率和测试集正确率。

表 5.1 SVM RFA 和 SVM RFE 特征选择方法对 Dermatology 数据集的特征排序
表 5.1 Ranked features of Dermatology dataset by SVM RFA and SVM RFE respectively

Dermatlogy		Ranked features
SVM RFA	Fold 1	34, 28, 21, 15, 16, 14, 5, 22, 20, 9, 4, 2, 7, 31, 30, 10, 26, 33, 3, 11, 24, 29, 19, 1, 17, 18, 13, 6, 27, 25, 12, 8, 32, 23
	Fold 2	34, 21, 28, 15, 16, 14, 5, 22, 20, 9, 7, 31, 30, 10, 2, 24, 33, 4, 17, 18, 11, 29, 27, 26, 6, 25, 12, 8, 3, 19, 13, 1, 32, 23
	Fold 3	34, 21, 28, 15, 16, 14, 5, 7, 9, 31, 30, 2, 20, 22, 10, 4, 26, 3, 33, 17, 24, 19, 11, 1, 29, 18, 25, 27, 12, 6, 8, 32, 23, 13
	Fold 4	34, 28, 21, 15, 16, 14, 5, 7, 22, 20, 9, 31, 30, 10, 4, 26, 24, 33, 17, 2, 1, 3, 19, 11, 18, 13, 29, 8, 6, 27, 25, 12, 32, 23
	Fold 5	34, 21, 28, 15, 16, 14, 5, 22, 20, 2, 7, 9, 31, 30, 10, 23, 4, 33, 1, 26, 19, 3, 11, 18, 17, 24, 29, 27, 25, 8, 12, 6, 32, 13
SVM RFE	Fold 1	34, 21, 28, 33, 15, 29, 16, 4, 27, 5, 9, 6, 20, 22, 14, 12, 3, 8, 25, 10, 2, 19, 26, 24, 7, 1, 23, 18, 31, 17, 30, 13, 32, 11
	Fold 2	34, 21, 28, 33, 15, 16, 29, 27, 6, 5, 20, 22, 9, 12, 4, 14, 3, 25, 8, 10, 2, 7, 24, 26, 19, 23, 1, 18, 17, 31, 32, 13, 30, 11
	Fold 3	34, 21, 28, 33, 15, 16, 27, 29, 6, 20, 5, 9, 14, 22, 4, 12, 25, 3, 8, 10, 2,

		7, 26, 23, 19, 24, 1, 31, 17, 30, 18, 32, 13, 11
	Fold 4	34, 28, 21, 33, 15, 16, 27, 29, 6, 5, 20, 9, 22, 4, 12, 14, 3, 8, 19, 25, 2, 10, 26, 24, 7, 1, 18, 23, 32, 31, 17, 30, 13, 11
	Fold 5	34, 21, 28, 33, 16, 15, 27, 29, 6, 9, 5, 22, 20, 4, 14, 12, 25, 3, 8, 2, 10, 19, 26, 7, 23, 17, 24, 1, 18, 31, 32, 30, 13, 11

表 5.2 SVM RFA 和 SVM RFE 两种特征选择方法对 Glass 数据集的特征排序

表 5.2 Ranked features of Glass dataset by SVM RFA and SVM RFE respectively

Glass		**Ranked features**
SVM RFA	Fold 1	7, 2, 6, 4, 3, 8, 5, 1, 9
	Fold 2	7, 6, 4, 2, 5, 3, 8, 1, 9
	Fold 3	7, 2, 4, 6, 3, 8, 5, 1, 9
	Fold 4	7, 2, 4, 6, 3, 8, 9, 5, 1
	Fold 5	7, 2, 6, 4, 3, 5, 8, 1, 9
SVM RFE	Fold 1	7, 2, 5, 6, 4, 3, 9, 8, 1
	Fold 2	7, 4, 6, 2, 5, 3, 9, 8, 1
	Fold 3	7, 3, 6, 2, 4, 5, 8, 9, 1
	Fold 4	7, 2, 4, 3, 6, 5, 8, 9, 1
	Fold 5	7, 3, 6, 4, 2, 5, 8, 9, 1

表 5.3 SVM RFA 和 SVM RFE 特征选择方法对 Handwrite 数据集的特征排序

表 5.3 Ranked features of Handwrite dataset by SVM RFA and SVM RFE respectively

Handwrite		**Ranked features**
SVM RFA	Fold 1	146,130,145,162,193,178,177,161,112,114, 98, 99,129, 83,113, 96, 95,111,128,160,144, 82, 79, 68, 80, 127, 143,159,232, 84,194,184, 97,185,169,168,103, 88,239,152,138, 73,137,153,163, 52, 66, 51, 67, 183, 167,136, 81,121,175,147, 37,131, 36, 63,151,105,166, 57,191,176,154, 89, 74,120,104,231,135, 50, 65,119,106,134, 35, 58, 90,122, 75, 59, 72,182, 53, 91,150, 64,200,210,209,118,203,199, 69, 87, 102, 238, 54,211, 38,195, 23, 71, 22,179, 76, 21,7,8,9, 10, 20, 62,107, 92,123, 93, 11, 108, 139, 170, 230, 124, 16, 24, 25,109, 26, 39,155,246,115, 12, 245,241,229,140,125,244,228,223,198,247, 77, 240, 141, 47, 60,227,202,254,212,255,6,224,233, 27,217,248,156,253,171,237, 15, 61,213,174,190, 32, 13, 100, 158, 85,218,216,204,142,189, 70,173,205,219, 40,188,181,234,165, 28,197, 55,116,252,242,5, 149, 19, 34, 49,225, 44,186,206,117, 43,220,235,249, 31,133,4, 14,256,222, 18,214,215, 30, 164, 196, 157, 148, 45,132,250, 42,208,3,192, 243,180, 33, 48, 56, 86,187, 94, 29, 78, 41,251,2, 17, 236, 126, 226, 221, 46,207,110,172,101,201,1
	Fold 2	112,145,146,161,162,178,130,177,193,111,128,114,113,144, 95,160, 74, 79, 96, 80,127,143, 24,194, 73,159, 58, 89, 82, 72, 88,103,102, 105, 75, 63,183, 57,104,119, 98, 83, 59, 90,129,118,134,120, 97,135,184, 99, 84,136,232, 87,175,151,168,121,152,167, 64,9, 163, 68,150,166,106, 10,137, 76, 91,8, 51, 66, 81, 67, 52, 50, 36, 37, 25,153,138,169,122, 92,107,182, 93, 53,185, 38, 23,191,176, 39,108, 123,139, 22,231,230,124, 16, 65, 35, 21,154,239,7,195,210,179, 69,238,223, 60, 54,170, 11,109, 77,240,190,198,199, 20,229,174,155,6,140, 15, 12,125,5,141, 19,246,245,211,228,212,227,247, 26,244,61,254,32,233,13,213,31,214,156,115,171,255,186,253,200,209,158,237,24 8,217,157,27,225,241,189,205,147,142,206,224,252,218,234,204, 28,216,131, 40,196,34,49,117,4,243,100,165,149,220,181,219,203,56,133,173,71,249,202,116, 29,164,242,235,201,221, 44, 47, 42, 85,188, 43, 86,222,256,215, 14, 18, 70,148, 55, 94,132, 46, 251, 62,192,197, 41,208,101,3, 33,2,172,250,207, 48,187, 30,110, 45,126, 17,180,226, 78,1,236
	Fold 3	146,145,161,130,162,177,178,112,193, 95,114, 83,111, 98, 99,113,194, 68, 82,129,128,159,143,127,144,160, 97, 79, 52, 51, 66, 67, 37, 36, 96,175, 81,184, 73, 16, 50,185,169,168,138,153,137,121,152,154,183, 35,163,147,167,136, 93, 80,105,122,109,151,191,106,238,239, 57, 74, 88, 58, 89,195, 65, 38, 23, 22,176, 21, 53,179, 63, 9, 90, 84,8,7,120,104, 72,103, 59, 75,210,223, 20,6,139, 64, 24, 91,107,123,170, 10,108, 76, 92,124, 54, 69, 39,231,209,232,166,182, 25, 77,240, 60, 11,211,230,229,247,246, 12,245,155,140,125,228, 244,227,212,255,141, 61,

		15,190,248,174,233,156,158,13,171,199,135,198,119,134,87,47,237,254,253,217,115,118,241,200,242,224,186,216,26,32,204,218,189,205,5,234,142,173,102,150,219,203,28,220,27,19,225,131,165,181,188,202,149,100,243,222,56,34,213, 4,31,235, 29, 49,249,256,214,206,196,157,252,133, 43, 44, 42,201, 71,101, 40,117,70,86,14,215,55,85,132,250,18,94,172,192,110,221,208, 45,126,197,226, 78, 48,164,148, 30,187,207,3, 33,251,236,116,2, 41, 46, 62,180,1, 17
	Fold 4	146,162,130,178,177,161,145,129,113,98,112,114,193,99,83,111,143,127,128,144, 160, 95, 68, 96,159, 97, 82, 84,194, 80, 16,121, 66, 51, 67, 36, 52, 37, 81, 79,136,232,152,168,184,185,137,153,169,138, 73,175, 50, 35, 88, 89,104,183,167, 64,105,103, 57, 74, 58,163, 147,179, 21,195, 65, 72,231,154,191, 63, 20, 22, 59, 23,122, 75,8,176,239,139,106,238, 53, 38,223,210, 90,7, 93,120,151, 69, 24, 123,107, 91,170,209, 76,108,9, 10, 92,240,124,109,6, 11, 25, 39, 54,166, 60,230,211,182,233,247,217,246,245,125, 77, 12,140,155,229,244, 13,228,227, 15, 61,141, 32,255,248,224,216,254,115, 26,212,253,237,174,171,252,156,190,135,119,118,102,186,234,213,158,214, 31,5,87,134,142,27,192,206,218,131,150,226,165,243,189,204,205,173,203,199,200,219,198,241,100,202, 70, 85,116, 55, 40,196,242, 44,149,220,235, 45,188, 43, 47,222,256,133, 48,117, 19, 34, 49,4,181,249, 28,225,164,148, 56, 42,201,221, 71, 86,101,157,197, 29, 94, 78,250, 62, 18,208, 14,3,215, 30,2, 33,180,251, 17,236, 46,126,132,110,207,172, 41,187,1
	Fold 5	146,130,112,193,145,129,177,178,162,143,161,232,114, 98,113,111,127,233, 82,128,159,144,160, 95, 99, 79, 88,103,175, 96, 97, 83, 16, 66, 52, 67, 81, 51, 36, 37, 22, 73, 24,185,184, 89,104, 74, 57, 50, 23, 80,168,136,163,105, 63,121,203, 68, 64, 10,152,137,153,138, 169,154,191,8,7,9,194, 21, 35, 58,170, 84,176, 25, 65, 11,183,151,120, 72,167, 87,210, 59, 90, 75,119,231,230, 20,106,122,134,6,166,118,135,223,179,239, 53, 38, 93, 76, 91,107,238,182,123,139,199,200, 92,108,109,124,195,147, 69, 12,198, 15,247,174,102,190, 205,189,204,229, 60, 77,246,228,211,150,209,245,248,155,125, 32,241,244, 13,140,227,212,141, 39, 54,158,186,156,171,173,142, 27, 26,5, 31, 61,115,240,242,225,224, 19,201,213,254,117,220, 44,188,218,149,206,133,234,219,165,217,181,131,197, 28,100,216,253,237,252, 34, 49,4,255, 47, 43,249,222,235, 70, 85,202, 14, 40,256, 56, 86, 42,101,132, 71,157, 18,148,214,196,164, 30,192, 62,3,250, 33, 208, 215,172, 55, 45, 94, 78,110,243,116,226, 48, 29,2,187,221,207, 41,126,180, 17, 46,236,251,1
SVM RFE	Fold 1	112,128,130, 68,162,160,193,238, 50, 80,230, 59, 98, 12,146, 79, 82, 99,144, 10,153,232,145, 95,194, 96,196,129,163, 97,111,178,246, 8, 89,136,239, 25,176,177,169,210,114,152, 58,247, 24, 66,254,138, 27,103, 52,119, 53,143, 11, 65,113,203,105,147,166,211,184, 88, 81,255, 16,127,223, 84,9, 159, 216,185,151, 74,131, 26,161,195, 83,154,187,124,175,253, 222,7, 63,4,168,231, 73,121,233,109,179, 104, 23,229,202,256, 15, 64, 19,204, 18,237,183,139,240, 72,245,137,102, 37, 70,3,209, 67,123, 60,6,167, 22, 90,141,110, 69, 93, 57, 36,182, 54,172,157, 28,241, 35, 94,125,213,224,242, 39,5,199,120,252,228,217,117,150, 38,164,106, 62, 49, 47,101,140, 32, 14,107,165, 51,108,191,205,115,180, 76,244,200, 75, 85, 71,132,170,171,225,212,122, 20,116, 43, 91, 92, 13,135,192,133,134, 77,198, 87, 86,100,189, 61,118,221,218, 40,249,234, 34, 17, 31,2,208,126,226,243, 45,190,155,235, 30, 29,251,197,227, 21,215,248, 56, 44,186, 206, 48,149,201,148,207,173, 55,250,219,214,158,142,181, 33,220, 46,236, 41,174,156, 42,188, 78,1
	Fold 2	112,111,113,128,146, 10, 89,108, 98, 80,160,193,130,144, 74,238,139, 11,145,153, 96,136, 58,230,183,239,162, 25, 12, 81, 79,232,194,105,9, 73,124,114, 82,119, 27, 53, 59,102,138,178,129, 24,121, 65,254,103,246,168, 16,199, 93, 63, 97,143,161,8, 99,184, 91,196, 95, 39, 19,4,229,163, 50,169,151,240,177, 57, 84,120,195,127,109,212,241, 66,134,152, 76,210, 68, 75,135,223,176,154, 92,202, 72,179,106, 49, 90,198, 26,150,159, 88,155,187,5,166, 69, 52,209,247, 83,233,123,137, 47, 28,182,245, 64,125, 54, 94,147,104,185,107, 234, 253, 60, 37, 38,231,118,175,200,167, 18,216,3, 51,255,180, 77, 40,206, 36,140, 23, 15,172,224,237, 71, 35,131,122,256,225,211, 67, 32,236,141, 42,222, 34, 13,149,242, 29, 20,235,219,170,203,101,189,248,220,213,217,208, 78, 86, 30, 46,165,7,6,226,191,214, 197, 228, 14,207,244, 17,2,133,251,190,205,

		87,243,249,116,117, 45,115,126,252, 48,204,157,174, 56,192, 70, 85,215,132,171,188, 22, 61,110,164, 41, 43,142, 21,227,158, 31,250,221, 33, 44,173,100,201, 55,218, 62,148,156,181,186,1
	Fold 3	128, 60,146,113, 96, 82,95,246,177, 50, 68,239,130,160,111,10,145, 66, 53,232,178,210,112,193,98,24,144,89,153,136,230, 97,195, 79,143,238,162,129, 12, 59,247,152,194,147,161,223,159, 83, 23, 52, 11,114,184,8, 121,151, 58,124,245,36, 99, 35,163,211,105,168,9, 73,203,109, 67, 81,127,7,196,102,137, 37,175,166, 16,139,131, 80,254,103, 38,154, 84, 74,185,229,233, 51, 19, 54, 94, 27,119, 237, 65,138,241, 88,6,4,212,169,256, 22,104,199, 93,209, 39, 15, 57,167,255,101,150, 43, 69,183,231,240,182,217,176, 64,179,132, 20, 72, 90, 70,86,253,207,77,123,25,216,200,228,248,28,122,242,157,244,76,172,34,141,187,1 65,164,106,198,197,135,170,75,191,92,49,204,30,63,71,120,18,252,3,222,202,224, 140,134,107,91,208,31,13,108,5,125,220,206,243,148,46,78,227,21,213,47,155,171 ,26,186,180,133,190,117,115,32, 48,181,251,188, 14,250,218,174,142, 44,55,214,225,173,61,215,126,189,192,29,40,87,158,234,62,219,118,116,149,201,2 21,56,45,235,85,156,33,41,100,110,249,236,226,205, 42, 17,2,1
	Fold 4	112, 68,129,162,209,160, 16,111, 80, 97,239,146,128, 11,178, 79,130,232,143,238,113,153, 96,193, 67,8, 136,139, 98, 50,138,163,196, 203,144, 69, 10, 53, 66,127,240,145,246,177,105, 82, 84, 95, 233, 104, 210, 247,152, 59,179, 24, 36, 64,114,147,195, 81,222,137, 58,154, 26,161,194, 89,131,253, 12, 99, 74, 121, 65,120,211,9,241,172,109, 52, 35, 83,159, 93,169, 88, 39, 94,254, 37,185,167, 25, 63, 73, 27,223, 51, 23,151,217,245,200,115,176, 19,224,119,103,132, 15,4,168,231, 47,6,170,237,110,184,7, 166, 216, 20,183, 38,181,125, 57,255, 60, 13, 32, 54,124,256,5,202,199, 49,212, 91,187,182,102, 22,108, 75, 242,123,175,197,116,206,218,220,189, 72,191, 62, 90, 70, 106, 204, 28, 244, 140, 174, 205, 208, 150, 164, 77, 71,100,122,141,173,180,229, 85,101, 31,190, 78,192,165,133,198, 34, 18,3, 21, 76, 158, 248, 92,135,107,230,157,243,118,228,188,252,234, 17,2,207,142,155,117,148, 30, 225, 215, 250, 236, 214, 226, 126, 55,149, 87,249, 33, 186,213, 46,235, 86,227, 48,134, 41,221, 44, 14, 42, 61,156, 56, 43, 219, 45, 29,171,201,251, 40,1
	Fold 5	112,129,178,128,138,232,160, 80, 68,145, 10,113, 95,194,103, 97,247, 59, 98, 27, 24,146, 50, 79, 82, 238, 89,130,162, 96, 11, 52,193, 144, 58,136,246,143,210,180, 16,114,163, 66, 88, 139, 230, 105, 203, 147, 111,8,179, 12,177, 99,152,211, 81,195,153,199,239, 19,216,159,4, 64, 7,222, 26, 74,161, 65, 69, 127, 176,223,151,245,196,9,121,124,185,233, 49,154,237,119, 37, 84,175,167,241,254, 73,137, 35, 53, 117,131,104, 25, 23,229,209,6,166,200,123,172,253,109,231, 94,184,183, 83, 39,5, 32,256, 60, 57, 168, 70,170,110,102,242,107, 36, 15,187, 67,182, 54,141,108,120,101,248,202, 20, 125, 169, 255, 204, 18, 63,3, 93,224,191,206,165,221, 22, 28,156,240, 90, 38, 86,150, 85, 72,192,228,140, 51, 34, 218, 155,188,164, 45,186,250,118,174,157,135,217, 76,198,189, 87,106, 14, 13,201, 21, 122, 226, 244, 158, 207, 92,134, 47,234, 77,225, 75,214, 30, 71,212,115,215, 91,149,220, 56,208, 17,2, 78,251, 61, 29, 249, 100, 46, 40,142,126,173,205, 235,227, 31,213,243,171, 44,236,148, 42,132,116,197,181,133, 48,219,252, 41, 43,190, 55, 62, 33,1

表 5.4 SVM RFA 和 SVM RFE 特征选择方法对 Inosphere 数据集的特征排序

表 5.4 Ranked features of Inosphere dataset by SVM RFA and SVM RFE respectively

Ionosphere		**Ranked features**
SVM RFA	Fold 1	3,8, 5, 22, 27, 1, 17, 7, 6, 16, 14, 12, 10, 19, 18, 4, 34, 20, 21, 25, 13, 33, 9, 26, 15, 11, 23, 28, 24, 31, 29, 32, 30, 2
	Fold 2	5, 3, 8, 7, 27, 1, 6, 17, 22, 14, 19, 4, 12, 10, 16, 33, 15, 18, 21, 23, 11, 25, 13, 20, 30, 26, 9, 24, 32, 28, 29, 31, 34, 2
	Fold 3	3, 5, 8, 22, 27, 1, 7, 17, 4, 10, 19, 12, 14, 16, 6, 33, 13, 26, 15, 11, 18, 23, 28, 25, 24, 32, 9, 34, 30, 20, 31, 29, 21, 2
	Fold 4	8, 5, 7, 3, 27, 1, 17, 4, 16, 14, 10, 12, 19, 21, 6, 22, 20, 18, 15, 11, 25, 13, 23, 9, 33, 28, 24, 31, 34, 29, 26, 32, 30, 2
	Fold 5	3, 5, 7, 8, 27, 1, 6, 14, 10, 12, 17, 19, 4, 16, 22, 18, 20, 15, 29, 23, 13, 26, 9, 21, 34, 24, 28, 30, 33, 31, 25, 11, 32, 2
SVM RFE	Fold 1	3, 27, 22, 7, 8, 5, 1, 10, 6, 13, 4, 18, 14, 16, 11, 17, 33, 31, 30, 34, 29, 12,

		24, 19, 15, 23, 9, 26, 20, 21, 25, 32, 28, 2
	Fold 2	5, 3, 27, 22, 8, 1, 10, 7, 6, 14, 4, 12, 11, 13, 17, 19, 16, 33, 15, 21, 18, 32, 30, 26, 34, 31, 20, 28, 29, 25, 23, 9, 24, 2
	Fold 3	3, 1, 8, 22, 5, 7, 27, 6, 4, 10, 31, 26, 34, 28, 24, 12, 16, 11, 13, 17, 14, 18, 32, 33, 19, 15, 25, 9, 20, 23, 30, 29, 21, 2
	Fold 4	3, 27, 8, 1, 5, 22, 7, 11, 6, 13, 4, 10, 15, 25, 17, 14, 16, 33, 31, 12, 19, 18, 34, 9, 21, 23, 29, 24, 28, 32, 30, 20, 26, 2
	Fold 5	5, 1, 3, 22, 6, 10, 8, 7, 27, 14, 34, 25, 18, 11, 17, 4, 33, 13, 15, 16, 12, 19, 21, 24, 31, 23, 20, 9, 26, 32, 30, 29, 28, 2

表 5.5 SVM RFA 和 SVM RFE 特征选择方法对 WDBC 数据集的特征排序

表 5.5 Ranked features of WDBC dataset by SVM RFA and SVM RFE respectively

WDBC		Ranked features
SVM RFA	Fold 1	24, 4, 23, 14, 3, 21, 1, 13, 22, 2, 11, 27, 26, 7, 28, 8, 6, 29, 25, 30, 9, 5, 17, 16, 18, 20, 12, 19, 10, 15
	Fold 2	24, 4, 23, 14, 3, 21, 1, 13, 22, 2, 11, 27, 26, 7, 28, 8, 6, 29, 25, 9, 5, 30, 17, 16, 18, 20, 15, 12, 19, 10
	Fold 3	24, 4, 23, 14, 3, 21, 1, 13, 22, 2, 11, 27, 26, 7, 28, 8, 6, 29, 25, 9, 17, 16, 30, 18, 5, 20, 12, 19, 15, 10
	Fold 4	24, 4, 14, 23, 3, 21, 1, 13, 22, 11, 2, 27, 26, 7, 28, 8, 6, 29, 25, 9, 30, 5, 17, 16, 18, 12, 19, 10, 15, 20
	Fold 5	24, 4, 23, 14, 3, 21, 1, 22, 13, 2, 11, 27, 26, 7, 28, 8, 6, 29, 25, 30, 17, 16, 9, 18, 5, 20, 15, 12, 19, 10
SVM RFE	Fold 1	24, 4, 23, 14, 3, 21, 22, 1, 2, 13, 11, 27, 26, 7, 28, 6, 8, 29, 12, 25, 9, 17, 30, 5, 16, 18, 10, 19, 15, 20
	Fold 2	24, 4, 23, 14, 3, 21, 1, 22, 2, 13, 11, 27, 26, 7, 28, 6, 8, 29, 12, 9, 25, 17, 30, 5, 16, 18, 15, 20, 10, 19
	Fold 3	24, 4, 23, 14, 3, 21, 22, 1, 2, 13, 11, 27, 26, 7, 28, 6, 8, 29, 25, 9, 17, 30, 16, 5, 12, 18, 20, 15, 19, 10
	Fold 4	24, 4, 23, 14, 3, 21, 22, 1, 2, 13, 11, 27, 26, 7, 28, 6, 8, 12, 29, 25, 9, 17, 30, 16, 5, 18, 20, 15, 19, 10
	Fold 5	24, 4, 23, 14, 3, 21, 22, 1, 2, 13, 11, 27, 26, 7, 28, 6, 8, 29, 25, 9, 17, 30, 16, 5, 18, 12, 19, 15, 20, 10

表 5.6 SVM RFA 和 SVM RFE 对标准化的 WDBC 数据集的特征排序

表 5.6 Ranked features of normalized WDBC dataset by SVM RFA and SVM RFE respectively

Normalized WDBC		Ranked features
SVM RFA	Fold 1	28, 8, 21, 23, 3, 1, 4, 24, 7, 27, 26, 6, 11, 13, 14, 22, 2, 18, 29, 25, 5, 30, 9, 16, 17, 19, 15, 12, 20, 10
	Fold 2	28, 8, 21, 23, 3, 1, 4, 24, 7, 27, 6, 26, 11, 13, 14, 22, 2, 25, 29, 18, 30, 5, 16, 9, 17, 19, 15, 12, 10, 20
	Fold 3	28, 8, 21, 23, 3, 1, 4, 24, 7, 27, 26, 6, 11, 13, 14, 22, 2, 18, 25, 29, 30, 16, 5, 9, 17, 19, 10, 15, 12, 20
	Fold 4	28, 8, 21, 23, 3, 1, 4, 24, 7, 27, 11, 6, 26, 13, 14, 22, 2, 18, 25, 29, 30, 5, 9, 16, 17, 19, 20, 10, 15, 12
	Fold 5	28, 8, 21, 23, 3, 1, 4, 24, 7, 27, 26, 6, 11, 13, 14, 22, 2, 25, 29, 30, 5, 18, 9, 16, 17, 19, 12, 10, 15, 20
SVM RFE	Fold 1	21, 28, 23, 8, 22, 3, 25, 1, 24, 29, 7, 4, 2, 27, 11, 26, 13, 14, 6, 10, 9, 5, 20, 18, 30, 17, 15, 19, 12, 16
	Fold 2	21, 28, 8, 23, 24, 25, 22, 3, 1, 7, 29, 4, 27, 11, 2, 26, 13, 14, 9, 6, 10, 5, 20, 30, 16, 18, 12, 17, 15, 19
	Fold 3	21, 28, 23, 22, 8, 24, 3, 25, 1, 29, 7, 4, 2, 27, 11, 26, 13, 10, 9, 14, 6, 20, 30, 5, 17, 12, 18, 19, 15, 16
	Fold 4	21, 28, 8, 23, 22, 3, 25, 1, 7, 24, 29, 4, 27, 2, 11, 26, 13, 9, 14, 6, 10, 5, 30, 20, 18, 19, 17, 16, 15, 12
	Fold 5	21, 28, 23, 8, 22, 1, 25, 3, 29, 24, 7, 4, 27, 2, 26, 11, 9, 10, 13, 14, 6, 20, 5, 30, 17, 16, 12, 19, 18, 15

表 5.7 SVM RFA 和 SVM RFE 特征选择方法对 WPBC 数据集的特征排序

表 5.7 Ranked features of WPBC dataset by SVM RFA and SVM RFE respectively

WPBC		Ranked features
SVM RFA	Fold 1	25, 5, 1, 15, 24, 4, 22, 2, 33, 32, 23, 13, 3, 7, 10, 8, 18, 28, 14, 11, 12, 19, 16, 6, 27, 20, 21, 9, 29, 30, 17, 26, 31
	Fold 2	25, 5, 1, 24, 15, 4, 22, 33, 3, 32, 23, 2, 13, 27, 18, 14, 19, 9, 12, 11, 16, 31, 17, 10, 21, 28, 30, 20, 26, 8, 6, 7, 29
	Fold 3	25, 5, 1, 24, 15, 4, 33, 22, 2, 32, 23, 14, 13, 18, 12, 3, 9, 19, 11, 16, 6, 20, 28, 30, 17, 21, 27, 7, 31, 8, 10, 26, 29
	Fold 4	25, 5, 1, 15, 24, 4, 22, 2, 33, 3, 13, 32, 23, 14, 17, 20, 18, 12, 9, 19, 6, 16, 11, 30, 26, 27, 8, 21, 28, 7, 29, 10, 31
	Fold 5	25, 5, 1, 24, 15, 4, 22, 33, 2, 13, 3, 27, 23, 32, 18, 8, 10, 19, 14, 11, 20, 16, 12, 28, 21, 9, 6, 7, 29, 17, 31, 30, 26
SVM RFE	Fold 1	25, 5, 1, 15, 24, 4, 22, 2, 33, 14, 23, 32, 3, 12, 13, 27, 30, 8, 9, 10, 31, 29, 28, 7, 11, 26, 18, 6, 19, 16, 21, 20, 17
	Fold 2	25, 5, 1, 24, 15, 4, 33, 22, 23, 3, 2, 32, 14, 13, 12, 28, 30, 9, 8, 10, 27, 29, 18, 26, 31, 11, 19, 20, 6, 17, 16, 21, 7
	Fold 3	25, 5, 1, 24, 15, 4, 33, 22, 23, 32, 2, 3, 14, 13, 12, 28, 29, 9, 10, 26, 8, 18, 30, 7, 11, 19, 31, 17, 20, 27, 21, 16, 6
	Fold 4	25, 5, 1, 15, 24, 4, 22, 33, 2, 32, 14, 3, 12, 27, 13, 23, 30, 9, 10, 29, 8, 7, 28, 31, 18, 11, 17, 26, 6, 20, 19, 16, 21
	Fold 5	25, 5, 1, 24, 15, 4, 33, 22, 2, 3, 32, 23, 14, 13, 12, 29, 28, 27, 30, 9, 18, 26, 10, 7, 8, 20, 6, 19, 11, 31, 16, 17, 21

表 5.8 SVM RFA 和 SVM RFE 对标准化的 WPBC 数据集的特征排序

表 5.8 Ranked features of normalized WPBC dataset by SVM RFA and SVM RFE respectively

Normalized WPBC		Ranked features
SVM RFA	Fold 1	1, 19, 18, 11, 3, 22, 25, 24, 5, 16, 8, 4, 10, 2, 30, 15, 14, 21, 31, 6, 13, 32, 17, 23, 7, 9, 12, 26, 27, 28, 33, 20, 29
	Fold 2	1, 11, 17, 21, 5, 4, 9, 14, 3, 16, 24, 10, 26, 18, 15, 8, 22, 27, 25, 19, 31, 2, 6, 33, 29, 23, 13, 28, 20, 30, 32, 12, 7
	Fold 3	1, 11, 18, 25, 24, 22, 19, 16, 5, 15, 9, 31, 4, 2, 14, 28, 21, 10, 12, 29, 3, 17, 6, 26, 8, 20, 7, 33, 32, 30, 27, 23, 13
	Fold 4	1, 10, 11, 18, 30, 22, 31, 17, 25, 24, 14, 5, 4, 8, 16, 6, 21, 13, 7, 3, 20, 19, 28, 33, 9, 15, 26, 29, 32, 2, 12, 27, 23
	Fold 5	1, 19, 11, 17, 25, 22, 24, 28, 26, 27, 9, 15, 30, 3, 14, 5, 12, 4, 8, 20, 29, 18, 16, 33, 6, 23, 7, 10, 2, 13, 21, 32, 31
SVM RFE	Fold 1	22, 1, 10, 3, 33, 19, 6, 13, 25, 18, 11, 26, 24, 7, 9, 20, 21, 15, 27, 31, 23, 30, 2, 32, 29, 14, 8, 5, 12, 16, 28, 17, 4
	Fold 2	22, 13, 8, 6, 1, 33, 17, 11, 19, 18, 21, 20, 10, 3, 24, 26, 14, 31, 32, 25, 15, 12, 30, 2, 29, 16, 27, 7, 4, 28, 5, 9, 23
	Fold 3	25, 1, 11, 32, 26, 3, 33, 13, 10, 6, 22, 18, 24, 9, 2, 27, 29, 20, 19, 17, 4, 7, 15, 31, 21, 30, 12, 23, 8, 28, 14, 5, 16
	Fold 4	13, 6, 1, 10, 22, 3, 25, 18, 26, 20, 33, 11, 24, 23, 19, 21, 8, 9, 2, 14, 17, 4, 12, 28, 31, 27, 15, 29, 16, 7, 30, 32, 5
	Fold 5	14, 19, 13, 1, 3, 25, 26, 18, 32, 6, 7, 24, 22, 2, 23, 21, 20, 10, 16, 11, 4, 17, 30, 29, 33, 8, 15, 27, 12, 5, 28, 31, 9

表 5.9 SVM RFA 和 SVM RFE 对 Wine 数据集的特征排序

表 5.9 Ranked features of Wine dataset by SVM RFA and SVM RFE respectively

Wine		Ranked features
SVM RFA	Fold 1	13, 5, 10, 4, 7, 12, 1, 6, 11, 2, 3, 9, 8
	Fold 2	13, 5, 4, 10, 7, 12, 1, 6, 11, 2, 9, 8, 3
	Fold 3	13, 5, 4, 10, 7, 12, 1, 6, 11, 2, 3, 9, 8
	Fold 4	13, 5, 10, 4, 7, 12, 1, 6, 11, 2, 3, 9, 8
	Fold 5	13, 5, 10, 4, 7, 12, 1, 6, 9, 2, 11, 8, 3
SVM RFE	Fold 1	13, 5, 4, 10, 7, 1, 12, 2, 6, 9, 11, 3, 8
	Fold 2	13, 5, 10, 4, 7, 1, 12, 2, 6, 9, 11, 3, 8

	Fold 3	13, 5, 4, 10, 7, 1, 12, 2, 6, 9, 11, 3, 8
	Fold 4	13, 5, 10, 4, 7, 1, 12, 2, 6, 9, 11, 3, 8
	Fold 5	13, 5, 4, 10, 7, 1, 2, 12, 6, 9, 11, 3, 8

表 5.10 SVM RFA 和 SVM RFE 对标准化的 Wine 数据集的特征排序

表 5.10 Ranked features of normalized Wine dataset by SVM RFA and SVM RFE respectively

Normalized Wine		Ranked features
SVM RFA	Fold 1	12, 7, 13, 1, 10, 6, 11, 9, 2, 8, 4, 5, 3
	Fold 2	13, 1, 10, 7, 12, 6, 11, 4, 8, 2, 9, 5, 3
	Fold 3	12, 10, 1, 13, 7, 6, 11, 2, 4, 8, 9, 5, 3
	Fold 4	12, 7, 10, 1, 13, 6, 11, 8, 4, 9, 2, 5, 3
	Fold 5	13, 12, 7, 10, 1, 6, 11, 2, 9, 8, 4, 5, 3
SVM RFE	Fold 1	7, 1, 12, 13, 10, 11, 6, 2, 4, 3, 8, 9, 5
	Fold 2	7, 12, 13, 10, 1, 11, 4, 6, 2, 3, 8, 9, 5
	Fold 3	7, 1, 12, 13, 10, 11, 4, 6, 2, 3, 8, 9, 5
	Fold 4	7, 12, 13, 10, 1, 11, 4, 6, 2, 3, 8, 9, 5
	Fold 5	7, 12, 13, 10, 1, 11, 6, 4, 2, 3, 9, 8, 5

表 5.11 SVM RFA 和 SVM RFE 对 thyroid disease 数据集的特征排序

表 5.11 Ranked features of thyroid disease dataset by SVM RFA and SVM RFE respectively

Thyroid disease		Ranked features
SVM RFA	Fold 1	1, 5, 4, 2, 3
	Fold 2	1, 5, 2, 4, 3
	Fold 3	1, 5, 2, 4, 3
	Fold 4	1, 5, 4, 2, 3
	Fold 5	1, 5, 4, 2, 3
SVM RFE	Fold 1	1, 5, 2, 4, 3
	Fold 2	1, 5, 2, 4, 3
	Fold 3	1, 5, 2, 4, 3
	Fold 4	1, 5, 2, 4, 3
	Fold 5	1, 5, 2, 4, 3

表 5.12 SVM RFA 和 SVM RFE 对 Heart disease 数据集的特征排序

表 5.12 Ranked features of Heart disease dataset by SVM RFA and SVM RFE respectively

Heart disease		Ranked features
SVM RFA	Fold 1	12, 10, 6, 13, 8, 11, 7, 9, 1, 3, 5, 4, 2
	Fold 2	13, 12, 3, 10, 6, 7, 8, 11, 2, 9, 4, 5, 1
	Fold 3	12, 13, 10, 3, 8, 6, 7, 9, 11, 1, 5, 4, 2
	Fold 4	12, 3, 10, 13, 8, 6, 9, 7, 11, 1, 5, 4, 2
	Fold 5	12, 3, 10, 13, 6, 8, 9, 7, 2, 4, 5, 11, 1
SVM RFE	Fold 1	10, 12, 13, 6, 3, 9, 8, 11, 7, 1, 4, 2, 5
	Fold 2	10, 12, 13, 7, 11, 3, 9, 6, 8, 2, 4, 5, 1
	Fold 3	12, 10, 9, 13, 11, 3, 7, 8, 6, 2, 1, 4, 5
	Fold 4	12, 10, 8, 13, 3, 6, 9, 11, 7, 2, 1, 4, 5
	Fold 5	12, 10, 3, 13, 9, 6, 11, 7, 8, 4, 2, 5, 1

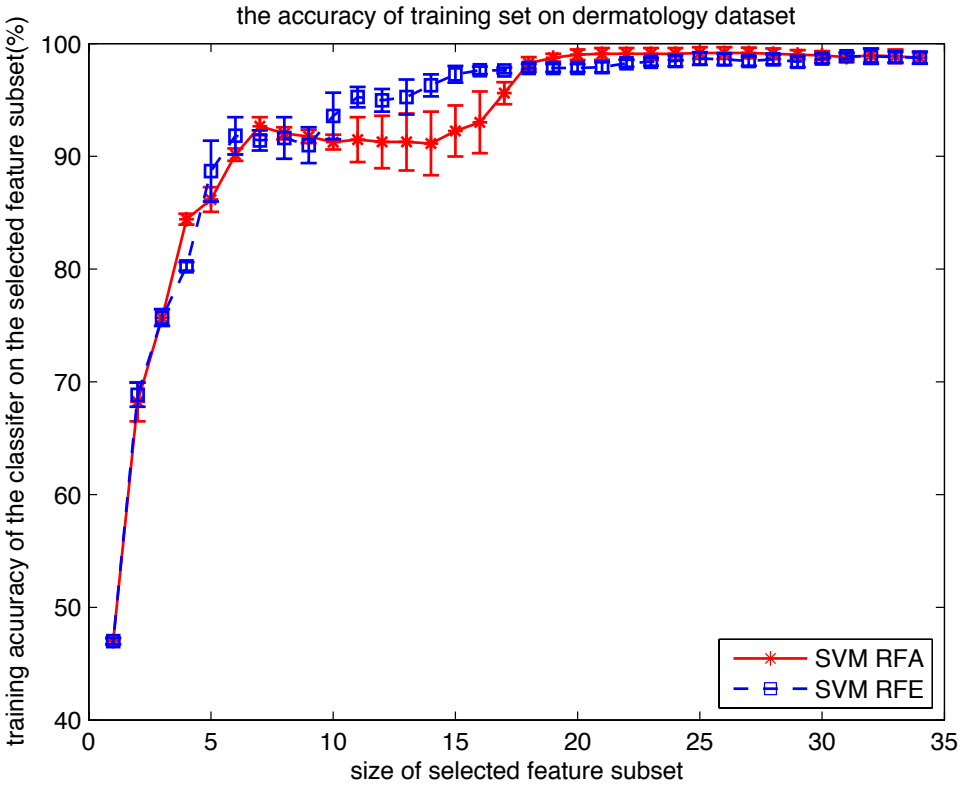


图 5.1　Dermatology 数据集上的训练正确率

Figure 5.1　The training accuracy on Dermatology dataset

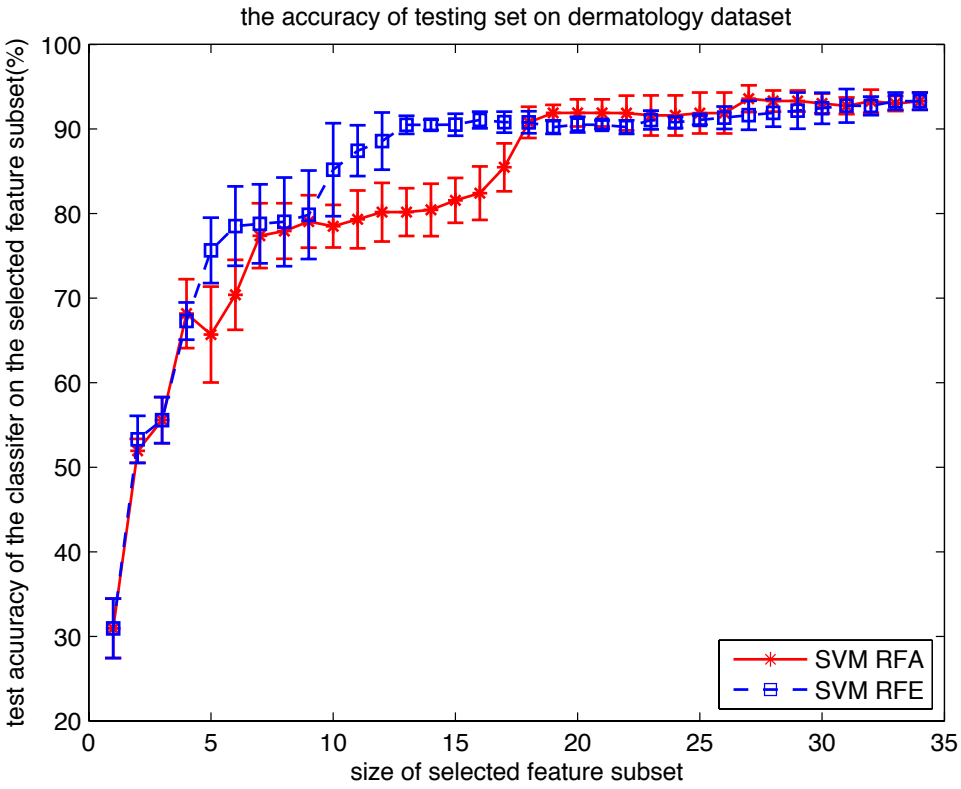


图 5.2　Dermatology 数据集上的测试正确率

Figure 5.2　The testing accuracy on Dermatology dataset

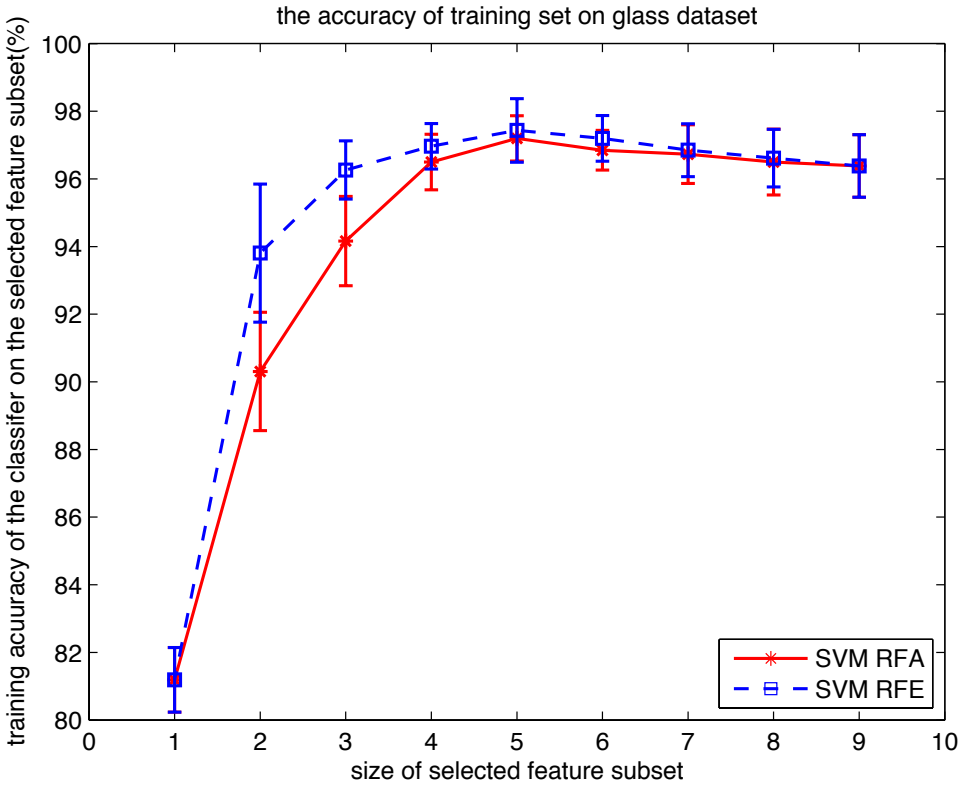


图 5.3　Glass 数据集上的训练正确率

Figure 5.3　The training accuracy on Glass dataset

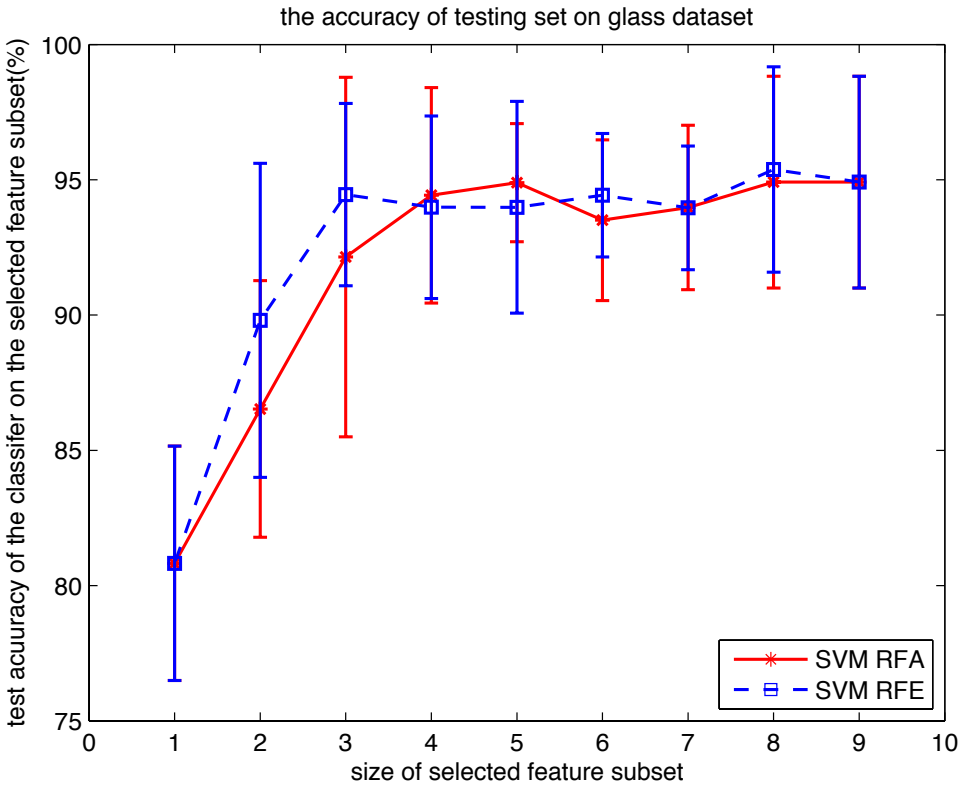


图 5.4　Glass 数据集上的测试正确率

Figure 5.4 The testing accuracy on Glass dataset

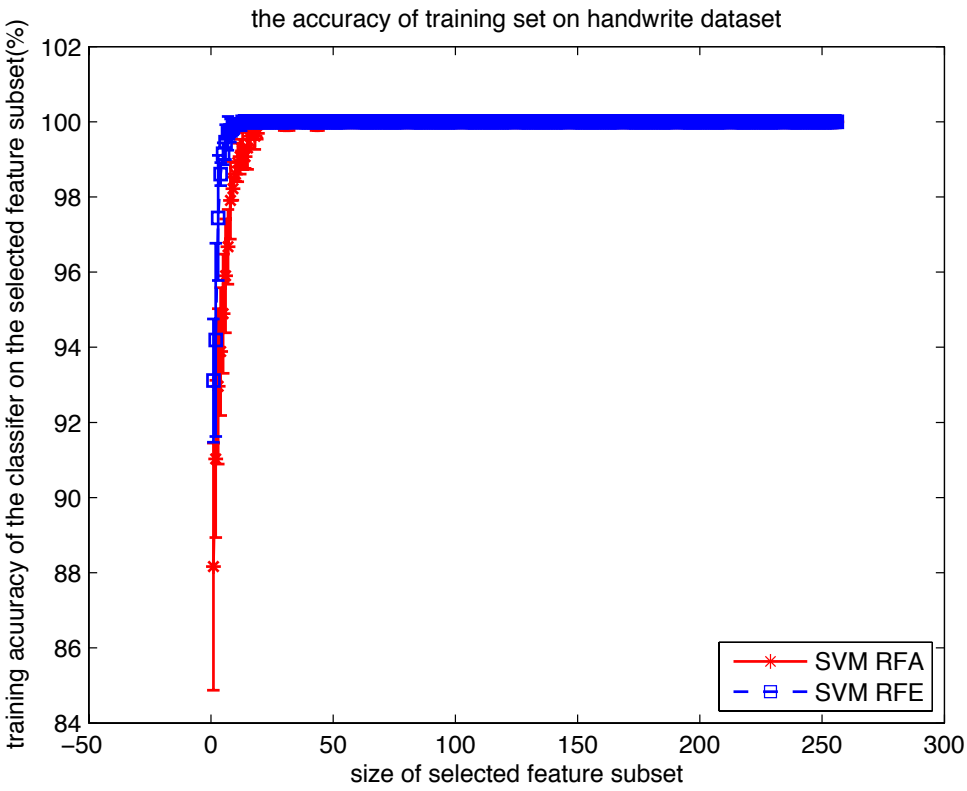


图 5.5　handwrite 数据集上的训练正确率

Figure 5.5　The training accuracy on handwrite dataset

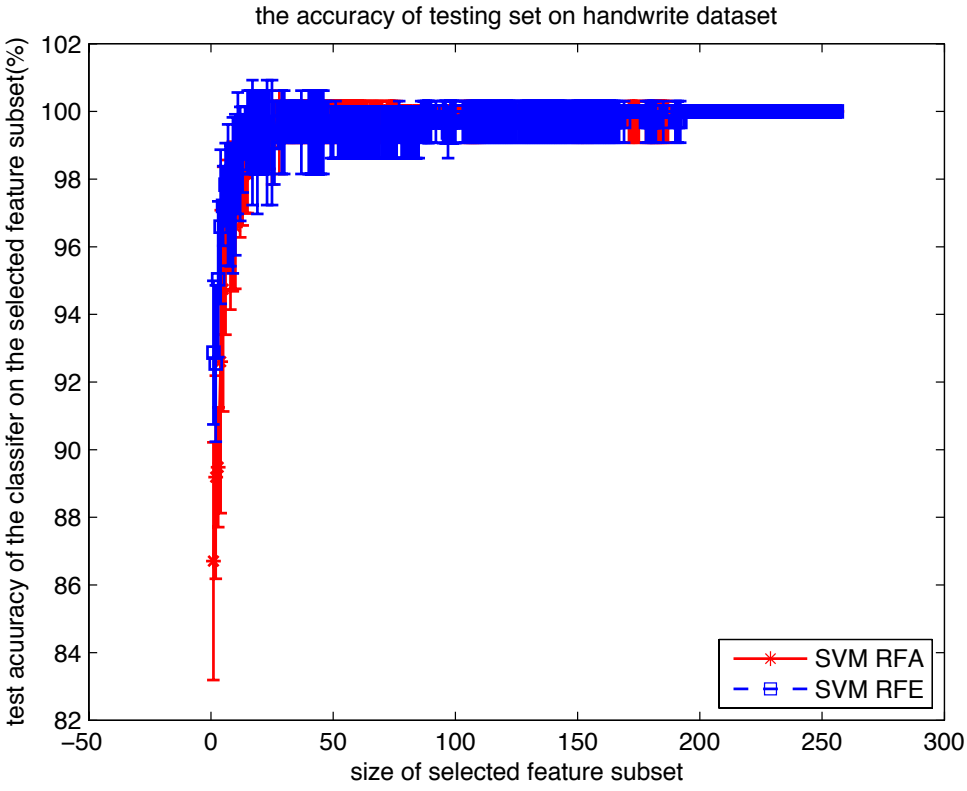


图 5.6　handwrite 数据集上的测试正确率

Figure 5.6 The testing accuracy on handwrite dataset

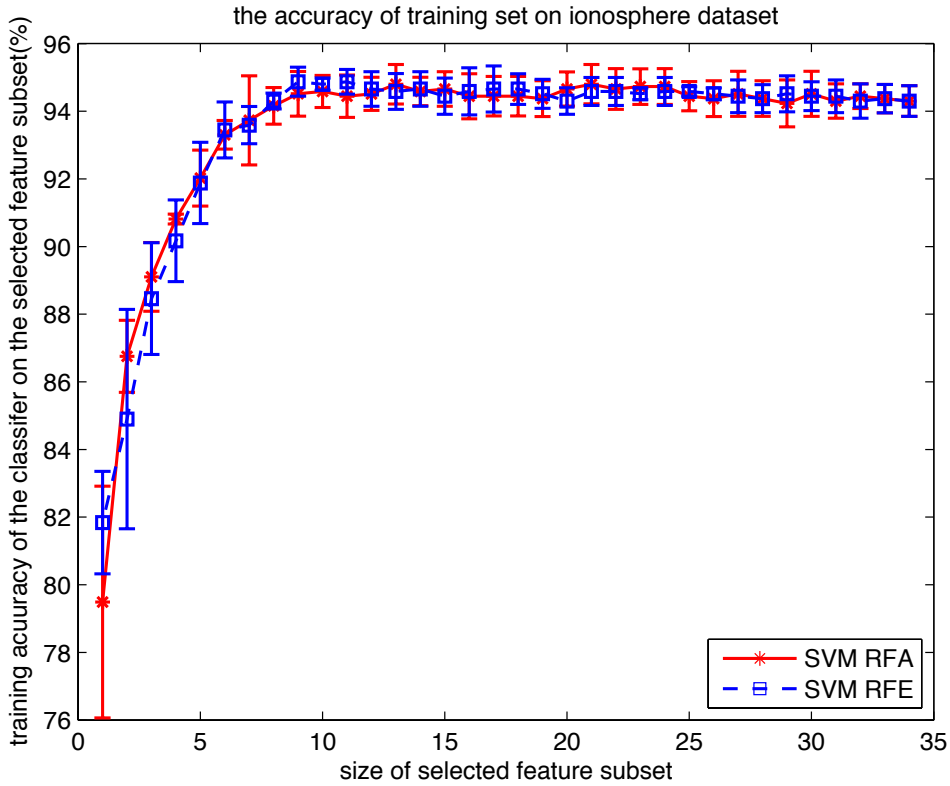


图 5.7　Inosphere 数据集上的训练正确率
Figure 5.7　The training accuracy on Inospher dataset

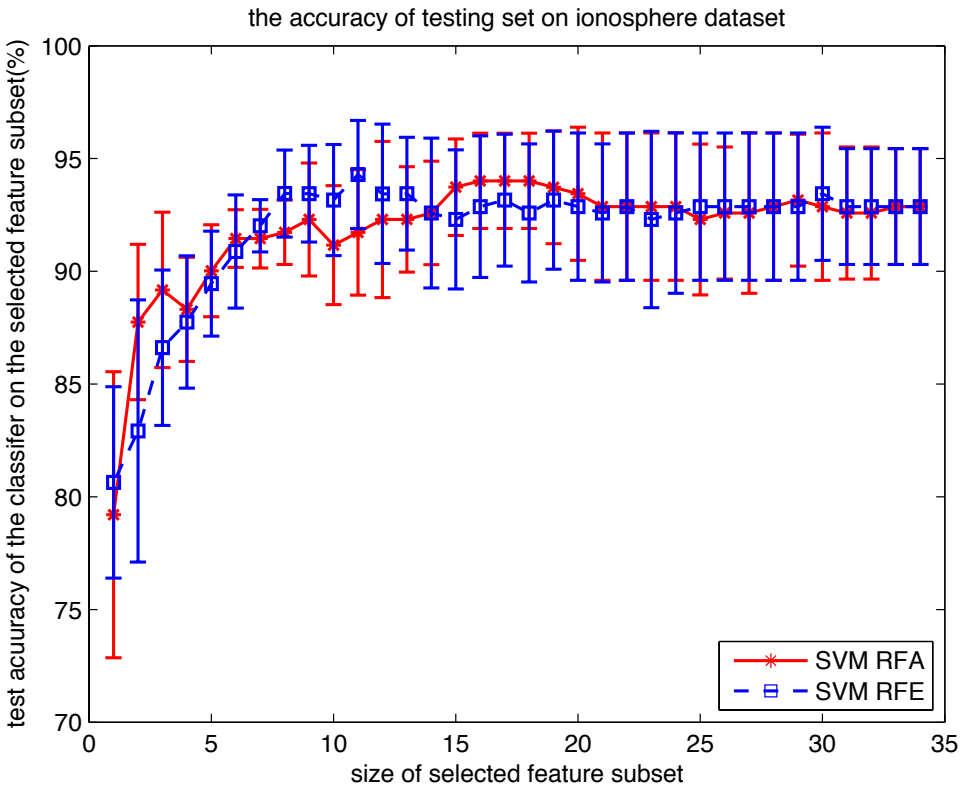


图 5.8　Inosphere 数据集上的测试正确率
Figure 5.8 The testing accuracy on Inosphere dataset

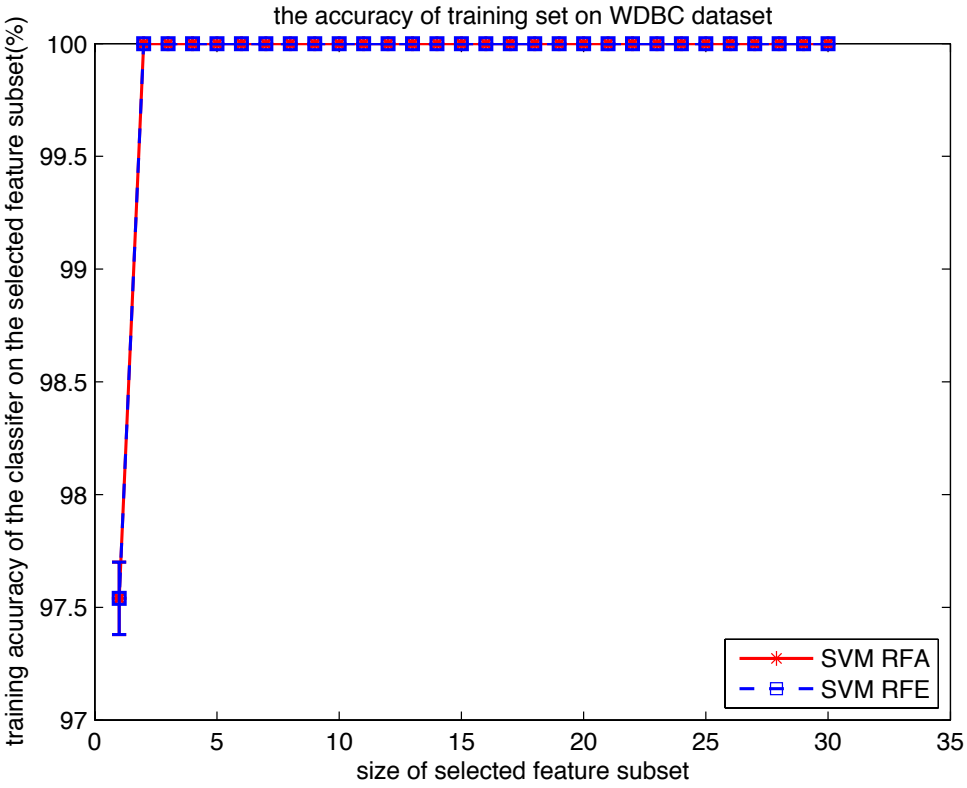


图 5.9　WDBC 数据集上的训练正确率

Figure 5.9　The training accuracy on WDBC dataset

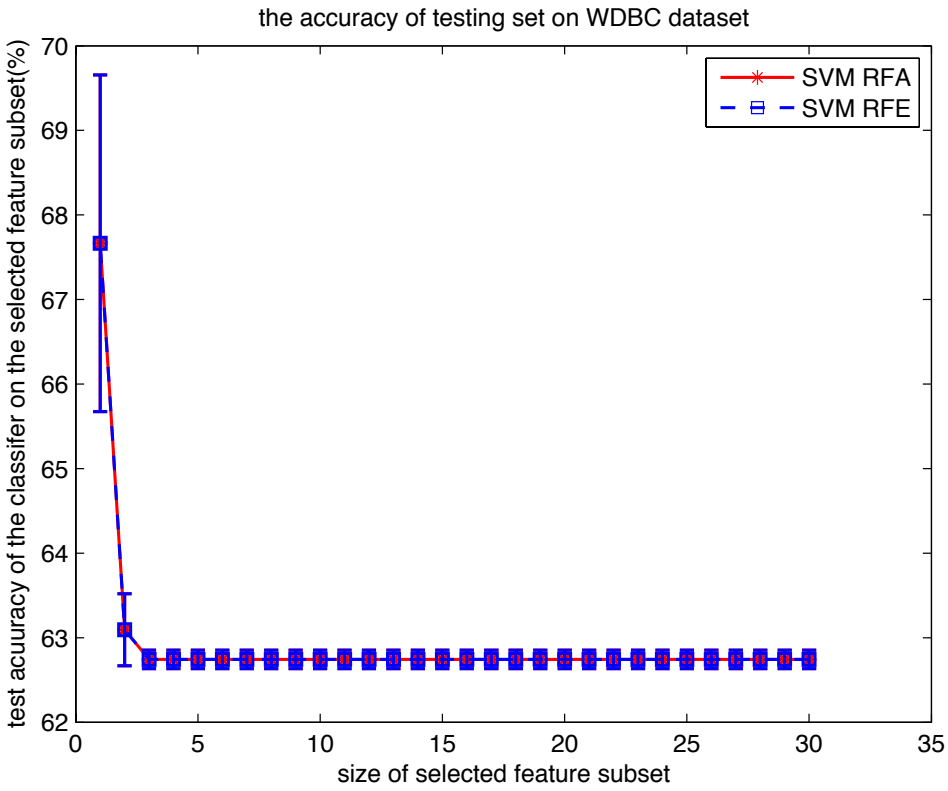


图 5.10　WDBC 数据集上的测试正确率

Figure 5.10 The testing accuracy on WDBC dataset

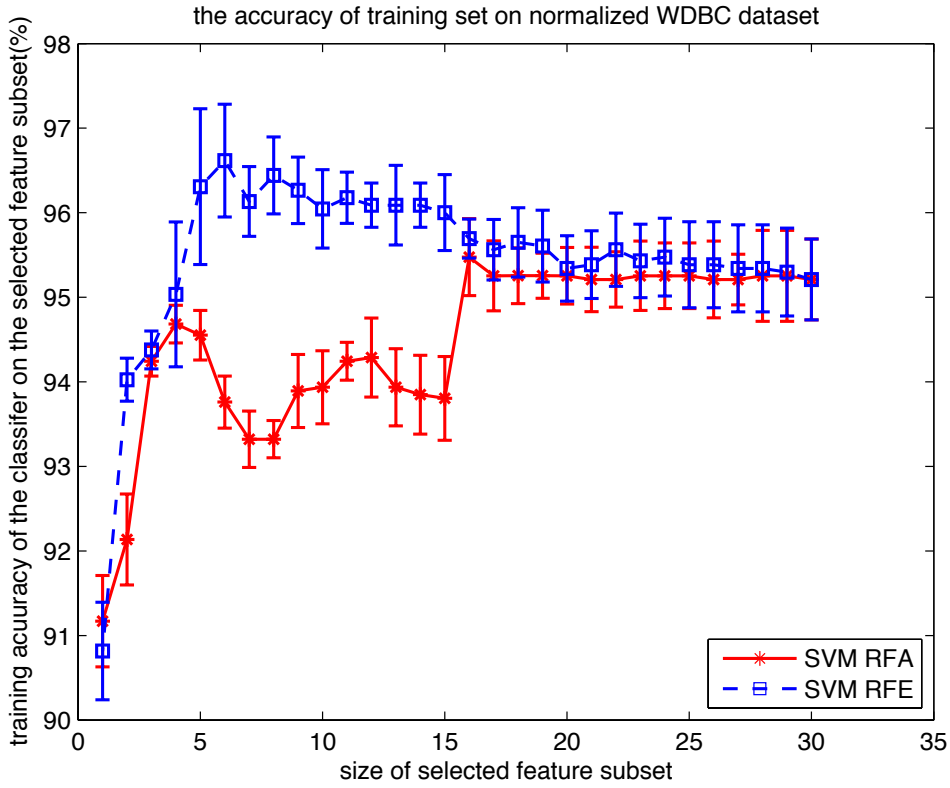


图 5.11　WDBC 的标准化数据集上的训练正确率

Figure 5.11　The training accuracy on normalized WDBC dataset

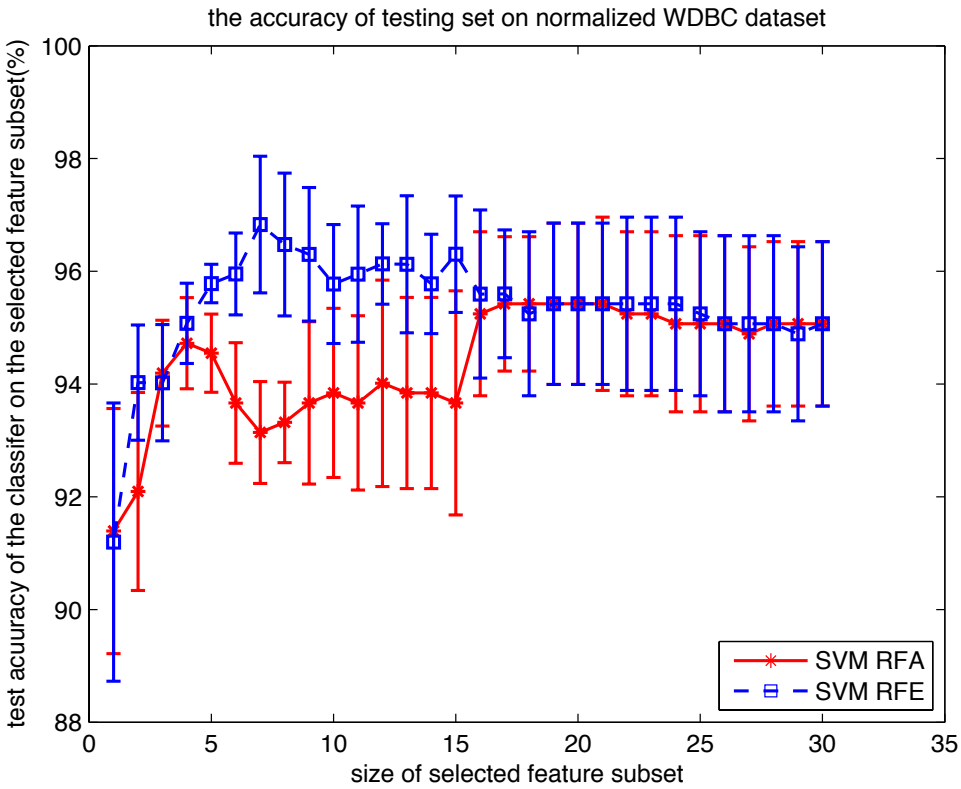


图 5.12　WDBC 的标准化数据集上的测试正确率

Figure 5.12 The testing accuracy on normalized WDBC dataset

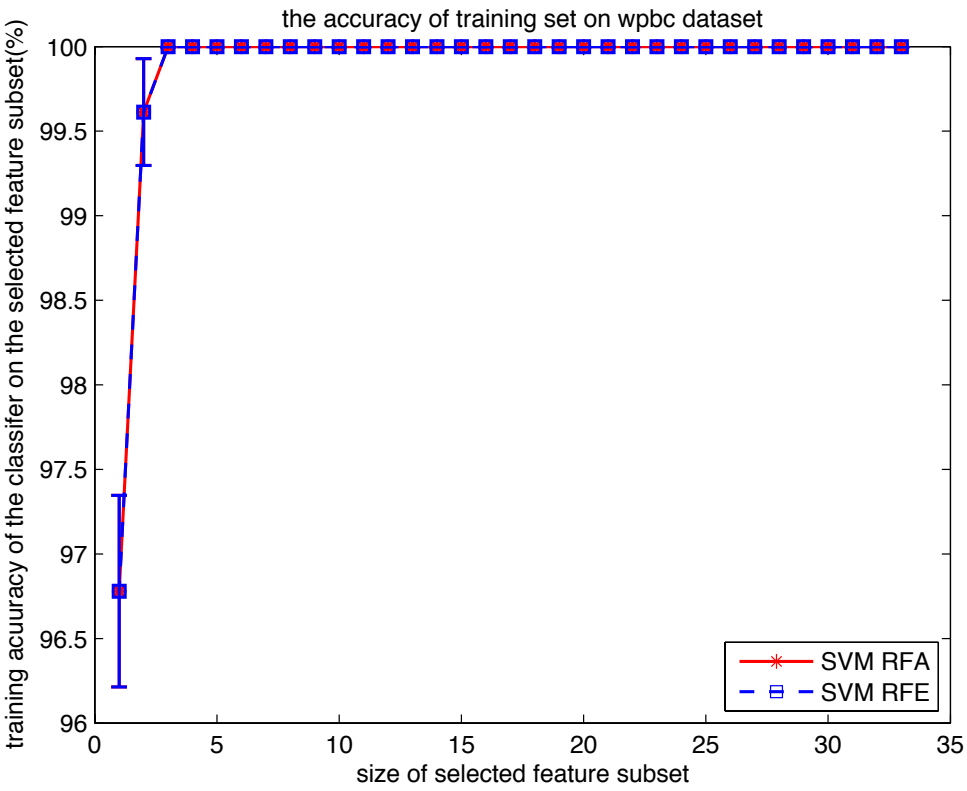


图 5.13　WPBC 数据集上的训练正确率

Figure 5.13　The training accuracy on WPBC dataset

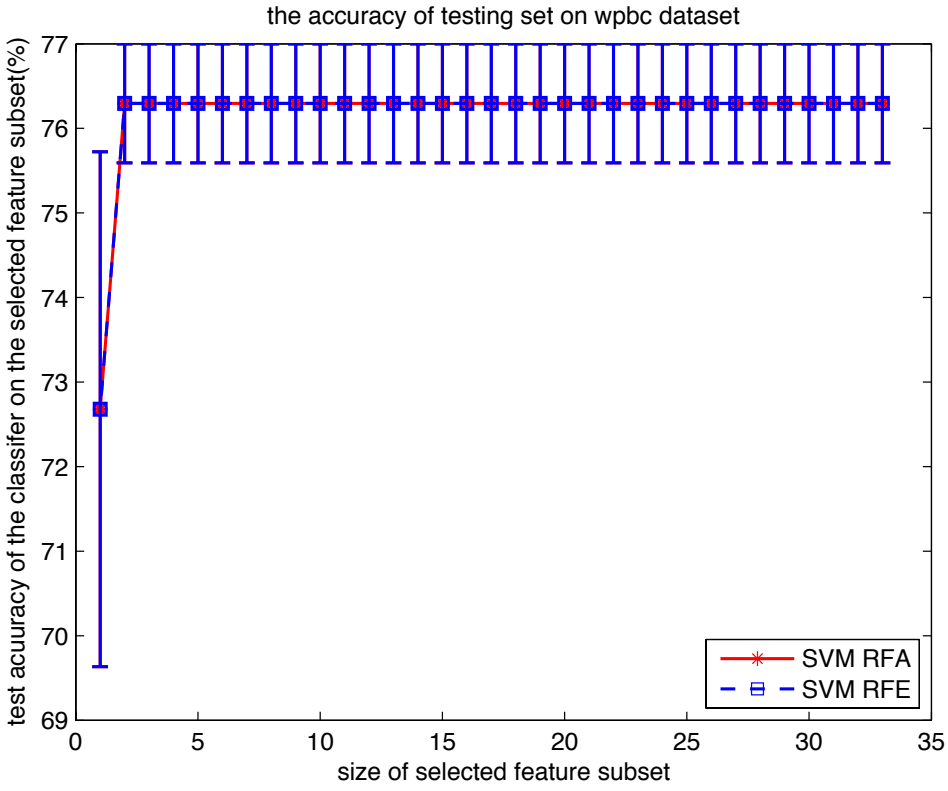


图 5.14　WPBC 数据集上的测试正确率

Figure 5.14 The testing accuracy on WPBC dataset

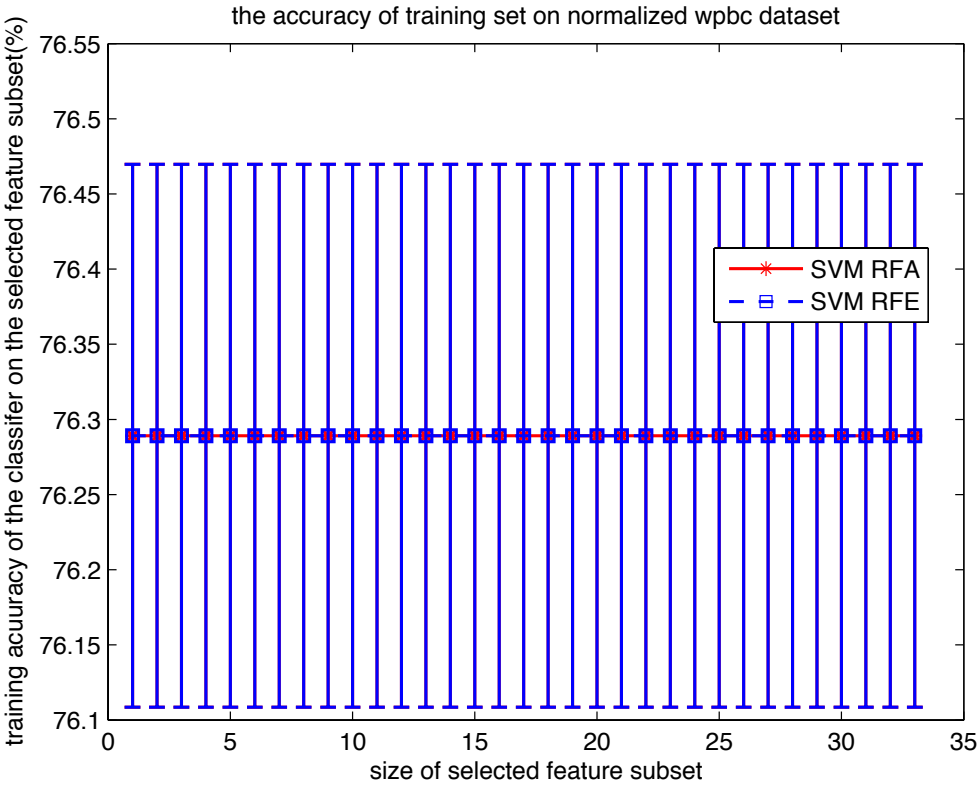


图 5.15　WPBC 标准化数据集上的训练正确率

Figure 5.15　The training accuracy on normalized WPBC dataset

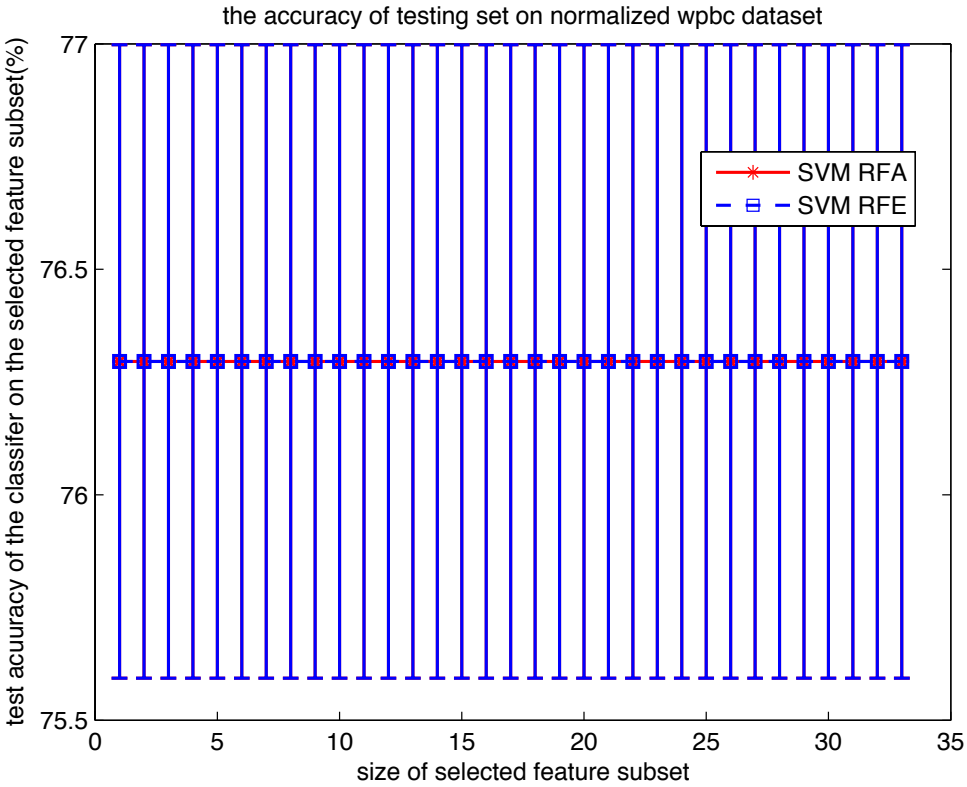


图 5.16　WPBC 标准化数据集上的测试正确率

Figure 5.16 The testing accuracy on normalized WPBC dataset

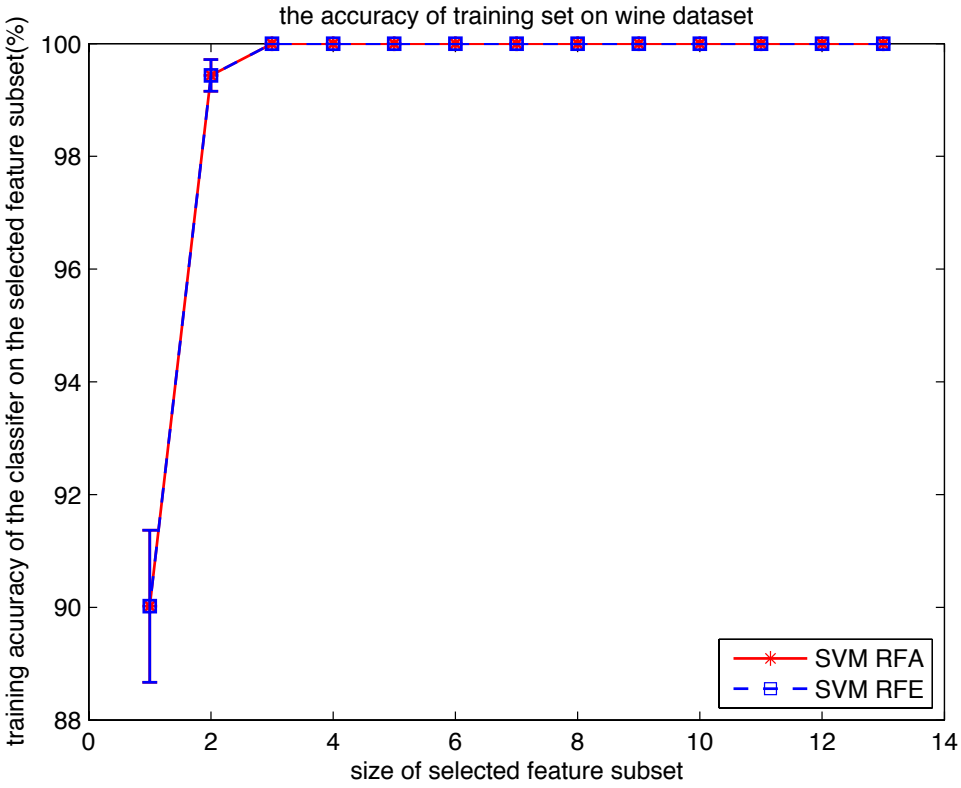


图 5.17　Wine 数据集上的训练正确率

Figure 5.17　The training accuracy on Wine dataset

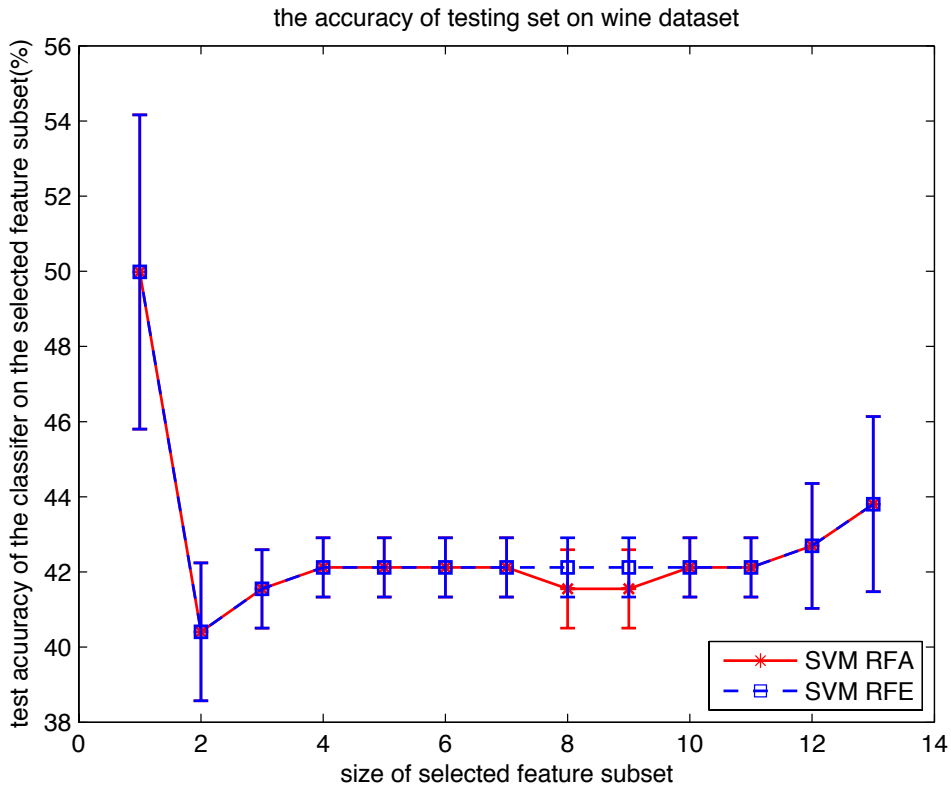


图 5.18　Wine 数据集上的测试正确率

Figure 5.18 The testing accuracy on Wine dataset

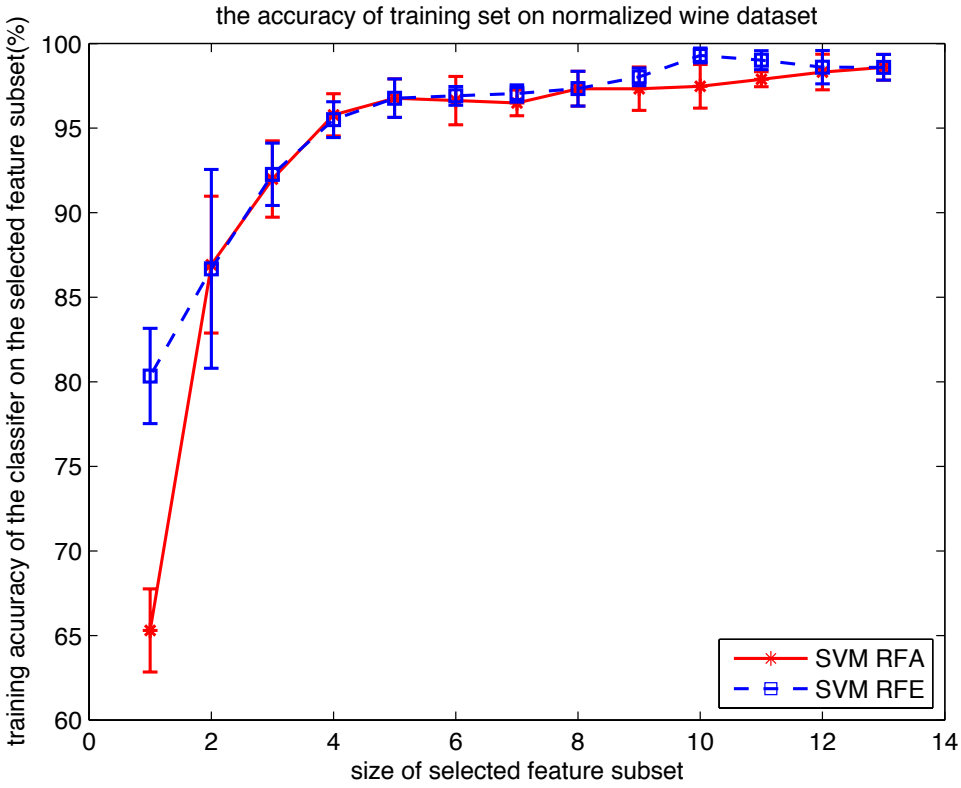


图 5.19　Wine 标准化数据集上的训练正确率

Figure 5.19　The training accuracy on normalized Wine dataset

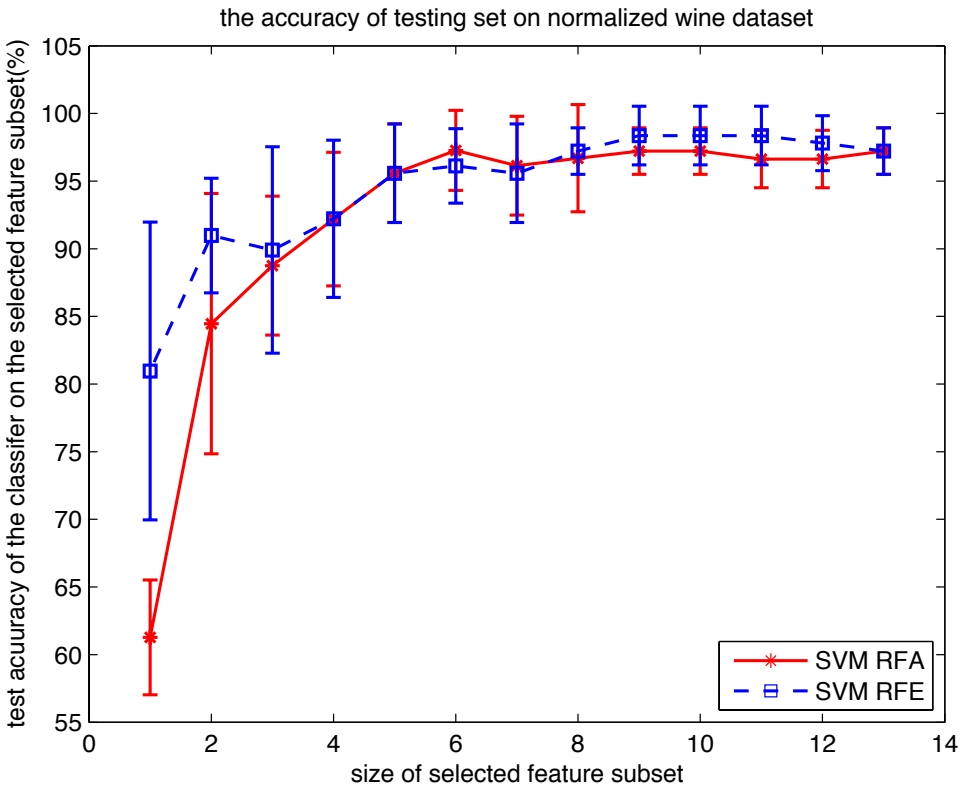


图 5.20　Wine 标准化数据集上的测试正确率

Figure 5.20 The testing accuracy on normalized Wine dataset

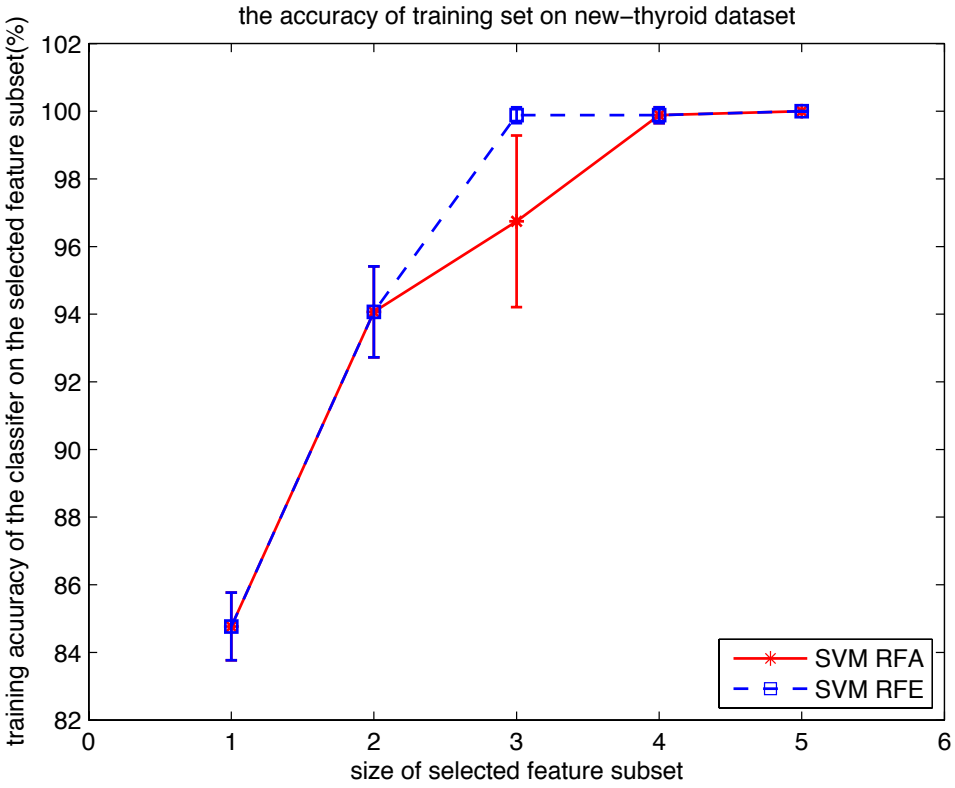


图 5.21　thyroid disease 数据集上的训练正确率

Figure 5.21　The training accuracy on thyroid disease dataset

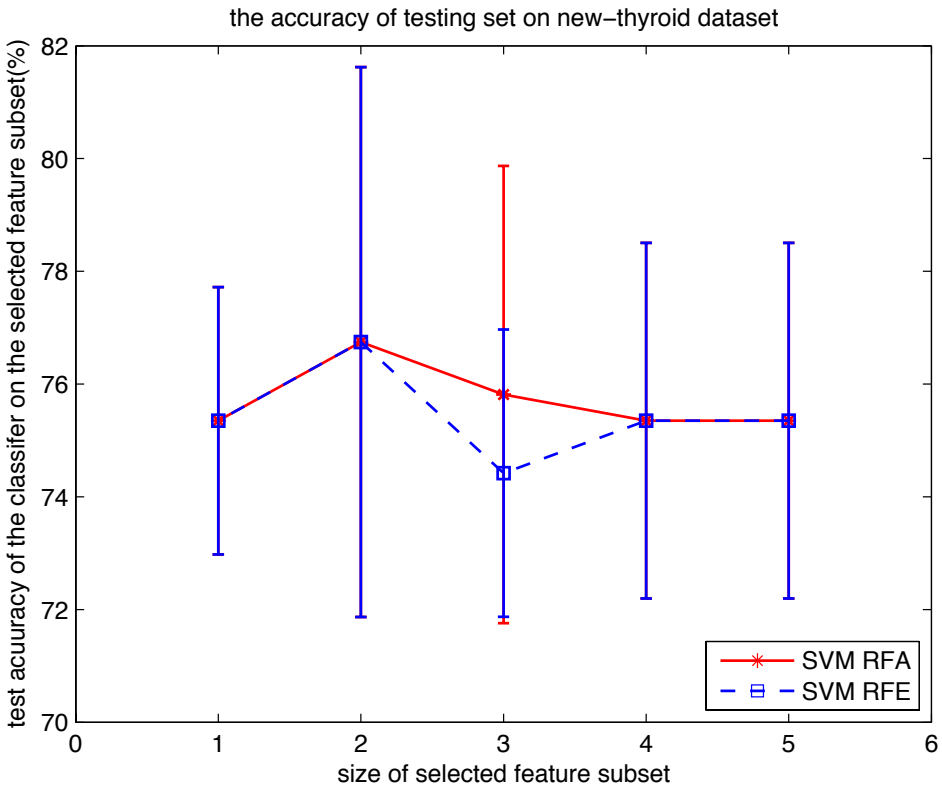


图 5.22　thyroid disease 数据集上的测试正确率

Figure 5.22　The testing accuracy on thyroid disease dataset

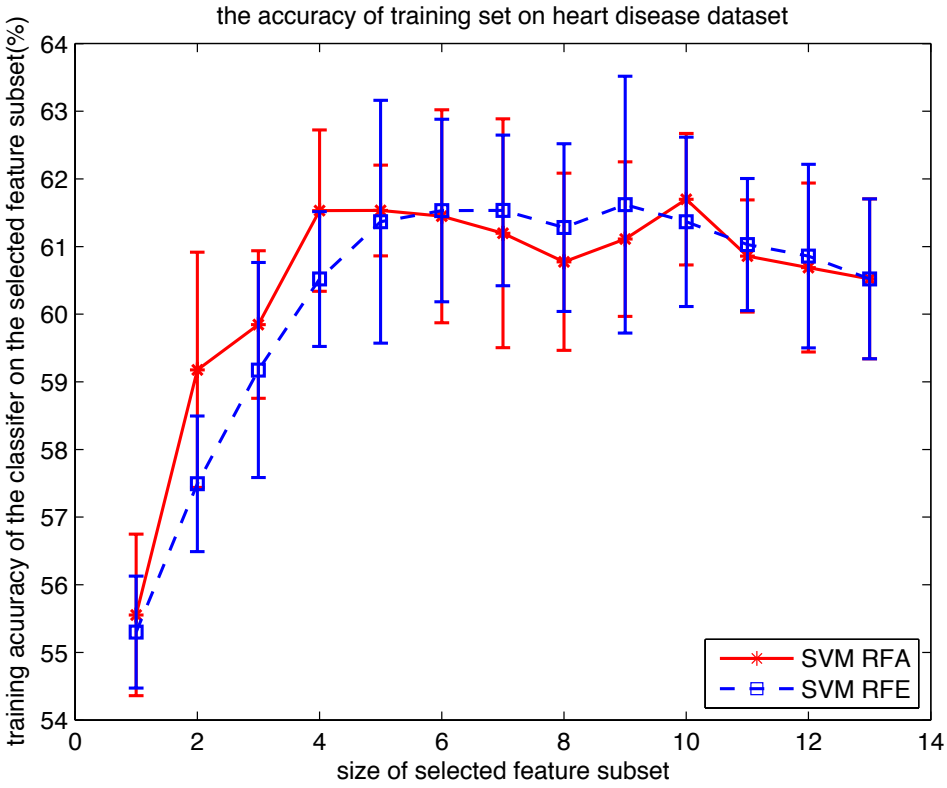


图 5.23　Heart disease 数据集上的训练正确率
Figure 5.23　The training accuracy on Heart disease dataset

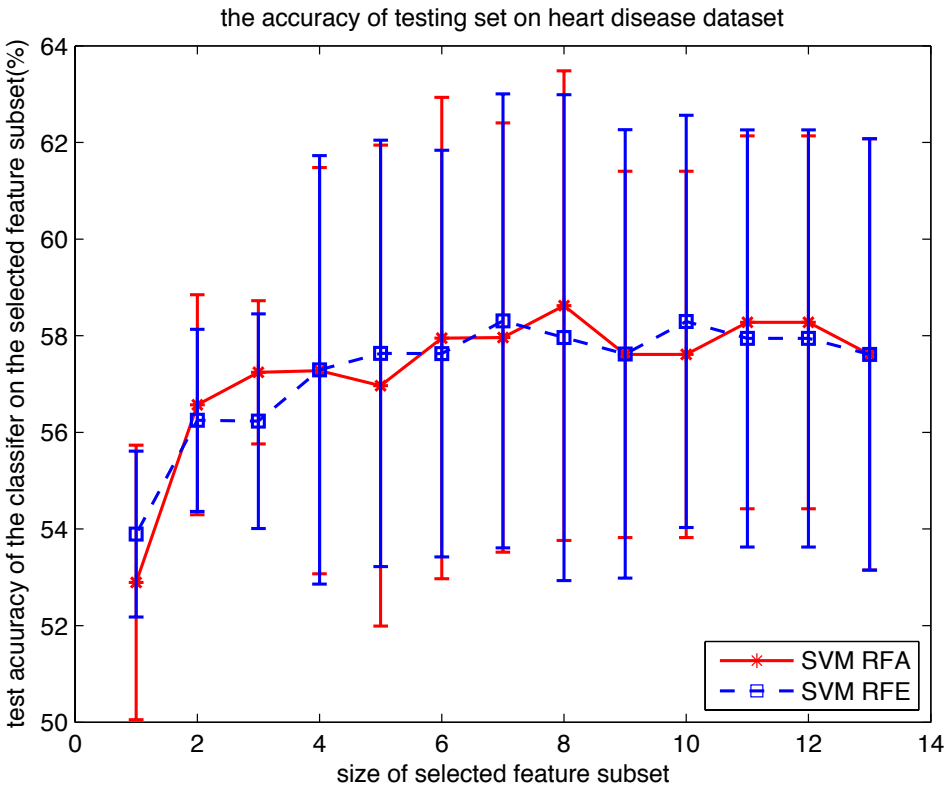


图 5.24　Heart disease 数据集上的测试正确率
Figure 5.24 The testing accuracy on Heart disease dataset

从图 5.1 可以看出 SVM RFA 和 SVM RFE 的训练正确率在选择前 6 个重要特征的时候达到了训练集正确率的第一个小高峰，一个局部的极值；此时图 5.2 的测试正确率比较显示：SVM RFE 方法的测试正确率高于本章提出的 SVM RFA。而当选择到前 18 个重要特征构成特征子集的时候，从图 5.1 可见两种方法的训练集正确率都达到近似全局最优值，此时从图 5.2 可看出：SVM RFA 方法的测试集正确率稍稍胜出 SVM RFE 方法的测试正确率。从图 5.1 和图 5.2 还可以看出，SVM RFE 只要前 16 个特征就可以达到 和 SVM RFA 选择 18 个特征相同的性能，SVM RFE 的第 17 和第 18 个特征对于提高分类正确率的作用不明显。然而无论是 SVM RFA 的选择前 18 个重要特征，还是 SVM RFE 的保留最重要的 16 个重要特征，两个特征选择算法的迭代次数都是 18 次。因为，SVM RFE 是后向剔除不重要的特征，而 SVM RFA 是前向加入重要的特征。这样加入 18 个特征循环 18 次，而后向剔除保留 16 个特征的循环次数是 34-16＝18 次。

从表 5.1 可见 SVM RFE 的 5 折交叉验证实验中，每折前 16 个的特征的交集构成特征子集｛34, 21, 28, 33, 15, 29, 16, 4, 27, 5, 9, 6, 20, 22, 14, 12｝，共 16 个特征；SVM RFA 的 5 折交叉验证实验，每折前 18 特征的交集构成特征子集{34, 28, 21, 15, 16, 14, 5, 22, 7, 20, 9, 4, 31, 30, 10}，共 15 个特征。他们的公共特征是｛34, 28, 21, 15, 16, 14, 5, 22, 20, 9, 4｝， 共 11 个特征。这 11 个特征对实现分类，即对 eryrhenato-squamous 疾病的正确诊断起重要作用。

以上分析可得出：对于 Dermatology 数据集 SVM RFE 和 SVN RFA 两种特征选择算法的性能区别不是很大。

图 5.3 的 Glass 数据集训练正确率比较显示：无论 SVM RFE 还是 SVM RFA 都在选择 5 个特征的时候达到最佳；图 5.4 的测试正确率比较说明：选择 5 个特征时，SVM RFA 的平均分类正确率为 94.89％，高于 SVM RFE 的分类正确率 93.99%，因此，SVM RFA 的泛化性能更优。

表 5.2 显示 SVM RFE 在 5 折实验中，有 3 折的前 5 重要特征是｛2，3，4，6，7｝，2 折的前 5 个重要特征是｛2，4，5，6，7｝；而 SVM RFA 的 5 折实验结果，其中有 4 折的前 5 个重要特征为｛2，3，4，6，7｝，只有一折的前 5 个重要特征为｛2，4， 5，6，7｝。结合图 5.4 的测试平均正确率分析可知：第 5 和第 3 个特征相比，第 3 个特征对于分类的意义更大。

由此可见，对于 Glass 数据集 SVM RFE 和 SVN RFA 两种特征选择算法的性能相比，后者泛化性能更优。

图 5.5 关于 handwrite 数据集的训练集正确率比较显示，SVM RFE 在选择前 13 个重要特征的训练正确率就到 100%，而 SVM RFA 要选择到前 20 个重要特征时，才达到 100%的训练正确率；从图 5.6 可见，两种特征选择算法的测试正确率在留下前 20 个重要特征，或者选择到第 20 个重要时都能达到 99%以上。但是，对于 SVM RFA 算法当选择前 13 个重要特征时，分类正确率稍稍弱于 SVM RFE 的只剩下 13 个特征时的分类正确率。然而，SVM RFA 是不断加入特征，SVM RFE 是迭代地剔除特征，对于该数据集，达到只选择前 20 个重要特征，SVM RFA 需要 20 次迭代，而 SVM RFE 却需要 256-20＝236 次迭代，达到只保留前 13 个重要特征的迭代次数是 256-13=243 次。因此，SVM RFA 的效率更高。

表 5.3 显示 SVM RFA 的 5 折交叉验证实验，5 折中前 20 个特征的交集为{146，130，145，162，193，178，177，161，112，114， 113， 111}，共 12 个特征；SVM RFE 的 5 折交叉验证实验，5 折的前 13 个特征的交集是｛128}，只有一个特征。实验中只有第三折当只剩下 7 个特征｛128，60，146，113，96，82，95｝的时候，训练集和测试集的分类正确率都达到了 100%。

由表 5.3 与图 5.5 和 5.6 的分析可见，对于 handwrite 这样的稍高维数据集，达到同样的分类性能时，SVM RFA 的效率更高。

从图 5.7 可见，对于 Inopshere 数据集，当选择前 9 个重要特征时，SVM RFA 和 SVM RFE 两个特征选择算法的训练集正确率都最优，但是从图 5.8 可见，这时的测试正确率 SVM RFE 稍稍高于 SVM RFA，前者的测试正确率为 93.44%，后者为 92.30%。然而，前者达到选择 9 个特征需要迭代地剔除掉 34-9=25 个特征，迭代次数是 25；而后者只要依次加入前 9 个高权重的特征，迭代次数为 9。

表 5.4 的实验结果显示，SVM RFA 算法的 5 折交叉验证实验，5 折的前 9 个特征的交集为｛3，8，5，27，1，7}，共 6 各特征。与此相应，SVM RFE 的 5 折交叉验证实验的各折前 9 个特征的交集为｛3，27，22， 7， 8， 5， 1， 6}，共 9 个特征。SVM RFE 和 SVM RFA 的 5 折交叉验证实验的平均测试正确率分别为 93.44%和 92.30%。

因此就分类正确率来讲，SVM RFE 在 Inosphere 数据集上略显优势，但是其迭代次数接近 SVM RFA 算法迭代次数的 3 倍。

图 5.9 关于 WDBC 的实验结果显示，当选择前 2 个重要的特征时，SVM RFE 和 SVM RFA 两种特征选择算法的训练正确率均达到 100％；从图 5.10 可见此时两个算法的测试集正确率也相同，都超过了 63％。另外，从这两个图展示的实验结果可见，该两个特征选择算法在选择前 1 个最终要的特征时，训练集的正确率超过 97.5%，测试正确率接近 68％。

表 5.5 的实验数据显示，该两算法的 5 折交叉验证实验，每一折的前两个重要特征都是第 24 个特征和第 4 个特征。根据 UCI 机器学习数据库[158]关于 WDBC 数据集的描述可知，这两个特征对应该数据集的特征是细胞核特征 d) area。由此可见，该特征对于乳腺癌（Breast Cancer）的诊断具有重要意义。

为了消除不同特征的不同测度对于其重要性判别的影响，我们对 WDBC 数据集进行了进一步的实验，先对所有特征采用最大最小标准化方法进行了标准化，然后重复 SVM RFA 和 SVM RFE 算法。从图 5.11 的实验结果来看，SVM RFE 算法在迭代剔除地只剩下 6 个重要特征的时候，训练正确率达到最优；SVM RFA 算法在选择到前 16 个重要特征的时候，训练正确率最优。相应地，图 5.12 显示：SVM RFE 在只有 6 个特征是的测试正确率接近 96%，为 95.95%，SVM RFA 对应 16 个特征的测试正确率超过 95％，但是略逊色于 SVM RFE 对应 6 个最佳特征的 95.95%，为 95.26%。但是对于该数据集的 30 个特征，SVM RFA 选择 16 个重要特征的迭代次数为 16，而 SVM RFE 选择 6 个重要特征的迭代次数为 30-6=24 次，比 SVM RFA 的迭代次数多$1/2$倍。

另外，从表 5.6 的数据可见，SVM RFE 算法 5 折交叉验证实验的前 6 个重要特征的交集为{21，28，23，8}，共 4 个特征，对应的相关细胞核特征为 a) radius (mean of distances from center to points on the perimeter)，c) perimeter 和 h) concave points (number of concave portions of the contour)。 SVM RFA 的 5 折交叉验证实验，每一折的前 16 个特征的交集是｛28，8，21，23，3，1，4，24，7，27，26，6，11，13，14，22｝，共 16 个特征，对应的细胞核特征是 a) radius (mean of distances from center to points on the perimeter)，b) texture (standard deviation of gray-scale values)，c) perimeter，d) area，f) compactness (perimeter^2 / area - 1.0)，g) concavity (severity of concave portions of the contour)和 h) concave points (number of concave portions of the contour)。可见 SVM RFE 方法的公共特征子集是 SVM RFA 方法的公共特征集合的子集。另外，SVM RFA 方法在未进行标准化数据集上选择的特征子集是在标准化后数据集上算法选择特征子集的

子集。而 SVM RFE 方法在标准化和非标准化数据集上所选择的特征子集的交集为空集。

以上关于 WDBC 的 5 折交叉验证实验结果分析可见：SVM RFA 的综合性能优于 SVM RFE。

图 5.13 显示，对于 WPBC 数据集，无论 SVM RFA 还是 SVM RFE 都在选择到前 3 个重要特征时，训练正确率达到 100%，此时的测试正确率为 76.3%。从图 5.14 还可以看到，测试正确率从选择到第 2 个特征开始就达到了最高值 76.3%，之后测试正确率一直保持这个最高值，不在随特征的增加变化。从表 5.7 可见，SVM RFA 和 SVM RFE 两种特征选择算法所选择前 3 个重要特征都是第 25 个，第 5 个和第 1 个特征。但是对于该数据集来说，选择到前 3 个重要特征，SVM RFA 需要 3 次迭代，而 SVM RFE 需要 33-3=30 次迭代。由此可见，对于 WPBC 数据集，SVM RFA 算法的性能更好。

同 WDBC 数据集，我们对 WPBC 数据集也进行了最大最小的标准化实验，图 5.15 和图 5.16 的实验结果显示，只要选择第 1 个重要属性，SVM RFA 和 SVM RFE 两个特征选择的算法的训练正确率和测试正确率就都达到最优值 76.3%，而且不再变化。从表 5.8 可看出 SVM RFA 的 5 折交叉验证实验，每一折的首要特征是第 1 个特征；然而 SVM RFE 的 5 折实验每一折的首要特征不完全相同，前两折的首要特征均为第 22 个特征，而第三折的首要特征为第 25 个特征，第三、四折的首要特征分别是第 13，14 个特征。以上分析显示，对于 WPBC 数据集，标准化似乎没有必要。

对于 wine 数据集，从图 5.17 可见，当选择到前 3 个重要特征时，SVM RFA 和 SVM RFE 的训练正确率达到 100%，此时的测试正确率从图 5.18 可见不足 42%。图 5.17 和 5.18 还显示，当选择最重要的一个特征时，SVM RFA 和 SVM RFE 的训练正确率达到 90%，测试正确率达到 50%。由此可见第 1 个权值最大的特征对于分类的重要性，由表 5.9 可知该特征对应 wine 数据集的第 13 个特征，即：13)Proline（脯氨酸）。

对 Wine 数据集的标准化数据集实验结果图 5.19 和图 5.20 显示，SVM RFE 在选择到前 10 个重要特征时，训练正确率达到最高，此时的测试正确率为 98.36%；从表 5.10 可见，此时被剔除的特征为第 5，8，9 三个特征。对应真实特征：5) Magnesium，8) Nonflavanoid phenols 和 9) Proanthocyanins。

相应的，图 5.19 显示的 SVM RFA 在 Wine 的标准化数据集上的实验结果是，当选择到第 5 个重要特征时，训练正确率达到一个局部最优 96.77%，对应的图 5.20 的测试正确率是 95.58%。表 5.10 显示 SVM RFA 的 5 折交叉验证实验的各折中前 5 个重要特

征的交集是{12，7，13，1，10}，共 5 个特征。另外，图 5.19 和表 5.10 显示，SVM RFA 的 5 折交叉验证实验的第 6 个重要特征，也即真实特征的第 6 个特征，对于训练正确率几乎没有贡献，但是却使测试正确率达到了一个局部极优值的分类正确率 97.27%。但是图 5.19 显示，SVM RFA 在全部特征被选择时达到全局最优的训练正确率 98.6%，而此时的 5 折交叉验证实验的平均测试正确率为 97.22%，不如只选择 6 个重要特征时的测试正确率。

以上分析显示，对于 Wine 数据集来说，SVM RFE 的分类性能优于 SVM RFA；但是所选择的特征子集规模不如 SVM RFA 算法。

图 5.21 和图 5.22 关于 thyroid disease 数据集的实验结果显示：SVM RFA 算法和 SVM RFE 算法在选择到 4 个属性的时候，训练正确率相同，都接近 100%；此时的测试正确率也相同，为 75.35%。然而 SVM RFE 在选择 3 个属性的时候训练正确率就达到和 SVM RFA 选择 4 个特征的训练正确率相同的水平；但是，SVM RFE 选择 3 个属性的相应 5 折交叉验证实验的平均测试正确率为 74.42%，不如 SVM RFA 选择 3 个重要属性时的 75.81%的分类正确率。由表 5.11 可知，SVM RFA 选择的前 4 个特征构成的特征子集，有 3 折为{1，5，4，2}，两折为{1，5，2，4}；而 SVM RFE 选择的前 4 个特征的特征子集，5 折实验的结果均为{1，5，2，4}。由此可见，第 4 个特征对于分类很重要。同时图 5-22 还显示，当取 3 个特征的时候，SVM RFA 的分类正确率比 SVM RFE 高出 1 个多百分点；当选择 4 个特征的时候，两个特征选择算法的 5 折交叉验证实验的平均测试正确率相同，均为 75.35%。以上分析显示，对于 thyroid disease 数据集，第 1 和 5 两个属性最重要，只有第 3 个属性对于分类的贡献可以忽略不计。

图 5.23 和图 5.24 的 Heart disease 数据集实验结果显示：当选择到前 4 个属性时，SVM RFA 的训练正确率达到了一个局部最优，当选择到前 10 个重要属性时，SVM RFA 的训练正确率达到全局最优；这两个最优点对应的测试集平均分类正确率分别为 57.29%和 57.61%。由表 5.12 可知，对应的各折前 4 个特征的交集为{12，13，10}，共 3 个特征；前 10 个特征的交集为{12，10，6，13，8，7，9，3}，包含 8 个特征。SVM RFE 在保留 7 个特征时，有一个局部最优的训练正确率，在保留 9 个特征时，达到全局最优的训练正确率；分别对应的测试即正确率为 58.31%和 57.62%。从图 5-24 还可以得出，在选择 9 个特征和 4 个特征时，SVM RFA 和 SVM RFA 两个特征选择算法的测试正确率相同，分别为 57.62%和 57.29%。由此可见，该两个特征选择算法对于 Heart disease 数

据集的效率和性能基本相同。

以上关于 SVM RFA 和 SVM RFA 特征选择算法在 9 个数据集的 5 折交叉验证实验结果的详细分析显示，对于较低维数据集两个算法的效率和分类性能差别不大，但是对于维数比较高的数据集进行为了分类的特征选择时，SVM RFA 特征选择算法在效率上高于 SVM RFE 算法。原因在于 SVM RFA 是通过迭代不断地加入重要特征；而 SVM RFE 算法则相反，是通过迭代，反复地剔除不重要的、对分类最没有贡献的特征。

5.4 小结

本章鉴于 SVM 对于非线性可分问题的最大泛化性能，提出基于 SVM 分类模型的适用于多类分类问题的特征选择算法 SVM RFE 和 SVM RFA，解决了第 2～4 章分别基于 G-score、D-score、DFS 与 SVM 的特征选择算法在处理非线性可分问题时，有可能使得具有有效区分性能的特征被误剔除的潜在缺陷；并解决了 Guyon 的 SVM-RFE 特征选择算法只适用于两类分类问题的缺陷。

UCI 机器学习数据库的 9 个经典数据集的 5 折交叉验证实验证明：本章提出的推广 SVM RFE 特征选择算法，和基于前向选择思想的适用于多类分类问题的 SVM RFA 特征选择算法能在保持或提高分类正确率的前提下，实现有效的特征选择。9 个 UCI 经典数据集的 5 折交叉验证实验，在 8 个数据集上 SVM RFA 算法优于 SVM RFE 算法。实验还表明，对于较低维数据集，两个特征选择算法的效率差别不大，但是对于维数比较高的数据集，为了进行有效分类，进行特征选择时，SVM RFA 特征选择算法的效率明显优于 SVM RFE 算法。

第六章　基于 SVM 分类模型的基因选择算法

第五章我们研究了基于 SVM 分类模型的特征选择算法，以训练所得 SVM 分类器的权重来度量相应特征的区分度，进行特征选择。给出了对于多类分类问题，如何利用 SVM 分类模型计算特征区分度的方法；提出了适用于多类分类问题的推广 SVM RFE（SVM Recursive Feature Elimination）特征选择算法，和基于前向顺序选择思想的 SVM RFA（SVM Recursive Feature Addition）特征选择算法； 并用 UCI 机器学习数据库的数据集对这两种特征选择算法进行了实验比较。5 折交叉验证实验表明，对高维数据集进行特征选择时，SVM RFA 的效率优于 SVM RFE。

生物技术的发展加速高维小样本的癌症基因数据集的产生[9, 25~31]。在基因数据集中，样本通常只有几十个，而作为描述每一个样本的基因却有成千上万之多[33, 25~29]。 特征选择成为对基因数据集分类分析的首要的步骤。若将 SVM RFE 或者 SVM RFA 直接应用于基因选择，每次迭代中依然逐个剔除或加入基因特征，时间消耗将成为屏障。如何对这些高维小样本的基因数据集进行特征选择，成为特征选择方法研究的新挑战。本章针对基因数据集的特点，结合上一章的研究结论，提出基于 SVM 分类模型的基因选择算法——SVM RRFA（SVM Recursive Random Feature Addition, RRFA）。该方法引入随机思想，能针对具体的基因数据集，在每次迭代中一次剔除或加入随机数个基因。同时，为了减少算法时间开销，提出了简化 SVM RRFA 基因选择算法。

本章内容组织如下，首先介绍基于 SVM 分类模型的随机特征选择的必要性；然后介绍基于 SVM 的随机特征（基因）选择算法 SVM RRFA 和简化 SVM RRFA 基因（特征）选择算法；最后用斯坦福大学的 3 个基因数据集对该基因选择算法进行了实验测试，并对实验结果进行了分析。

6.1　基于 SVM 分类模型的随机特征选择的必要性

基因数据集的特点是：样本个数远远少于特征维数，样本空间是一个高维的稀疏空间。对该类数据集进行分类分析，发现区分癌症患者和正常人的关键区分基因，建立癌症患者和正常人群之间的最优分类超平面，以便对未知患者进行分类预测和判断。为此，特征选择（基因选择）成为基因数据分类分析研究的首要步骤。

第二～五章研究的特征选择方法的共同特点是：每次迭代中，特征的加入或剔除是一个一个进行地。在基因选择中如果每次迭代依然采用逐个加入或者删除特征的方法，那么，当面对成千上万的基因特征时，时间开销将变得巨大。因此，必须考虑基因数据

集的高维小样本特征，研究针对基因数据集的特征选择方法。Guyon 在其研究[33]中指出，因为基因数据集的维数很高，在 SVM-RFE 每次迭代剔除特征时，可以一次剔除掉上百个特征。但是，Guyon 的研究没有给出每次迭代中，剔除特征的具体个数或者依据；同时，不同基因数据集的维数差别较大，比如，结肠癌数据集[26]的规模为 62*2000，而乳腺癌数据集[29]的规模是 47*24481。对含有几万个基因特征的基因数据集，每次迭代剔除的基因规模依然采用和含有几千基因特征的基因数据集相同的剔除规模，显然是不合适。另外，现有基于 SVM-RFE 框架的基因选择算法[33,71~74,104,116,122~124,130,140,148~153]，还没有关于特征选择过程中每次迭代剔除特征数目的最佳数研究。

因此，有必要研究一种新的适用于基因数据集特点的基因选择方法；该方法能根据具体基因数据集的维数规模确定每次迭代过程中合适的加入或者剔除基因个数的多少，以便得到不同基因数据集的最佳基因子集。

6.2 基于 SVM 分类模型的随机特征选择算法

本小节将提出基于 SVM 的前向顺序随机特征选择算法 SVM RRFA（SVM Recursive Random Feature Addition, SVM RRFA）。由第五章的研究知道，基于 SVM 分类模型的特征选择算法，SVM RFA 优于 SVM RFE。因此，本章在第五章研究的基础上提出基于 SVM 分类模型的随机特征选择算法——SVM RRFA。该算法的思想是：对算法 SVM RFA 进行改进，在其中加入随机思想，根据基因数据集的基因规模，每次迭代中，产生一个随机数，选择随机数规模的基因特征加入；并考虑加入特征对于分类正确率的影响，以进一步减少加入的特征数。算法的具体详细步骤描述如下。

step 1: 初始化被选特征子集 X 为空集，备选特征子集 Y 为全集，备选特征子集的特征规模 *number* 为全部基因个数。

step 2: 用含有备选特征子集特征的训练集训练 SVM，根据一对一的方法得到一个（或多个，对于多类分类问题）两类分类的最优 SVM 分类模型。

step 3: 根据该 SVM 分类模型，计算各特征的区分度；对于多类问题依据 5.1 提出的特征重要性计算方法公式（5-1），计算各特征的权重（即特征区分度），根据权重将各特征降序排序。

step 4: 根据相应基因数据集的基因规模，产生一个随机数 p，若 $p \le number$，试将 Y 中前 p 个特征并入被选特征子集 X；否则，若 $p \succ number$，则置 $p = number$，试将 Y 中前 p 个特征并入被选特征子集 X。

step 5: 对只含有当前被选特征子集 X（第一次迭代时，X 中包含有试加入的前 p 个特征；以后迭代，X 中包含有试加入的前 p 个特征和已经并入 X 中的特征）中特征的训练样本集进行训练，得到相应的 SVM 分类模型；

step 6: 若训练集的分类正确率上升，则产生一个更小的随机数 $Q\,(5 \le Q \le 10)$；否则，若训练集的分类正确率没有上升，则产生一个 $0 \sim 5$ 之间的随机数 Q。

step 7: 若 $Q \le number$，则对被选择子集 X 试并入 step 4 选择的前 p 个特征中的前 Q 个；否则，选择 $number$ 个特征试并入被选特征子集 X。

step 8: 对只含有被选特征子集 X 中（X 中包含有试加入的前 p 个特征中的前 Q 个特征与已经加入 X 中的特征）特征的训练集进行 SVM 训练，得到相应的分类模型。

step 9: 若该模型的分类正确率上升，则将这 Q 个特征加入到被选特征子集 X 中；否则，不加入。

step 10: 从备选特征子集 Y 删除这 p 个特征。

step 11: 若备选特征子集 Y 不空，则转 step 2 ；否则算法结束 。

然而，上述 SVM RRFA 基因选择算法，在每次迭代中都要对包含当前备选特征子集特征的训练集训练新的 SVM 分类器，并重新计算当前备选特征子集中各特征的权重，即各特征的区分度，对其进行重新排序；然后再随机选取相应特征。为了减少算法的时间开销，省去 SVM RRFA 算法在每次迭代中的以上步骤。即，每次迭代不再重新训练新的 SVM 分类模型，各特征的区分度在第一次计算后，以后的迭代中不再重新计算。这样修改，即将上述 SVM RRFA 算法的 step 11 中的转 step 2 修改为转 step 4。修改后的 SVM RRFA 算法被称为简化 SVM RRFA 算法（Simple SVM RRFA）。

6.3 基于 SVM 分类模型的随机特征选择实验结果与分析

为验证本章提出的基因选择算法 SVM RRFA，以及简化 SVM RRFA 算法的性能，选择了普林斯顿大学基因表达工程[164]的三个基因数据集对 SVM RRFA 算法和简化 SVM RRFA 算法进行测试，并对两个算法的实验结果灵敏度、特异性，以及 Matthews 相关系数[165, 166]进行分析。

实验所用基因工程数据集中的三个基因数据集分别是 Carcinoma Data[167]、Adenoma Data[167]，和 Colon Data[25]三个维阵列基因数据集。Carcinoma 数据集含有 36 个样本，18 个癌症患者样本，18 个正常样本；每个样本含有 7457 个基因。Adenoma 数据集含有 8 个样本，4 个癌症病变样本，4 个正常样本；每个样本的基因数为 7086 个。Colon 数

据集含有的样本数为 62 个，其中 40 个样本是 Colon 癌变样本，22 个正常样本；每个样本的特征数为 2000 个基因。

实验采用留一法进行，实验所用 SVM 工具箱为台湾林智仁教授等开发的 LibSvm 工具箱[160]的线性核函数，参数 C＝1：2：50，训练集上采用 2 折交叉验证确定最佳参数 C。

SVM RRFA 算法在 Adenoma、Carcinma 和 Colon 三个基因数据集的详细实验结果分别如表 6.1～表 6.3 所示，相应地该算法三个基因数据集所得的混淆矩阵分别如表 6.4～表 6.6 所示。简化 SVM RRFA 算法在 Adenoma、Carcinoma 和 Colon 三个基因数据集上的详细实验结果如表 6.7～表 6.9 所示；相应的混淆矩阵见表 6.10～6.12 所示。

表 6.1 SVM RRFA 算法在 Adenoma 数据集的实验结果
Table 6.1 the experimental results of SVM RRFA on Adenoma data

所选特征	特征个数	训练集正确率	测试集正确率	C
3886,3127,2910,3887,2510,3885	6	100	100	1
3886,3127,2910,3890,2510	5	100	100	1
3886,3127,2910,3887,2510,3885	6	100	100	1
3886,3127,2910,2510,3887	5	100	100	1
3886,3127,2910,2510,3887,3885,3093,3890,3837,3169	10	100	100	1
3886,3127,2910,2510,3885,3887,3093,3890,3837	9	100	100	1
3886,2910,3127,3887,2510,3885,3093,5701	8	100	100	1
3886,3127,2910,2510,3887,3885	6	100	100	1

表 6.2 SVM RRFA 算法在 Carcinoma 数据集的实验结果
Table 6.2 the experimental results of SVM RRFA on Carcinoma data

所选特征	特征个数	训练集正确率	测试集正确率	C
3011,4952,3791,3304,1694,3235	6	100	100	1
3011,3235,1694,4952,3790,373,3304	7	100	100	1
3011,4952,3304,5068,3235,1694,2514,2503	8	100	100	1
4952,3011,373,3304,5203,3235	6	100	100	1
3011,4952,3304,5068,5697,2514,3235,2642	8	100	100	1
4952,5068,3011,3304,3791	5	100	100	1
3011,3304,3235,4952,2514,1694,2503,5697,4975,3790	10	100	100	1
4952,3011,5068,3304,2514,3235,1694	7	100	100	1
3011,4952,3304,5068,2514,3235,1694,2642,2503,3791	10	100	100	1
3011,4952,5068,3304,3235,1694,2514,2503,5120,3790	10	100	100	1
3011,3304,4952,5068,2514,1694,2503,3235,5120,3791	10	100	100	1
3011,4952,3304,5068,1694,2514,3235,2642,2503	9	100	100	1
3011,4952,3791,3304,1694,3235,2503,5120	8	100	100	1
3011,4952,5068,3304,3791,3235	6	100	100	1
3011,4952,3304,5068,2514,1694,3235	7	100	100	1

3011,4952,3304,5068,2514	5	100	100	1
3011,4952,3304,5068,2514,1694,3235,2503,2642	9	100	100	1
3011,4952,3304,5068,1694,2514,3235,2642,2503	9	100	100	1
3011,4952,3304,5068,3235,1694,2514	7	100	100	1
3011,4952,5068,3304,3791,3235,1694,2503,2514,5120	10	100	100	1
3011,4952,3791,2514,5120,3304	6	100	100	1
3011,4952,5068,3304,3235,1694,2514,2503,5120,3790	10	100	100	1
4952,3011,5068,3304,3791,1694	6	100	100	1
5068,3011,3304,5697,1705,2514,5203,3473,4963	9	100	0	1
3011,4952,3304,5068,3235,1694,2514,2503,5120,5697	10	100	100	1
3011,4952,3304,5068,3235	5	100	100	1
3011,4952,5068,3304,3235	5	100	100	1
3011,4952,3791,3304,1694,3235,2503,5120,2514	9	100	100	1
3011,4952,3304,5068,2514,3235,1694,2503,2642,5203	10	100	100	1
4952,3011,1694,3304,2503	5	100	100	1
3011,4952,5068,3304,3235	5	100	100	1
3011,4952,3304,5068,3235,1694,2514	7	100	100	1
3011,4952,3304,5068,2514,3235	6	100	100	1
3011,4952,3304,5068,3235,1694,2514,2503,5120,5697	10	100	100	1
3011,4952,3304,5068,3235,1694,3791,2514,2642,5697	10	100	100	1
3011,4952,3304,3235,5068	5	100	100	1

表 6.3 SVM RRFA 算法在 Colon 数据集的实验结果

Table 6.3 the experimental results of SVM RRFA on Colon data

所选特征	特征个数	训练集正确率	测试集正确率	C
1423,792,765,493,1924,1668,377,1482,1597,350,1570,611,654,1548,1866,856,186,665,875,918,1348,1096,1022,1773,472,401,1164,822,1386	29	96.72131	100	5
974,1769,1346,1580,1772,1740,1024,1641,43,353,590,522,698,1743,1729,815,1521,325,138,405,107,1792,1069,385	24	96.72131	100	3
1769,43,1346,974,1772,1953,504,1798,1378,1110,1042,627,384,1549,1221,987,1954,1048,513,1579,780,1260,1067,234,1582,1887,1534,1808,1601,1727,1366,1098,1168,1867,621,1914,1068,1718	38	95.08197	0	1
1769,175,1740,799,1346,1870,974,353,187,682,590,1823,1177,1885,698,1622,601,1584,685,427,31,627,399,1806,161,1123,1048,795,513,1562,1886	31	96.72131	100	3
974,47,1772,341,1823,1769,175,353,1325,1870,118,1743,1466,698,1993,522,1729,1177,915,1531,1587,1697,601,576,491,1102,1970,539,1732,724,1308,285,1254,1804,767,1312,413,1305,500,683,1966,313,246,1082,421	45	96.72131	100	1
1346,1772,974,1769,341,1740,175,1870,353,590,1743,1024,1466,1993,1221,625,1799,1546,1334,1395,1797,657,1107,401,227,717,178,338,619,1605,1236,467,1397	33	98.36066	100	5
974,1641,1769,43,1772,1740,11,1953,353,590,522,1325,1743,187,698,279,1993,1177,1370,682,427,915,306,223,1584,118,1608,1920,1954,1891,652,1048,234,151,727,1395,1334,1579,1067,1064,1797,734,1680,338,1107,982,1864,1856,640,1741,197,1159,1755,1084	54	95.08197	100	1

974,1772,1921,1769,11,516,175,590,698,1743,1024,187,1466,1823,43,915,1993,1042,627,682,1916,1697,1226,1521,597,1812,72,1896,1083,26,405,1030,1334,1546,1147,1189,734,1155,727	39	96.72131	100	1
1772,1740,974,43,1870,1769,1042,1916,682,1697,427,118,1370,1582,1634,763,188,1799,1145,1067,444,1239,388,1406,1405,964,702,992,1826	29	98.36066	100	3
974,43,1346,1769,1772,175,1641,353,1953,1325,590,698,1564,1743,1729,187,522,279,1177,1806,1042,1697,118,682,576,1896,380,1339,597,653,138,234,1608,1920,652,1023,1226,304,515,325,1654	41	96.72131	100	3
974,43,1346,1772,1769,590,279,1743,522,915,187,1564,682,306,1916,1042,1370,164,1859,118,427,1771,1757,815,1102,1872	26	95.08197	100	13
974,1772,43,1346,281,1859,915,1584,1435,1823,590,682,1177,1042,1697,1353,427,1419,764,1798,752,878,384,627,1110,1370,1920,515,1339,1896,552,1372,657,1582,159,1239,1624,62,1614,391,513,1366,226,1444,1616	45	98.36066	0	7
974,1769,175,1772,1346,1641,353,11,187,590,522,516,1870,43,1024,1993,915,427,682,1370,1564,1697,1177,118,534,1473,153,1771,661,23,1562,987,72,527,1521,652,1896,1339,835,380,1083	41	96.72131	100	7
974,1769,1772,1740,1346,43,1325,698,1743,522,187,1024,1466,427,1606,306,1993,1440,1564,390,534,580,1110,627,1916,1042,1859,1370,955,118,164,1697,597,107,1339,72,405,987,1030,380,385,1896,527,1812,1747,652,1521,585,1379,1837,1822,343,1867,1079	54	100	100	5
1769,175,974,1772,1641,1580,1740,341,1870,11,516,627,43,1370,1466,1743,1325,955,698,1798,682,1970,576,338,1916,1549,1042,1440,1993,31,635,1256,1972,1098,1472	35	100	100	5
974,47,1769,1641,43,1325,1729,915,1584,279,187,2000,306,1110,1896,627,576,752,764,26,1747,1226,1886,1792,304,127,1954,542	28	95.08197	0	7
1772,974,1769,1466,1641,175,516,1953,1743,1870,467,1798,43,1729,1916,1110,306,187,1584,1177,1954,1872,1196,72,1970,1549,1030,1048,1896,764,31,1102,1313,727,771	35	95.08197	100	3
1772,1740,974,43,1870,1769,590,1584,1743,698,1221,279,1466,1435,1823,534,1110,955,1353,601,1549,764,878,153,1473,1562,757,72,1889,1352,1313,107,1339,527,138,380,835,1896,385	39	96.72131	100	5
974,1772,1641,1769,11,281,698,306,118,682,427,1859,164,1564,1177,390,601,1606,576,1042,1697,1378,1353,1872,1102,668,1004,1951	28	96.72131	100	11
1772,43,974,1346,1740,341,353,175,1769,1564,1743,698,1584,590,1729,661,1549,1798,23,1771,987,1367,1920,380,405,1083,329,138,1226,527,385,304,1048,1521,1954,515,1797,652,1064,1608,727,1579,780,1812,1217,1546,734,1334,717,639,1534,1098,1060,1808,228,1310,882,1294,467,1558,1851,1472,1650,1170,994,1257	66	100	100	5
1772,974,1769,11,1870,353,1641,175,281,1916,1440,1993,1370,118,427,1177,661,815,1004,1562,653,1473,1771,1889,1313,987,153,72,527,1083,597,1896,1521,1920,504,1608,1249,1068,1942,1447,1420,1430,1103,861,1000,953,507,1652,1290,1305,1734,821	52	100	100	3
974,1769,43,1772,175,1346,590,915,1584,698,1146,976,522,516,1042,1993,682,1564,1353,1916,1370,427,1697,306,661,385,578,72,652,1226,1783,1339,1896,405	34	98.36066	0	13

974,1772,43,1769,1870,175,1641,1357,281,11,187,1564,590,915,1859,1729,1584,698,682,1110,1916,627,1697,1993,427,1920,1030,1608,835,1646,527,1557,1750,1716,1048,625,1799,1693,614,1680,401,1927,657,1476	44	96.72131	100	3
1769,974,11,175,353,1370,1772,1221,915,1346,580,1367,1697,43,1622,1042,1564,1916,47,1473,815,1859,1123,601,1579,62,906,1064,1799,769,1088,780,1155,1365	34	96.72131	100	5
1772,43,974,1740,1346,590,698,1584,1823,915,1466,1622,1042,1697,682,1370,1993,1564,1916,306,427,1859,384,1549,1872,878,1798,1951,1757,1920,1608,234,1023,1226,1747,1521,1088,515,922	39	96.72131	100	11
974,1769,43,1772,1641,11,682,1564,187,698,306,1177,1370,427,1584,764,1042,627,955,1110,1920,1608,1226,1334,153,329,652,1582,625,489,1876,619,1442,1239,3,162,444	37	96.72131	100	3
974,1769,1641,43,1772,175,11,1325,590,698,1743,187,1466,522,1729,1042,1606,1177,1370,390,1757,815,1102,685,595,783,107,31,653,1747,62,380,835,652,265,1030	36	96.72131	100	5
974,1769,43,1772,1641,11,915,1024,1743,1584,590,698,47,187,15,522,682,1564,1440,427,1177,390,1370,1119,164,1479,764,955,1110,1353,752,1042,1606,1798,1806,1226,515,265,578,1747,922,1030,1626,1557	44	96.72131	100	7
1772,43,974,1346,1740,341,1870,2000,281,1769,682,1378,1993,1697,1042,427,164,625,1634,619,489,657,840,1730,1150,778,1405,830,600,1775,57,1109	32	98.36066	100	3
1772,974,1769,1740,11,1870,1221,516,522,43,1325,1346,1024,590,1993,1466,427,279,698,306,187,601,764,1549,1872,878,685,1196,399,384,1771,1757,31,815,1951,1102,653,1534,717,1563,1601,1801,1098,226,251	45	96.72131	100	5
1772,43,974,1740,1346,1584,590,698,915,1466,1743,1823,187,682,1697,1042,1916,1564,427,380,405,1896,265,1339,385,1088,138,1783,1226,329,1920,1747,652,515,1048,513,1579,1797,1067,1546,763,1450,1068,639,1585,1534,1637,1912,464	49	96.72131	100	5
974,1346,43,1769,1772,590,1584,915,1729,1743,1823,522,1798,601,1353,1180,627,955,1747,1226,1920,652,542,1023,1861,539,498,1139,165,686,212	31	96.72131	100	3
974,1769,353,11,1953,281,1870,1740,1221,1772,516,1325,522,153,661,1473,1771,815,72,107,385,1339,835,597,987,405,1048,151,1608,1546,727,1334,1147,468,1064,1068,1957,1299,697,1657,1474,1456,170,74,1428	45	98.36066	100	3
974,1769,353,11,1953,281,1870,1740,1641,175,522,1772,698,187,1325,590,516,1608,1920,922,1747,515,652,329,1954,1579,468,234,1634,1680,1147,780,1155,1067,1582,513,1799,619,1145,1876,1442,625,1927,429,444,489,1366,1407,1037	49	98.36066	100	5
974,43,1346,1769,1772,175,1641,590,1823,915,1729,1743,516,698,1622,1466,1177,1119,427,534,1479,878,1549,1798,764,384,1806,1757,1004,685,399,653,661,1771,1473,23,31,1951,1102,1562,1313,1896,597,1226,527,72,1954,1891,304,542,1334,325,652,1064	54	96.72131	100	5
1772,43,974,1346,1740,341,353,175,1769,1743,590,279,1584,522,698,1466,915,1562,1889,1352,597,651,987,1313,757,288,72,1896,1339,1783,380,138,265,835,513,1799,1579,1067,347,1680,1567,1074,105,1671,258,1637,1436,1599,1590,190,649,1867,1764,1153,1046,959	56	100	100	5
974,1769,1772,1641,11,1953,353,175,1740,590,187,516,1325,1221,1466,522,43,1024,1920,1226,652,380,527,922,1747,1521,329,1891,515,304,1557,1954,1048,1395,1064,1608,1334,62	38	95.08197	100	1
974,1772,1769,1641,11,353,175,1325,1953,590,1743,43,187,47,915,1584,661,653,1473,1771,23,1102,1889,1078,1951,1048,771,1579,734,468,780,1067,1546,325	34	95.08197	100	3

1772,974,43,1346,1769,341,2000,281,1580,1740,698,915,522,590,516,427,1993,1042,118,955,1370,1859,1606,1177,1353,1564,1102,1110,1951,878,153,1435,384,661,1549,987,653,1123,1562,1872,578,1582,1634,1799,619,1395,1226,1650,1727,903,1972,1032,360,1770,1992,1036,530,486,131,1198	60	100	100	3
974,43,1346,1772,1769,175,1641,1953,1743,590,698,279,187,1466,522,1042,1916,1564,627,1606,1110,764,1196,72,380,597,1896,405,661,385,835,527,1521,1891,1954,1608,1395,304,542,1249	40	96.72131	100	5
974,1769,43,175,1772,353,1641,1740,682,1042,427,118,1370,187,1110,1798,627,1564,107,878,1177,1353,815,1697,1916,138,1549,384,601,1102,576,1896,1196,734,1954,683,727,515,1891,987,1886,1305,257,634,282,1631,1050,1448	48	100	0	3
1772,1769,974,1641,175,353,516,1325,1993,1743,1870,522,1470,1221,630,1346,1378,764,1353,1042,1951,1872,1016,1196,384,1110,1921,1521,1859,595,1834	31	98.36066	100	5
974,1769,1772,1346,1580,43,2000,1177,1757,1697,1606,601,391,1534,326,649,780,513,1563,1868,228,1315,679,497	24	98.36066	0	5
974,1769,1772,1641,11,1953,175,590,1743,187,43,522,1466,1729,627,1042,1110,1606,1916,955,878,1798,1549,399,1872,384,815,601,1123	29	96.72131	100	27
974,1772,1325,341,1769,1740,1024,1221,43,522,187,590,815,595,1697,1110,1771,1859,1794,1036,360,1558,1535,1914,1039,1610,287,460,1422,579,1978,1685,1448,1698,1966,1861,454,1576,1487,1410	40	96.72131	0	1
1772,974,1769,341,11,175,1325,1641,281,1870,516,1024,915,1466,43,522,1221,1823,1920,1608,1747,922,578,1886,138,515,1023,527,1048,780,1334,771,1579,1395,1312,197,949,1290,634,1946	40	96.72131	100	1
1772,43,974,1740,1346,1584,590,1823,698,915,1743,279,1177,1729,1221,1042,1697,682,1564,1916,118,1859,427,164,1370,1562,1352,225,1313,680,1929,288,305,659,25,1781,87	37	100	100	3
974,1769,1772,175,11,353,1641,1953,1325,187,590,1743,43,1110,627,1042,1564,1970,601,1606,1916,1353,815,1747,878,1771,1697,1102,1549,26,1646,1806,504,72,1339,597,35,1527,9,709,224,1994,1634,987,380,1896,652,329,835,1232,1886,1605,1048,1546,1579,1680,1693,1334,62,311,1608	61	100	0	7
1870,43,281,974,601,1346,1769,1370,698,1772,1353,1531,405,1466,279,1916,1585,1473,1196,917,399,385,627,1951,1378,661,1920,1334,653,1797,1902,1385,1954,1839,1826,1679,1732,1236,1485,1651,1046	41	98.36066	0	3
974,1346,1772,43,1769,341,764,1110,682,1993,1564,427,306,1177,534,118,1697,1378,1042,1916,1353,1859,1123,1473,384,504,1008,138,1102,1920,1896,1646,771,987,1954,31,23,835	38	96.72131	0	7
1769,1743,279,974,1870,1353,1346,43,1325,1772,187,630,1697,682,915,601,1951,1521,1466,1916,390,306,1806,955,595,1102,763,1703,1083,752	30	96.72131	0	7
974,1772,1346,43,1769,341,1740,590,698,1584,1743,1729,522,279,661,1473,1771,31,1004,1078,653,1951,1102,72,1562,1896,987,597,405,1048,1891,1608,1395,1334,1966,486,1216,503,131,74,90,305,1456,218	44	98.36066	100	3
1772,974,1769,1325,11,1024,1221,522,1622,187,915,1042,1606,601,1110,627,1747,763,1425,652,515,329,1521,1557,304,1226,1048,504,734,1334,1954,804,727,1147,1064,1891,1563,733,1444,1200,567,213,1661,1903,1899,270	46	96.72131	100	1
974,1769,175,1346,1641,1772,353,590,698,187,1743,522,1042,1177,682,118,1564,1993,427,1562,26,1352,1473,1646,1889,1113,72,288,1896,107,138,265,835,1339,552,380,385,625,1407,489,733,619,1820,943,388,3	46	98.36066	100	7

1772,1769,43,590,974,353,764,1757,1859,1549,187,698,516,878,685,1916,1123,581,1606,1745,1073,1048,661,1196,578,915,15,1102,668,390,1473,461,580,1050,1993,1964,1666,1370,1783,162,1562,1030,1998,277,1927,347,1226,1200,1342,820,1079,1136,75,1464,614,1297	56	98.36066	0	5
341,1024,1870,1772,353,1346,764,590,47,1177,627,1951,1068,347,2000,1916,187,23,1151,1996,1579,1549,1783,1115,1355,1837,1298,1748,520,550,1276,1034,241,197,498,1447,726,730,1405,1240,978,700	42	98.36066	0	1
974,1772,1769,1346,1740,1580,43,1024,187,1325,1916,1743,590,915,1466,516,1584,1993,118,698,1859,427,682,1920,987,1334,835,380,1064,1372,391,1499,1464,1599,1444,1070,1805,1712	38	100	100	3
43,1772,1346,974,1740,341,1769,1870,281,1697,164,1993,118,1378,682,1916,1872,878,1549,1757,384,1951,1367,1078,601,840,1293,1099,778,745	30	96.72131	100	3
974,1772,1769,1641,11,590,698,1024,1729,43,522,1466,1622,1743,1993,427,1177,1370,187,306,223,661,1473,1562,107,680,1889,224,1352	29	95.08197	100	3
974,1769,1641,1772,43,11,353,590,187,1221,1466,1024,1622,1729,1743,915,1346,1564,1993,698,1177,427,1110,682,627,1042,1697,1916,1353,1562,1889,1352,23,680,987,1313,1113,225,1896,653,527,1521,380,1891,1608,1064,734,504,771,1954,1048,1334,467,1168	54	98.36066	100	5
1772,974,43,1769,353,1740,1325,11,341,1024,1798,1916,601,427,1042,306,1564,764,1048,1920,405,304,652,1030	24	98.36066	0	7
974,1743,1769,1870,1641,353,1772,1466,1872,279,915,427,187,1729,522,2000,590,698,1916,764,399,1042,1564,627,1177,1110,1798,783,576,1580	30	96.72131	0	19

表 6.4 SVM RRFA 算法在 Adenoma 数据集的混淆矩阵

Table 6.4 the confusion matrix of SVM RRFA algorithm on Adenoma data

Adenoma		**Prediction outcome**	
		Patient (Positive)	Normal people (Negative)
True value	Patient (Positive)	4	0
	Normal people (Negative)	0	4

表 6.5 SVM RRFA 算法在 Carcinoma 数据集的混淆矩阵

Table 6.5 the confusion matrix of SVM RRFA algorithm on Carcinoma data

Carcinoma		**Prediction outcome**	
		Patient (Positive)	Normal people (Negative)
True value	Patient (Positive)	18	0
	Normal people (Negative)	1	17

表 6.6 SVM RRFA 算法在 Colon 数据集的混淆矩阵

Table 6.6 the confusion matrix of SVM RRFA algorithm on Colon data

Colon		**Prediction outcome**	
		Patient (Positive)	Normal people (Negative)
True value	Patient (Positive)	34	6
	Normal people (Negative)	9	13

表 6.7 简化 SVM RRFA 算法的 Adenoma 数据集的实验结果
Table 6.7 the experimental results of simple SVM RRFA on Adenoma data

所选特征	特征个数	训练集正确率	测试集正确率	C
3886,3127,2910,3887,2510	5	100	100	1
3886,3127,2910,2510,3885,3890,3887	7	100	100	1
3886,3127,2910,3887,2510,3885,3096,2746	8	100	100	1
3886,3127,2910,2510,3887,3885,3093,3890	8	100	100	1
3886,3127,2910,2510,3887,3885,3093	7	100	100	1
3886,3127,2910,2510,3885,3887,3093,2735,3890	9	100	100	1
3886,2910,3127,3887,2510,3885,3093	7	100	100	1
3886,3127,2910,2510,3887	5	100	100	1

表 6.8 简化 SVM RRFA 算法的 Carcinoma 数据集的实验结果
Table 6.8 the experimental results of simple SVM RRFA on Carcinoma data

所选特征	特征个数	训练集正确率	测试集正确率	C
3011,4952,3791,3304,1694,3235	6	100	100	1
3011,3235,1694,4952,3790,373,3304	7	100	100	1
3011,4952,3304,5068,3235,1694,2514,2503	8	100	100	1
4952,3011,373,3304,5203,3235	6	100	100	1
3011,4952,3304,5068,5697,2514,3235,2642	8	100	100	1
4952,5068,3011,3304,3791	5	100	100	1
3011,3304,3235,4952,2514,1694,2503,5697,4975,3790	10	100	100	1
4952,3011,5068,3304,2514,3235,1694	7	100	100	1
3011,4952,3304,5068,2514,3235,1694,2642,2503,3791	10	100	100	1
3011,4952,5068,3304,3235,1694,2514,2503,5120,3790	10	100	100	1
3011,3304,4952,5068,2514,1694,2503,3235,5120,3791	10	100	100	1
3011,4952,3304,5068,1694,2514,3235,2642,2503	9	100	100	1
3011,4952,3791,3304,1694,3235,2503,5120	8	100	100	1
3011,4952,5068,3304,3791,3235	6	100	100	1
3011,4952,3304,5068,2514,1694,3235	7	100	100	1
3011,4952,3304,5068,2514	5	100	100	1
3011,4952,3304,5068,2514,1694,3235,2503,2642	9	100	100	1
3011,4952,3304,5068,1694,2514,3235,2642,2503	9	100	100	1
3011,4952,3304,5068,3235,1694,2514	7	100	100	1
3011,4952,5068,3304,3791,3235,1694,2503,2514,5120	10	100	100	1
3011,4952,3791,2514,5120,3304	6	100	100	1
3011,4952,5068,3304,3235,1694,2514,2503,5120,3790	10	100	100	1
4952,3011,5068,3304,3791,1694	6	100	100	1
5068,3011,3304,5697,1705,2514,5203,3473,4963	9	100	0	1
3011,4952,3304,5068,3235,1694,2514,2503,5120,5697	10	100	100	1
3011,4952,3304,5068,3235	5	100	100	1
3011,4952,5068,3304,3235	5	100	100	1
3011,4952,3791,3304,1694,3235,2503,5120,2514	9	100	100	1

3011,4952,3304,5068,2514,3235,1694,2503,2642,5203	10	100	100	1
4952,3011,1694,3304,2503	5	100	100	1
3011,4952,5068,3304,3235	5	100	100	1
3011,4952,3304,5068,3235,1694,2514	7	100	100	1
3011,4952,3304,5068,2514,3235	6	100	100	1
3011,4952,3304,5068,3235,1694,2514,2503,5120,5697	10	100	100	1
3011,4952,3304,5068,3235,1694,3791,2514,2642,5697	10	100	100	1
3011,4952,3304,3235,5068	5	100	100	1

表 6.9 简化 SVM RRFA 算法的 Colon 数据集的实验结果

Table 6.9 the experimental results of simple SVM RRFA on Colon data

所选特征	特征个数	训练集正确率	测试集正确率	C
1423，792，765，377，1668，1924，493，1976，510，1231，1570，622，1613，1027，1885，663，1981，721	18	96.72131	100	9
974，1769，1346，1580，1772，1740，1024，1641，43，590，1729，1466，522，187，698，1221，1743，516，1440，159，835，652，922，515，468，1557，1883，1608，188，1920，1334，1747	32	96.72131	100	3
974，1769，1772，43，1953，353，1870，281，1357，1740，698，1346，590，1743，1993，1564，341，1466，513，1221，1954，1579，380，1079，1634，1067，780，151，224，906，1145，178，1970，1393，943，162，543	37	96.72131	0	7
1772，1769，974，353，1953，187，1622，682，698，1823，1729，516，1798，504，1993，1042，1353，1226，601，1584，31，1221，1806，627，1562，467，1372，1896，1777，513，1799，882，1200，1360，987，625，1079，1951，1599，489，1145，1153，1365，197，477，1825，1724，1642，1134，790，686，21，1077，153，421	55	100	100	5
974，1772，1823，175，341，1769，47，353，1870，1325，1743，1993，698，1466，1346，522，43，1729，661，1249，1042，578，1889，1828，1812，1798，630，1473，1226，527，1110，384，399，72，1752，1747，1716，1817，75，1915，965，1036	42	96.72131	100	1
974，1769，1346，1772，43，627，1916，682，1697，955，601，1370，1110，1042，1353，1004，380，504，1048，1634，153，304，727，1546，1145，552，1886，1239，1013，1703，1891，1412，1730，649，1900，950，1227	37	96.72131	100	3
974，1769，1346，1641，43，1772，175，516，353，187，1743，279，590，1823，915，1622，955，1110，1798，1549，764，601，878，769，752，384，1562，23，987，651，1352，894，107，1313，288，1889，1685	37	98.36066	100	3
974，1772，1769，175，11，1921，516，1870，1325，1584，1743，698，590，1024，187，1580，915，1587，1110，955，601，576，1806，764，380，1896，72，1226，1521，527，1783，835，1083，1339	34	96.72131	100	3
974，43，1346，1769，1772，590，1584，915，1729，698，1743，1823，682，1177，1993，1370，390，1110，1042，1916，1564，118，1606，1697，627，164，661，72，1562，1896，597，1634，1582，1239，625，619，489，683，1693，162	40	95.08197	100	5
974，1772，1769，1641，11，353，1325，1740，175，281，590，187，1743，698，522，43，1466，1221，427，1370，306，	47	98.36066	100	5

1993, 1564, 2000, 534, 223, 1119, 1110, 955, 1042, 627, 1798, 682, 1916, 1353, 1806, 576, 1920, 1226, 1896, 380, 835, 1783, 1339, 385, 527, 138				
1772, 1740, 974, 43, 1870, 1769, 353, 281, 2000, 1697, 1042, 1916, 682, 118, 427, 164, 1606, 1757, 661, 1771, 1102, 815, 685, 1562, 1473, 23, 153, 1896, 405, 72, 1339, 138, 1792, 1030, 987	35	96.72131	100	11
974, 1772, 1346, 43, 281, 1993, 1221, 915, 1823, 1622, 1791, 187, 682, 1584, 1177, 1042, 1353, 427, 601, 1697, 223, 1859, 1798, 1196, 627, 384, 399, 161, 576, 1606, 595, 752, 1608, 1226, 652, 304, 329, 1783, 1048, 1954, 763, 1546, 1797, 1605, 734, 1067, 1334, 1579, 347, 356, 780, 1680, 1724, 412, 1785, 1055, 1187	57	98.36066	0	5
1346, 1772, 43, 974, 1769, 1740, 281, 341, 1325, 353, 1993, 427, 1370, 118, 1916, 1479, 682, 164, 534, 1110, 601, 1806, 752, 1353, 1042, 1697, 1549, 878, 576, 1634, 234, 1799, 780, 1067, 1680, 468, 789, 1794, 141, 1887, 1360	41	98.36066	100	3
974, 1769, 1772, 1346, 43, 1740, 1325, 1024, 1729, 1743, 1221, 915, 1622, 878, 31, 1549, 601, 1951, 685, 153, 661, 1473, 815, 652, 1889, 1562, 72, 1771, 107, 1048, 1395, 277, 727, 311, 1388, 804, 771, 1922, 1249, 1366, 1582, 429, 1927, 1372	44	98.36066	100	7
1769, 1024, 1772, 43, 974, 1870, 1740, 955, 1798, 1916, 682, 1196, 1110, 1042, 1353, 698, 752, 1771, 1562, 399, 1889, 1473, 1634, 1872, 1605, 138, 1606, 347, 23, 329, 1896, 1799, 467, 1315, 717, 552, 542, 763, 1671, 1608, 177, 131, 1947, 509, 797, 387, 1811, 1988, 1216, 212	50	98.36066	100	3
1772, 974, 11, 281, 1769, 1641, 175, 1870, 1325, 590, 516, 1823, 47, 43, 1743, 1177, 580, 1370, 427, 764, 1042, 661, 468, 1102, 1123, 1896, 161, 26, 1872, 1478, 1405, 1732	32	98.36066	100	5
974, 1772, 1769, 175, 1346, 1859, 590, 467, 1993, 915, 1584, 1697, 1729, 187, 576, 1798, 1353, 1024, 955, 1608, 878, 1042, 1030, 595, 72, 1473, 627, 1634, 1579, 734, 1680, 159	32	95.08197	100	5
974, 1769, 1772, 1641, 11, 353, 1953, 1740, 175, 187, 43, 1325, 522, 1221, 516, 1024, 590, 1743, 1370, 427, 1993, 1564, 2000, 306, 1110, 682, 1353, 1798, 1042, 627, 1606, 1180, 1196, 1559, 982, 1447	36	100	100	5
1769, 974, 2000, 175, 1580, 1953, 187, 590, 279, 1743, 915, 698, 1110, 601, 627, 955, 878, 384, 1549, 1608, 1954, 1083, 26, 31, 1634, 304, 338, 1797, 265	29	95.08197	100	7
1772, 43, 974, 1346, 1740, 1743, 522, 590, 1823, 1564, 915, 1584, 1729, 187, 1110, 601, 384, 955, 399, 1587, 1549, 161, 627, 1771, 661, 653, 1798, 1102, 815, 1473, 72, 1562, 1004, 1886, 1896, 138, 527, 325, 1339, 835, 107, 1030, 1083, 385	44	98.36066	100	17
974, 1769, 43, 1772, 2000, 1641, 175, 1346, 1953, 1870, 668, 1110, 1872, 1367, 1353, 1606, 1771, 1473, 1951, 31, 1757, 661, 527, 72, 1226, 1562, 1339, 1896, 1030, 597, 1312, 1211, 285, 697, 1351	35	98.36066	100	3
974, 1769, 1953, 11, 353, 1870, 590, 1325, 1772, 43, 522,	48	98.36066	100	5

698, 47, 1221, 1743, 1584, 1993, 682, 1916, 1042, 1697, 1564, 427, 118, 1110, 601, 764, 752, 1806, 576, 1196, 955, 1872, 31, 1549, 1951, 1521, 652, 1747, 1425, 1064, 1608, 1868, 1414, 989, 1065, 1385, 1914				
1870, 974, 1772, 1769, 175, 187, 915, 1370, 522, 1859, 1729, 427, 1564, 1221, 1743, 1473, 1798, 304, 1334, 356, 734, 489, 1145, 72, 1747, 384, 1048, 1799, 625, 347, 1546, 1579, 515	33	96.72131	100	5
974, 1769, 1772, 175, 11, 1870, 281, 1614, 915, 1584, 1221, 43, 1370, 1346, 164, 1697, 1743, 590, 1353, 1470, 1440, 1325, 1042, 1916, 1872, 1378, 1921, 601, 1798, 329, 1531, 679, 1608, 1970, 652, 1067, 1562, 1812, 1568, 1541, 1196, 1875, 1616, 31, 1226, 1799, 1680, 668, 1935, 384, 780, 1797, 1703, 1309, 1603, 503, 305	57	100	100	3
974, 43, 1346, 1772, 1769, 175, 1641, 1953, 11, 590, 1743, 915, 1584, 1729, 279, 187, 698, 976, 522, 1110, 601, 1798, 627, 955, 384, 399, 1606, 1353, 1102, 1549, 1872, 1951, 1004, 685, 815, 1367, 1078, 1757, 1896, 1521, 1608, 72, 380, 1812, 1339, 405, 385, 1783, 1747, 1920, 1226, 325, 922, 542, 515, 1891, 1954, 504, 652, 468	60	96.72131	100	5
974, 43, 1346, 1769, 1772, 175, 590, 915, 1823, 1729, 698, 1221, 1177, 118, 427, 1993, 1119, 1435, 682, 955, 1042, 1564, 1110, 1549, 399, 601, 764, 1970, 1757, 878, 661, 1562, 597, 1896, 26, 987, 1339, 757, 1313, 1646, 1954, 1064, 1920, 515, 151, 304, 1425, 329, 1521, 652	50	96.72131	100	5
1769, 1772, 974, 1641, 11, 175, 682, 1110, 1196, 1806, 1916, 1606, 627, 769, 1872, 955, 1004, 685, 764, 1549, 1564, 815, 1473, 153, 23, 1008, 1562, 107, 1771, 661, 652, 234, 1339, 515, 1634, 795, 1582, 1799, 1366, 1200	40	98.36066	100	5
1772, 974, 1769, 1641, 11, 1325, 175, 281, 353, 1743, 1584, 187, 915, 590, 1042, 1110, 955, 601, 1606, 1806, 1521, 380, 1896, 1812, 1920, 1048, 734, 1546, 1797, 1334, 780	31	95.08197	100	3
974, 1769, 43, 1772, 1641, 11, 915, 1743, 1024, 590, 1584, 698, 522, 1042, 682, 1697, 1564, 1916, 118, 1440, 1177, 427, 1859, 1798, 31, 1549, 1757, 1771, 1492, 1389, 1733	31	96.72131	100	3
1769, 974, 43, 1772, 1641, 175, 2000, 1580, 47, 1870, 1697, 682, 1042, 1916, 427, 1564, 1177, 1370, 118, 1440, 504, 1064, 1048, 304, 1891, 151, 1334, 329, 1521, 1546, 1966, 1084, 93, 505, 1342, 1381, 1592, 634, 1856, 1274, 530, 421, 503, 1219, 615	45	100	100	3
974, 1346, 1769, 43, 1772, 175, 1740, 1221, 590, 1743, 915, 516, 698, 1531, 1823, 187, 1042, 682, 427, 1993, 1370, 1564, 1110, 955, 601, 627, 764, 752, 1806, 1606, 661, 1473, 153, 23, 1771, 1634, 1799, 1605, 1067, 763, 460, 1351, 1950, 298, 1112	45	98.36066	100	3
974, 1769, 43, 1772, 1641, 1743, 915, 590, 1584, 1024, 522, 1729, 187, 279, 597, 1896, 72, 380, 527, 835, 1783, 138	22	95.08197	100	7
974, 43, 1772, 1769, 1641, 698, 1743, 590, 1622, 1729, 915, 1042, 682, 1564, 1993, 427, 1370, 1353, 1916, 1549, 1872, 31, 1951, 1102, 685, 815, 1078, 527, 72, 1339, 835, 597, 1896, 1030, 380, 405, 385, 1747, 1608, 1920, 652,	43	96.72131	100	5

542，1521				
974，1772，1769，1641，11，590，698，1729，1743，1024，1747，1920，1954，1023，1886，1226，1521，515，652，304，62，1579，347，1680，1145，1107，234，1054，250，1969，1822，1862，147，713，447	35	96.72131	100	3
974，1772，1769，1641，11，175，281，1325，187，682，1564，427，1370，1993，1584，306，1177，1896，527，138，1339，597，652，385，1562，391，1372，1815，1013，429，1718，656，1366，1260，1298，1607，1355，659，352，88，463，940	42	95.08197	100	1
974，1769，1772，1641，11，353，1953，590，1325，1743，43，522，187，698，1466，915，987，1562，1352，680，107，1343，651，894，1313，1048，1891，1634，727，1147，780，734，1395，234，62	35	96.72131	100	7
974，43，1346，1769，1772，175，1641，353，1953，1325，590，1743，1823，915，698，516，1466，1042，682，1993，306，1370，1564，1608，1747，1954，1226，1920，1557，652，265，946	32	98.36066	100	3
974，1769，1953，353，11，281，1641，1870，1740，522，590，1325，1772，698，187，43，1743，1466，516，1993，427，1042，1370，1916，682，1697，118，1353，380，1896，653，405，527，1226，265，835，72	37	98.36066	100	11
974，1346，1772，43，1769，1177，955，1606，769，1042，1110，764，1798，1697，661，815，399，347，1646，601，1562，1771，1249，682，1954，138，1527，727，304，1455，527，1582，1927，1634，1799，1048，1820	37	96.72131	100	5
974，1772，1769，1641，11，175，590，1024，915，43，1466，1743，1221，47，682，1564，1177，427，187，1697，661，1473，1771，1004，815，653，1562，31，1646，1951，1226，1920，1747，325，922，1886，1088，585，265，1023	40	96.72131	100	5
974，1769，175，1641，2000，353，682，590，187，1370，1729，223，1042，1110，627，698，522，1177，341，107，1872，1916，815，1008，399，138，1564，1549，595，1680，882，1797，1334，1013	34	93.44262	100	1
974，1769，1772，1346，1870，1378，1367，1951，1042，764，427，405，1587，1549，1606，1110，399，955，1353，1102，1896，680，627，1521，1798，380，338，384，601，1435，987，1771，734，1473，1455，515，304，288，1088，356，445，1365，640，1492，1712，1208，1696	47	98.36066	0	3
974，1769，1772，1346，1580，43，1729，590，698，915，1743，1564，1177，1606，1042，1993，1757，1419，118，1697，682，164，1806，1110，1473，399，955，685，661，625，1747，1067，1023，1632，107，1886，489，578，356，1730，1761，1822，437，1540，520，1372，1699，1298，1486，1406，1355	51	98.36066	100	5
974，1769，353，11，1953，281，1870，1740，1641，1370，427，1466，187，682，1372，518，1952，1013，1541，1616	20	95.08197	100	3
1772，974，1325，1769，43，1466，1729，522，1221，1584，630，1743，187，15，1357，1916，1771，878，1110，1102，661，153，72，1521，955，1562，1123，1799，1582，1068，1145，1226，1579，1107，356，1165	36	96.72131	0	1
974，1772，1346，43，1769，590，1584，915，1729，1823，1743，516，522，698，187，661，153，1473，23，653，815，72，1562，597，1896，1339，26，385，987，835，1920，527，	36	95.08197	100	9

234, 380, 1083, 515				
974, 1772, 1346, 43, 1769, 341, 1740, 353, 281, 698, 1743, 590, 187, 1466, 279, 1221, 915, 1110, 1353, 955, 627, 1606, 1042, 1180, 576, 1916, 1370, 1634, 513, 619, 1680, 1145, 1605, 1911, 1239, 162, 1312, 559, 1655, 197, 1084, 982, 505, 854, 857, 1488, 1702, 1445, 50	49	96.72131	100	1
974, 1769, 43, 1772, 1641, 516, 1740, 590, 1325, 1584, 698, 1622, 915, 1729, 187, 1743, 1353, 1747, 1916, 1859, 399, 752, 1549, 1634, 987, 504, 72, 1783, 1339, 385, 1920, 1608, 1896, 1954, 329, 515, 652, 1605, 265, 1557	40	98.36066	100	5
974, 1870, 1346, 281, 1370, 1769, 516, 1772, 43, 601, 1466, 522, 175, 1993, 1771, 1042, 1606, 527, 1470, 1315, 1473, 468, 1872, 627, 1239, 752, 1671, 1632, 62, 1912, 1168, 639, 1954, 1902, 958, 1839, 227, 1397, 649	39	98.36066	0	5
1346, 1772, 974, 43, 1769, 1740, 1584, 1743, 915, 279, 1993, 698, 187, 590, 1622, 682, 1697, 427, 1123, 1378, 1896, 23, 31, 1226, 597, 1646, 1313, 1994, 1527, 1799, 1747, 1030, 1557, 652, 1088, 107, 513, 1048, 1147, 1546, 1521, 1334, 1478, 1485, 1121, 1447, 1831, 1569, 1741, 787, 1690, 331, 1264, 1505, 921, 1166, 477, 1474, 59	59	100	100	5
974, 1769, 1743, 43, 799, 279, 1641, 175, 281, 1325, 1772, 1353, 1859, 1697, 47, 915, 682, 976, 187, 1521, 1466, 601, 240, 31, 1470, 390, 118, 1370, 1798, 1899, 1123, 138, 1042, 1951, 783, 878, 1378, 1673, 1634, 1920, 1771, 715, 1473, 325, 542, 1237, 1491, 1249, 486, 1738, 562, 1296, 305	53	98.36066	0	3
1772, 43, 974, 1346, 1740, 341, 1769, 353, 1870, 175, 590, 1743, 522, 279, 1823, 187, 698, 915, 1584, 1466, 1110, 1872, 764, 1549, 685, 1876, 1582, 625, 1239, 489, 1366, 1927, 657, 468, 1200, 1904, 1541, 444	38	96.72131	100	5
1769, 1772, 974, 1641, 11, 1325, 590, 1729, 187, 1622, 1221, 43, 1024, 1564, 1177, 1584, 15, 427, 153, 653, 1473, 987, 1771, 661, 815, 31, 26, 1608, 1896, 578, 380, 72, 138, 1567, 795, 1808, 780, 1563	38	96.72131	0	5
1346, 974, 1772, 43, 341, 1769, 1953, 590, 915, 1584, 1743, 1729, 516, 1177, 1042, 682, 1993, 1859, 118, 601, 1564, 1771, 815, 1102, 661, 1004, 1757, 1078, 1970, 1752, 31, 380, 527, 1083, 138, 405, 385, 652, 1654, 597, 653	41	98.36066	0	17
43, 1769, 590, 769, 974, 341, 353, 11, 1580, 1641, 1549, 1772, 1757, 764, 1859, 1870, 187, 1325, 698, 516, 26, 795, 47, 1771, 228, 955, 682, 1797, 306, 522, 1226, 882, 789, 1083, 304, 639, 1605, 385, 1238, 1699, 1424, 1965, 192, 1034, 945, 1663, 1723, 1108, 1098, 1752, 1710, 1793, 595	53	100	0	3
1024, 974, 1325, 1772, 1993, 1870, 341, 353, 1641, 764, 1346, 47, 1177, 590, 1196, 1798, 1634, 1747, 107, 504, 1568, 399, 23, 1549, 1996, 1830, 1927, 427, 838, 1896, 1823, 226, 72, 795, 1983, 380, 138, 661, 1582, 1054, 1730, 964, 1039, 1868, 785, 1648, 1319, 447	48	98.36066	0	3
1772, 1769, 974, 1641, 1740, 1580, 1953, 353, 1870, 522, 516, 43, 1325, 1531, 1024, 427, 1859, 698, 1916, 2000, 1606, 1221, 682, 1470, 467, 943, 1013, 1952, 700, 1176, 1037, 1820	32	96.72131	100	1

1772，43，974，1740，1346，1870，2000，281，590，1859，1584，1743，698，915，1729，1177，47，1697，118，1110，576，1353，427，682，306，1564，1757，601，878，1367，1951，1771，1102，1406，1478，964，1575，1355，830，206	40	100	100	3
974，1772，1769，1641，11，353，1325，175，1740，1110，601，1042，627，1806，380，527，1088，265，138，922	20	96.72131	100	15
974，1769，1641，1346，175，187，590，1743，698，522，1823，279，1042，1564，1177，1697，682，118，1828，427，31，1102，1872，1757，661，1367，1004，1473，153，1771，815，1562，23，26，987，513，3，1634，1799，1067，338，1145，619，1680，444，1324，1932，1803，1235，861，1396	51	98.36066	100	3
1772，974，43，1769，353，1740，1325，1024，11，341，698，1743，187，1993，915，522，1466，47，1353，1697，1123，627，26，1196，1606，1587，752，661，1110，878，1473，685，1757，1078，31，1102，595，1799，771，1579，1147，1334，1680，325，1886，1693，1068	47	96.72131	0	3
974，1772，353，1743，1346，1466，1641，799，1870，1872，1916，627，1177，1564，1042，1709，752，1110，1798，1580，576，1102，1473，601，1994，955，1196，1771，515，1671，513，625，489，1200，733，1238，444，1237，391，679，833，1875，1407，1718，1632，961，647，702，466，323，1958，1355	52	98.36066	100	3

表 6.10 简化 SVM RRFA 算法在 Adenoma 数据集的混淆矩阵

Table 6.10 the confusion matrix of simple SVM RRFA algorithm on Adenoma data

Adenoma		**Prediction outcome**	
		Patient (Positive)	Normal people (Negative)
True value	Patient (Positive)	4	0
	Normal people (Negative)	0	4

表 6.11 简化 SVM RRFA 算法在 Carcinoma 数据集的混淆矩阵

Table 6.11 the confusion matrix of simple SVM RRFA algorithm on Carcinoma data

Carcinoma		**Prediction outcome**	
		Patient (Positive)	Normal people (Negative)
True value	Patient (Positive)	18	0
	Normal people (Negative)	1	17

表 6.12 简化 SVM RRFA 算法在 Colon 数据集的混淆矩阵

Table 6.12 the confusion matrix of simple SVM RRFA algorithm on Colon data

Colon		**Prediction outcome**	
		Patient (Positive)	Normal people (Negative)
True value	Patient (Positive)	34	6
	Normal people (Negative)	5	17

从表 6.1 所示 SVM RRFA 基因选择算法在 Adenoma 数据集上的实验结果可见，SVM RRFA 算法在该数据集上达到了非常好的结果，训练集、测试集的正确率均为 100%。选择的特征子集的平均规模为 6.875，不足 7 个，是原始基因数 7086 的千分之一弱。留一法实验所选择的共有特征为 4 个特征，分别是第 3886，3127，2910 和第 2510 个特征。

表 6.2 的实验结果显示，对于 Carcinoma 数据集，SVM RRFA 算法的训练正确率为 100％，留一法实验的测试结果有一个正常样本被误分为癌症样本。从实验结果分析可得：第 3011 和第 3304 个基因是共有基因，即所有特征子集的交集，是正确实现分类的两个最重要区分基因。从表 6.2 的实验结果还可看出，第 4952 个基因也很重要，因为除了被误诊的样本，其他任一样本作为测试样本时，所选择的特征子集都含有第 4952 个基因。参照数据集描述，可知第 3011 个基因对应"Human ribosomal protein S25 mRNA, complete cds"；第 3304 个基因为"Homo sapiens hnRNP-C like protein mRNA, complete cds"；第 4952 个基因表示"yb55h04.s1 Homo sapiens cDNA clone 75127 3' similar to gb:M16660 HEAT SHOCK PROTEIN HSP 90-BETA (HUMAN) "。实验所选择的平均基因数为 7.6389，不足 8 个，相对与原始数据集的 7457 个基因，大约只占其千分之一，但是平均分类正确率达到了 97.22%。表 6.1 和 6.2 的共同特征是所选择的参数 C 均为 1。

表 6.3 关于 Colon 癌症数据集的实验结果显示：有 15 个样本被错分，平均分类正确率是 $\frac{62-15}{62}\times 100\% = 75.81\%$ 。其中 6 个癌症患者被漏诊，9 个正常人被误诊。平均选择的特征数是 35.8585 个，相对于原始数据集的 2000 个基因特征，被选特征子集的规模不到原始特征集的五十分之一。

从表 6.4 关于 Adenoma 数据集的混淆矩阵可得，SVM RRFA 算法在该数据集的灵敏度（sensitivity），即对癌症患者的诊断准确率，为：$sensitivity = \frac{true\ positive}{true\ positive + false\ negative} = \frac{4}{4+0}\times 100\% = 100\%$；特异性（specificty），即对正常人的诊断正确率，为：$specifity = \frac{true\ negative}{false\ positive + true\ negative} = \frac{4}{0+4}\times 100\% = 100\%$；Matthews 相关系数 MCC（Matthews correlation coefficient）的值达到了最大值 1。MCC 的计算公式如下：$MCC = \frac{true\ positive \times true\ negative - false\ positive \times false\ negative}{\sqrt{PNP'N'}}$，其中 $P = true\ positive + false\ negative$，$P' = true\ positive + false\ positive$，$N = true\ negative + false\ positive$，$N' = true\ negative + false\ negative$。Matthews 相关系数的最优值为 1，最差值为－1，分别代表了分类器的全部分类正确和全部分类错误两种极端状态。通常情况下 $-1 \le MCC \le 1$，代表了分类器的分类性能，显然其值约接近 1，表示分类器的性能约好。

从表 6.5 的 Carcinoma 数据集混淆矩阵可得，SVM RRFA 算法在该数据集的灵敏度 $sensitivity = \frac{true\ positive}{true\ positive + false\ negative} = \frac{18}{18+0} \times 100\% = 100\%$，特异性 $specifity = \frac{true\ negative}{false\ positive + true\ negative} = \frac{17}{1+17} \times 100\% = 94.44\%$，Matthews 相关系数 $MCC = \frac{18 \times 17 - 1 \times 0}{\sqrt{(18+1) \times (18+0) \times (17+1) \times (17+0)}} = \frac{18 \times 17 - 1 \times 0}{\sqrt{19 \times 18 \times 18 \times 17}} = 0.9459$，接近最优值 1。

从表 6.6 关于结肠癌数据集 Colon 的混淆矩阵可得，SVM RRFA 算法在该数据集的灵敏度、特异性和 Matthews 相关系数分别为：$sensitivity = \frac{34}{34+6} \times 100\% = 85\%$，

$specifity = \frac{13}{9+13} \times 100\% = 59.1\%$，

$MCC = \frac{34 \times 13 - 9 \times 6}{\sqrt{(34+9) \times (34+6) \times (13+9) \times (13+6)}} = \frac{34 \times 13 - 9 \times 6}{\sqrt{43 \times 40 \times 22 \times 19}} = 0.4576$，不到 0.5。但是，SVM RRFA 算法在 Colon 数据集的灵敏度高大 85%。在一个分类器不能使灵敏度和特异度都优的情况下，我们更希望它的灵敏度高，这样不会漏诊真正的癌症患者。而且特异度低只是将正常人误判为癌症患者，这样的损失显然比漏诊小。

以上关于 SVM RRFA 基因选择算法在 3 个基因数据集的实验结果的详细分析表明：本章提出的 SVM RRFA 基因选择算法是一种有效的基因选择算法。从 3 个基因数据集的实验结果可以看出，SVM RRFA 算法对于前两个基因数据集的性能更好，对于结肠癌数据集的分类性能和所选择基因子集的规模相比前者略弱。

从表 6.7 所示的简化 SVM RRFA 基因选择算法在 Adenoma 数据集上的实验结果可见，简化 SVM RRFA 算法在该数据集上同样达到了非常好的结果，训练集、测试集的正确率均为 100%。选择的特征子集的平均规模为 7，大约是原始基因数 7086 的千分之一。实验所选择的共有特征为 5 个特征，分别是第 3886，3127，2910，3887，和第 2510 个特征。相比与本章提出 SVM RRFA 算法，简化 SVM RRFA 算法的共有特征多了一个，即第 3887 个基因；所选择的特征子集的平均规模也略大。

表 6.8 显示：简化 SVM RRFA 算法与本章提出的 SVM RRFA 算法在 Carcinoma 数据集上取得了相同的实验效果。

表 6.9 关于简化 SVM RRFA 算法在 Colon 癌症数据集的实验结果显示：有 11 个样

本被错分，平均分类正确率是$\frac{62-11}{62}\times 100\% = 82.26\%$。其中 6 个癌症患者被漏诊，5 个正常人被误诊。平均选择的特征数是 40.34 个，相对于原始的 2000 个基因特征，所选特征子集的规模大约是原始特征集规模的五十分之一。与本章提出的 SVM RRFA 基因选择算法相比，简化的 SVM RRFA 算法选择的特征子集的平均规模略大；平均分类正确率虽然略高，但是漏诊患者相比与误诊的正常人数要高。

从表 6.10 和表 6.11 可见，简化 SVM RRFA 算法在 Adenoma 和 Carcinoma 数据集上的灵敏度、特异性，以及 Matthews 相关系数的值与本章提出的 SVM RRFA 算法相同。

从表 6.12 所示简化 SVM RRFA 算法在 Colon 数据集的混淆矩阵可分别得到相应的灵敏度、特异性和 Matthews 相关系数的值为：$sensitivity = \frac{34}{34+6}\times 100\% = 85\%$，

$specifity = \frac{17}{5+17}\times 100\% = 77.27\%$，

$MCC = \frac{34\times 17 - 5\times 6}{\sqrt{(34+5)\times(34+6)\times(13+5)\times(13+6)}} = \frac{34\times 17 - 5\times 6}{\sqrt{39\times 40\times 18\times 19}} = 0.7502$。与本章提出的 SVM RRFA 基因选择算法相比，简化 SVM RRFA 算法的后两个参数的值要高。但是简化 SVM RRFA 算法并没有使得癌症患者的正确识别率得到提高。

6.4 小结

本章在第五章研究的基础上，根据基因数据集的特点，对第五章的 SVM RFA 特征选择算法进行改进，提出基于 SVM 分类模型的基因选择算法 SVM RRFA。该算法引入随机思想，能根据具体的基因数据集特点，在每次迭代中随机加入若干当前最重要的，也即最具有区分度的基因，实现基因选择。同时，在此基础上，为减少算法的时间开销，提出了简化的 SVM RRFA 基因选择算法。

普林斯顿大学基因表达工程的 3 个基因数据集实验测试和比较证明：本章提出的 SVM RRFA 基因选择算法能实现有效的基因选择，发现基因数据集的关键分类基因，实现有效的癌症诊断；简化 SVM RRFA 算法提高了 SVM RRFA 基因选择算法的分类正确率、特异性和 Matthews 相关系数；但是对于癌症患者的分类正确率并没有提高。

第七章　结论和展望

本研究论述了特征选择研究的内容、研究的必要性、研究的意义与面临的挑战；分析了选择 SVM 这一分类工具进行特征选择研究的原因；深入阐述了基于 SVM 的特征选择研究的现状和趋势；指出了现有基于 SVM 的特征选择算法研究中所存在的问题。针对现存问题，提出了 4 种不同的特征重要性评价准则，以及基于相应准则与 SVM 的混合特征选择算法；并针对基因数据集的特点，提出了基于 SVM 分类模型的基因选择算法。下面将对本研究的所有工作进行总结，并对未来研究进行展望。

7.1　研究结论

1. 提出了可用于任意类分类问题的基于 G-score 与 SVM 的特征选择算法，解决了基于 F-score 与 SVM 的特征选择算法只适用于两类分类问题的缺陷；并针对几种经典的特征搜索策略的时间复杂性高问题，提出了推广的前向顺序搜索策略 GSFS、推广的前向顺序浮动搜索策略 GSFFS，以及推广的后向顺序浮动搜索策略 GSBFS。以 G-score 度量特征对于分类的贡献大小，以 SVM 的分类性能评价特征子集的分类性能引导特征选择过程。UCI 机器学习数据库数据集的实验测试证明了提出的基于 G-score 与 SVM 的混合特征选择算法能实现有效的特征选择。其中，基于 G-score 与 SVM 的集中特征选择方法相比，就特征子集规模来看，前向顺序浮动混合特征选择算法效果最佳；但就分类正确率，即分类模型的泛化性能来看，前向顺序混合特征选择算法最优。

2. 提出了基于 D-score 与 SVM 的特征选择算法，解决了基于 G-score 与 SVM 的特征选择算法在衡量特征在类别间辨别能力大小时，只是基于类内、类间距离（或离散系数）之比值，没有考虑不同的特征测量量纲对特征区分度大小影响的缺陷。UCI 机器学习数据库的 9 个特征选择算法测试常用数据集实验，以及与基于 G-score 与 SVM 的相应特征选择方法的 5 折交叉验证实验比较显示：提出的基于 D-score 与 SVM 的特征选择算法所选择的特征具有更好的分类效果，其分类性能优于基于 G-score 与 SVM 的特征选择方法，达到了保持数据集辨识能力不变情况下进行维数压缩的目的。基于 D-score 与 SVM 的 3 种混合特征选择算法相比，就特征子集规模来看，前向顺序浮动混合特征选择算法最好；但分类器的泛化性能而言，前向顺序特征选择算法性能最好。

3. 提出了基于 DFS 与 SVM 的特征选择算法，解决了基于 G-score 与 SVM 和基于 D-score 与 SVM 的特征选择算法只是衡量单个特征在类间辨别能力的大小，没有考虑特征之间的相关性对于特征辨别能力大小影响的缺憾。DFS 通过计算多个特征构成的特征

子集的联合 G-score 值，考虑了特征子集中特征的联合作用，判断特征子集中所有特征对于分类的联合贡献，以此度量特征子集的类间区分能力大小，是一种特征子集区分度。同时，根据现有特征子集区分度评价方法 CFS 的特征相关性的正、负相关之分，提出不考虑特征之间相关性正、负之分，只考虑其是否相的特征子集区分度方法 CFSPabs。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：提出的基于 DFS 与 SVM 特征选择算法所选择的特征分类效果好，其分类性能优于基于 CFS 与 SVM，和基于 CFSPabs 与 SVM 的特征选择方法，达到了保持数据集辨识能力不变情况下进行维数压缩的目的;但是就特征子集规模来看,基于 CFSPabs 与 SVM 的特征选择算法最优。

4. 利用 SVM 对于非线性可分问题的最大泛化性能，提出了基于 SVM 分类模型的适用于多类分类问题的特征选择算法 SVM RFE 和 SVM RFA ，解决了分别基于 G-score、D-score 和 DFS 与 SVM 的特征选择算法的以类间距离（离散系数）与类内方差（离散系数）之比来衡量特征对于分类的贡献，在非线性可分的问题中有可能造成有效区分特征的误剔除缺陷； SVM RFE 解决了 Guyou 的 SVM-RFE 特征选择算法只适用于两类分类问题的缺陷；同时基于前向顺序选择思想的基于 SVM 分类模型的特征选择算法 SVM RFA 解决了 SVM RFE 对于高维数据集进行特征选择的时间效率问题。UCI 机器学习数据库 9 个经典数据集的 5 折交叉验证实验表明：提出的 SVM RFA 特征选择算法，以及适用于多类分类问题的 SVM RFE 特征选择算法,能在保持或提高分类正确率的前提下，实现有效的特征选择。实验结果同时显示：SVM RFA 算法优于 SVM RFE 算法。实验还表明：对于较低维数据集，该两个特征选择算法的效率差别不大，但是对于维数比较高的数据集进行特征选择时，SVM RFA 算法的效率明显优于 SVM RFE 算法。

5. 针对基因数据集中，样本通常只有几十个，而作为描述每一个样本的基因却有成千上万之多的特点，结合上衣研究结论，引入随机思想，提出基于 SVM 分类模型的基因选择算法——SVM RRFA，解决了基因选择过程中，每一次迭代应该剔除或加入的合适基因个数问题。同时，提出了简化的 SVM RRFA 基因选择算法，减少了 SVM RRFA 算法的时间开销。普林斯顿大学基因表达工程的 3 个基因数据集实验测试和分析比较表明：提出的 SVM RRFA 基因选择算法能实现有效的基因选择，发现基因数据集的关键分类基因，实现有效的癌症诊断；简化 SVM RRFA 算法提高了 SVM RRFA 基因选择算法的分类正确率、特异性和 Matthews 相关系数；但是对于癌症患者的分类正确率并没有提高。

7.2 研究展望

本研究虽然针对基于 SVM 的特征选择算法研究所存在的若干问题，进行了特征区分度准则的研究，提出了 4 种特征重要性的计算方法，以及相应的若干基于这些特征区分度与 SVM 的特征选择算法；并针对基因数据集的高维稀疏分布特点，提出基于 SVM 的两种基因选择算法 SVM RRFA 和简化 SVM RRFA。然而，在研究中还存在以下的问题需要进一步研究。

第一，如何确定最优的 SVM 核函数参数？本研究采用的网格搜索，虽然能解决问题，但是时间开销很大。因此，确定最优的核参数的方法问题有待进一步研究。

第二，如何选择核函数？本研究中核函数的选择也是根据前人研究的建议，对于高维的基因数据集采用线形核函数，其他采用 RBF 核函数。但是，到底选择哪种核函数？选择的依据是什么？这些都有待进一步研究。

第三，基于 SVM 的随机基因选择算法，虽然实现了有效的基因选择，但是实验结果有赖于随机数。如何根据具体的基因数据集产生最佳的随机数，有待进一步研究。

第四，尽管 Alon 等[25]对结肠癌基因数据集进行了聚类分析，Ben-Hur 等[168]和 Winters-Hilt 和 Marat[169]研究了 SVM 聚类算法，但基于 SVM 聚类思想的特征选择方法与基因选择研究还有待进一步研究。关于这方面的工作，我们已经进行了一部分实验。目前我们对于 Colon 数据集的 SVM 聚类实验结果在最佳时候只用 6 个基因就可以实现 80％以上的分类正确率。期望未来在这一部分的研究能有进一步的理论上的突破。

第五，现有基于 SVM 的多折交叉验证特征选择算法所选择的特征子集缺乏稳定性和鲁棒性。如何得到稳定和鲁棒的特征子集也有待进一步研究。

致　谢

在毕业论文完成之际，我非常感谢我的恩师谢维信教授！

谢老师对我论文的选题、进展和完成给予了亲切的关怀、鼓励和非常耐心地指导。每次交流谢老师的指点都会让我茅塞顿开，让我一下子豁然开朗，是谢老师的点拨我才有今天。师从谢老师近乎 8 年的过程，我从谢老师身上学到了很多，我非常敬佩谢老师高瞻远瞩的眼光和能力，非常钦佩谢老师深邃的洞察力，非常感谢谢老师对我的宽容和厚爱。是谢老师给我的宽松环境和谆谆教诲让我由一个没有自信心的大龄博士生慢慢走上了科研之路。让我不仅可以像过去一样继续在大学教书，而且找到了自己的研究方向，并且喜欢这份研究工作；让我站在讲台上时比过去更有信心；而且因为思路的拓宽，也使我可以将我的教学工作做得更好。完成毕业论文的过程中，我经常想到：“一日为师，终生为父”这句话，这句话深深表达了我对谢老师的感激。无论我在何方，谢老师在我心中永远是如父般的恩师！

我还要感谢我的师兄高新波教授！高老师不仅在学业上点拨和指导我，同时也在生活上给予我帮助、指导和鼓励。很钦佩高老师各方面的能力，在我心中，师兄一直是我学习的榜样。

感谢我的硕士导师刘芳教授，以及焦李成教授！感谢他们对我的启蒙和帮助！感谢刘老师和焦老师对我的厚爱！同时我要感谢陕西师范大学计算机科学学院的肖冰博士、西安电子科技大学电子工程学院的李阳阳博士，感谢他们所给予我的帮助和鼓励！

感谢英国 Brunel 大学的 Xiaohui Liu 教授，感谢他提供给我机会，让我在智能数据分析中心进行了一年的访问研究，感谢他在我访学英国一年期间给予我的无微不至的关怀、照顾和帮助，并感谢他在研究工作上给我的指导。

感谢我的工作单位——陕西师范大学计算机科学学院提供给我留学的机会，感谢国家留学基金委的资助，让我有机会去英国学习一年。

最后，我要感谢我的家人所给予的关爱！因为有爱，我才有源源不断的动力！

谢娟英

12-4-18

参考文献

[1] Fu K S, Min P J, Li T J. Feature selection in recognition. IEEE Transactions on systems science and cybernetics, 1970, SSC-6(1): 33-39.

[2] Widodo A, YANG B S. Application of nonlinear feature extraction and support vector machines for fault diagnosis of induction motors, Expert Systems with Applications, 2007, 33 (1): 241-250.

[3] Amaldi E, Kann V. On the approximation of minimizing non zero variables or unsatisfied relations in linear systems, Theoretical Computer Science, 1998, 209(1-2): 237–260.

[4] Brown M P, Grundy W N, Lin D, Cristianini N, Sugnet C W, Furey T S, Ares M Jr, Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS, 2000, 97(1): 262-267.

[5] 何小晨, 徐守时. 基于关联规则的特征选择方法. 红外与激光工程, 2002, 31(6): 504-508.

[6] 朱明, 王俊普, 蔡庆生. 一种最优特征集的选择算法. 计算机研究与发展, 1998, 35(9): 803-805.

[7] Blum A, Langley P. Selection of relevant features and examples in machine learning. Artificial Intelligence, 1997, 97 (1-2): 245-271.

[8] Kohavi R, John G. Wrappers for feature selection. Artificial Intelligence, 1997, 97(1-2): 273-324.

[9] Hua J P, Tembe W D, Dougherty E R. Performance of feature-selection methods in the classification of high-dimension data. Pattern Recognition, 2009, 42 (3): 409-424.

[10] Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm. AAAI-92 proceedings, pp. 129-134.

[11] Foithong S, Pinngern O, Attachoo B. Feature subset selection wrapper based on mutual information and rough sets. Expert systems with applications, 2012, 39(1): 574-584.

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

[13] Lal T N, Chapelle O, Weston J, Elisseeff A. Embedded methods. In: Guyon I, Gunn S, Nikravesh M, Zadeh L A (Eds.), Feature Extraction: Foundations and Applications. Studies in Fuzziness and Soft Computing, vol. 207, Springer, Berlin, Heidelberg, 2006, pp. 137-165.

[14] Breiman L, Friedman J H, Olshen R A, Stone C J. Classification and Regression Trees. Wadsworth, Monterey, CA, USA, 1984.

[15] Reunanen J. Overfitting in making comparisons between variable selection methods. Journal of machine learning research, 2003, 3: 1371–1382.

[16] Whitney A W. A direct method of nonparametric measurement selection. IEEE Transactions on Computers, 1971, 20(9): 1100-1103.

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

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

[19] Somol P, Pudil P, Novovicova J, Paclik P. Adaptive floating search methods in feature selection. Pattern Recognition Letters, 1999, 20(11-13): 1157-1163.

[20] Somol P, Pudil P. Oscillating search algorithms for feature selection. In: Proceedings of the 15th international conference on pattern recognition, Sanfeliu A, Villanueva J J, Vanrell M eds., IEEE Computer Society, Los Alamitos, 2000, pp. 406-409.

[21] Somol P, Pudil P, Grim J. Branch & bound algorithm with partial prediction for use with

recursive and non-recursive criterion forms. Lecture Notes in Computer Science LNCS 2013, Springer, 2001, pp. 230-239.

[22] Somol P, Pudil P, Kittler J. Fast branch & bound algorithms for optimal feature selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(7): 900-912.

[23] Nakariyakul S, Casasent D P. An improvement on floating search algorithms for feature subset selection. Pattern Recognition, 2009, 42(9): 1932-1940.

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

[25] Alon U, Barkai N, Notterman D A, Gish K, Ybarra S, Mack D, Levine A J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon cancer tissues probed by oligonucleotide arrays. PNAS, 1999, 96(12): 6745–6750.

[26] Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P, Coller H, Loh M L, Downing J R, Caligiuri M A, Bloomfield C D, Lander E S. Molecular classification of cancer: class discovery and class prediction by gene expression mMonitoring. Science, 1999, 286(5439): 531-537.

[27] Fodor S. DNA SEQUENCING: Massively Parallel Genomics, Science, 1997, 277(5324): 393-395.

[28] Alizadeh A A, Eisen M B, Davis R E, Ma C, Lossos I S, Rosenwald A, Boldrick J C, Sabet H, Tran T, Yu X, Powell J I, Yang L M, Marti G E, Moore T, Jr J H, Lu L S, Lewis D B, Tibshirani R, Sherlock G, Chan W C, Greiner T C, Weisenburger D D, Armitage J O, Warnke R, Levy R, Wilson W, Grever M R, Byrd J C, Botstein D, Brown P O, Staudt L M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 2000,403:503-511.

[29] Vant's Veer L J, Dai H, Van de Vijver M J, He Y D, Hart A A M, Mao M, Petersen H L, Kooy K, Marton M J, Witteveen A T, Schreiber G J, Kerkhoven R M, Roberts C, Linsley P S, Bernards R, Friend S H. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415, 530-536.

[30] Wheeler D A, Rothberg J M. The complete genome of an individual by massively parallel DNA sequencing, Nature, 2008, 452: 872-876.

[31] Durbin R M, Burton J, Carter D M, Churcher C, Wang J, et al. A map of human variation from population scale sequencing. Nature, 2010, 467(7319): 1061-1073.

[32] Armstrong S A, Staunton J E, Silverman L B, Pieters R, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics, 2002, 30(1): 41-47.

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

[34] Vinciotti V, Tucker A, Kellam P, Liu X H. The robust selection of predictive genes via a simple classifier. Applied Bioinformatics, 2006, 5(1): 1-11.

[35] 王明怡, 吴平, 王德林. 基于相关性分析的基因选择算法. 浙江大学学报(工学版), 2004, 38(10):1289-1292.

[36] Liu Y H. Dimensionality reduction and main component extraction of mass spectrometry cancer data. Knowledge-Based Systems, 2012, 26 (1): 207-215.

[37] Bolón-Canedo V, Sánchez-Maronõ N, Alonso-Betanzos A. An ensemble of filters and classifiers for microarray data classification. Pattern recognition, 2012, 45(1): 531-539.

[38] Li J T, Jia Y M. An improved elastic net for cancer classification and gene selection. ACTA

Automatica Sinica, 2010, 36(7): 976-971.
[39] 张树波, 赖剑煌. 基于融合信息的癌症相关基因选择方法. 计算机科学, 2010, 37(12): 171-175.
[40] Le Thi H A, Nguyen V V, Ouchani S. Gene selection for cancer classification using DCA. Journal of Frontiers of Computer Science and Technology, 2009, 3(6): 61-620.
[41] 马宁, 张正国. 一种基于 Gene Ontology 注释信息的基因选择算法. 中国生物医学工程学报, 2009, 28(5): 696-671.
[42] 李建更, 高志坤, 严志, 阮小刚. 基于双基因分析的结肠癌标志基因选择. 中国生物医学工程学报, 2009, 28(5): 691-695.
[43] 李建更. 高志坤. 随机森林:一种重要的肿瘤特征基因选择法. 生物物理学报, 2009, 25(1): 51-56.
[44] 叶奇明, 罗飞, 刘娟. 基于多目标 EDA 的特征基因选择. 计算机应用研究, 2009, 26(8): 2891-2894.
[45] 蔡丽君, 蒋林波, 易叶青. 基于蚁群优化算法的基因选择. 计算机应用研究, 2008, 25(9): 2754-2757.
[46] Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. Jornal of bioinformatics and computational biology, 2005, 3(2): 185-205.
[47] Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max relevance, and min-redundancy. IEEE transactions on pattern analysis and machine intelligence, 2005, 27(8): 1226-1238.
[48] Yeoh E J, Ross M E, Shurtleff S A, Williams W K, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell, 2002,1(2): 133-143.
[49] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 2004, 5:1205-1224.
[50] Hochreiter S, Obermayer K. Nonlinear feature selection with the potential support vector machine. In: Guyon I, Gunn S, Nikravesh M, Zadeh L. (eds.) Feature extraction, Studies in Fuzziness and Soft Computing, 2006, vol. 207, pp. 419-438, Springer Berlin / Heidelberg publisher.
[51] Li F, Yang Y, Xing E P. From Lasso regression to Feature Vector Machine. Advances in Neural Information Processing Systems 18, Weiss Y, Schólkopf B, Platt J (Eds.), MIT press, pp. 779-786, 2006.
[52] Cheng H B, Chen H F, Jiang G F, Yoshihira K. Nonlinear feature selection by relevance Feature Vector Machine. Lecture Notes in Computer Science, 2007, vol. 4571, Machine learning and data mining in pattern recognition, pp. 144-159.
[53] Liu H, Liu L, Zhang H. Feature selection using mutual information: an experimental study. Lecture Notes in Computer Science, vol. 5391, pp. 235-246. PRIC AI 2008: Trends in artificial intelligence.
[54] 张军英, Wang Y J, Khan J, Clarke R. 基于类别空间的基因选择. 中国科学（E 辑）, 2003, 33(12): 1125-1137.
[55] 李霞, 张天文, 李丽, 郭政. 决策树特征基因选择方法对 SVM 有效性的研究. 中国生物医学工程学报, 2004, 23(1): 66-72.
[56] Zhang Junying, Liu Shenliang, Wang Yue. Gene association study with SVM, MLP and

cross-validation for the diagnosis of diseases. Progress in Natural Science, 2008,18(6): 741-750.

[57] 张丽娟, 李舟军. 微阵列数据癌症分类问题中的基因选择. 计算机研究与发展, 2009, 46(5):794-802.

[58] 耿耀君, 张军英. 一种基于监督降维和形状分析的基因选择方法. 西安电子科技大学学报（自然科学版）, 2010,38(3): 121-127.

[59] Yu L, Han Y, Berens M E. Stable Gene Selection from Microarray Data via Sample Weighting. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(1): 262-272.

[60] Karabatak M, Ince M C. A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases, Expert Systems with Applications, 2009, 36(10): 12500-12505.

[61] Liu H W, Ssun J G, Liu L, Zhang H J. Feature selection with dynamic mutual information, Pattern Recognition, 2009, 42(7): 1330-1339.

[62] Übeyli E D. Multiclass support vector machines for diagnosis of erythemato-squamous diseases. Expert Systems with Applications, 2008, 35(4): 1733-1740.

[63] Polat K, Günes S. A new feature selection method on classification of medical datasets: Kernel F-score feature selection. Expert Systems with Applications, 2009, 36(7): 10367-10373.

[64] 王刚, 刘元宁, 陈慧灵, 董浩, 朱晓冬. 粗糙集与支持向量机在肝炎诊断中的应用. 吉林大学学报(工学版), 2011, 41(1): 160-164.

[65] Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifier. In: Fifth Annual Workshop on Computational Learning Theory, ACM, 1992, pp. 144 -152.

[66] Vapnik V. Statistical Learning Theory. Wiley, 1998.

[67] Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel_based learning methods. Cambridge University Press, 2000.

[68] Miranda J, Montoya R, Weber R. Linear penalization support vector machines for feature selection. In: Pal S K *et al*. (Eds.), PReMI 2005, LNCS, vol.3776, Springer-Verlag, Berlin Heidelberg, 2005, pp. 188-192.

[69] Schölkopf B, Smola A J. A tutorial on support vector regression. NeuroCOLT2 Technical Report Series, NC2-TR-1998 -030, 1998.

[70] Burges C J. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 121-167.

[71] Huang T M, Kecman V. Gene extraction for cancer diagnosis by support vector machines-An improvement. Artifical Intelligence in Medicine, 2005, 35(1-2): 185-194.

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

[73] Li F, Yang Y M. Analysis of recursive gene selection approaches from microarray data. Bioinformatics, 2005, 21(19): 3741-3747.

[74] 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.

[75] Wiliński A, Osowski S. Gene selection for cancer classification. COMPEL: The International Journal for Computation and Mathematics in Electrical and Electric Engineering, 2009, 28(1): 231-241.

[76] Lin C F, Wang S D. Fuzzy support vector machines. IEEE Transactions on Neural Networks, 2002, 13(2): 464-471.

[77] Inoue T, Abe S. Fuzzy support vector machines for pattern classification. In proceedings of International Joint Conference on Neural Networks (IJCNN'01), 2001, vol. 2, pp: 1449-1454.
[78] Abe S, Inoue T. Fuzzy support vector machines for multiclass problems. In: ESANN'2002 proceedings-European Symposium on Artificial Neural Networks, Bruges (Belgium), 2002, pp.113-118.
[79] Jayadeva, Khemchandani R, Chandra S. Twin support vector machines for pattern classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5): 905-910.
[80] Peng X J. A *v*-twin support vector machine (*v*-TSVM) classifier and its geometric algorithms. Information Sciences, 2010, 180(20): 3863-3875.
[81] Xie Juanying, Xie Weixin, Gao Xinbo, Hone Kate, Shi Yong, Liu Xiaohui. Extending Twin Support Vector Machine Classifier for Multi-Category Classification Problems. Intelligent data analysis, 2012, in press.
[82] Xu J H. An efficient multi-label support vector machine with a zero label. Expert systems with applications, 2012, 39(5): 4796-4804.
[83] 陈开周. 最优化计算方法. 西安电子科技大学出版, 1984.
[84] 毛勇, 周晓波, 夏铮, 尹征, 孙优贤. 特征选择算法研究综述. 模式识别与人工智能, 2007, 20(2): 211 - 218.
[85] Modrzejewski M. Feature selection using rough sets theory. Lecture notes in computer science, Proceedings of the European Conference on Machine Learning, Springer-verlag, London, UK, 1993, Vol. 667, pp. 213-226.
[86] Chan C C. A rough set approach to attribute generalization in data mining. Information Sciences, 1998, 107(1-4): 169-176.
[87] 苗夺谦, 胡桂荣. 知识约简的一种启发式算法. 计算机研究与发展, 1999, 36(6): 681-684.
[88] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简. 计算机学报, 2002, 25(7): 759-767.
[89] Questier F, Arnaut-Rollier I, Walczak B, Massart D L. Application of rough set theory to feature selection for unsupervised clustering. Chemomertics and Intelligent Laboratory Systems, 2002, 63(2): 155-167.
[90] 刘少辉, 盛秋戬, 吴斌, 史忠植, 胡斐. Rough 集高效算法的研究. 计算机学报, 2003, 26(5): 524 -529.
[91] 谢娟英, 谢维信, 高新波. 基于树结构的属性约简方法. 模糊逻辑与计算智能研究进展（上册）, 中国模糊逻辑与计算智能联合学术会议论文集. 中国科学技术大学出版社, 2005, pp. 360-364.
[92] 谢娟英, 李楠, 乔子芮. 基于邻域粗糙集的不完整决策系统特征选择算法. 南京大学学报（自然科学版）, 2011, 47(7): 383-390.
[93] 亢婷, 魏立力. 一种基于粗糙集理论的启发式特征选择算法. 计算机工程与应用, 2008, 44(30): 77-79.
[94] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. 软件学报, 2008, 19(3): 640-649.
[95] 蒙祖强, 史忠植. 一种新的基于简化二进制可辨矩阵的相对约简算法. 控制与决策, 2008, 23(9): 976-980.
[96] 朱颢东, 钟勇. 基于粗糙集和灰色关联度的综合性特征选择. 计算机工程与应用, 2009, 45(35): 6-11.
[97] Hu Q H, Pedrycz W, Yu D R, Lang J. 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.

[98] 赵军阳, 张志利. 基于模糊粗糙集信息熵的蚁群特征选择方法. 计算机应用, 2009, 29(1): 109-113.

[99] Swiniarski R W, Skowron A. Rough set methods in feature selection and recognition. Pattern Recognition Letters, 2003, 24(6): 833-849.

[100] Caballero Y, Lvarez D, Bello R, Garia M M. Feature selection algorithms using rough set theory. Seventh International Conference on Intelligent Systems Design and Applications, 2007, pp. 407-411.

[101] Hu Q H, Yu D R, Liu J F, Wu C X. Neighborhood rough set based heterogeneous feature subset selection. Information Sciences, 2008, 178(10): 3577-3594.

[102] 孙丽君, 苗夺谦. 基于粒度计算的特征选择方法. 计算机科学, 2008, 35 (4): 14-15, 39.

[103] Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, Vapnik V. Feature selection for SVMs. Advances in Neural Information Processing Systems 13, MIT Press, 2000, pp. 668-674.

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

[105] Chen Y W, Lin C J. Combining SVMs with various feature selection strategies. NIPS 2003 feature selection challenge, 2003, pp. 1-10.

[106] 陈光英, 张千里, 李星. 特征选择和 SVM 训练模型的联合优化. 清华大学学报(自然科学版), 2004, 44(1): 9-13.

[107] Liu Y, Zheng Y F. FS_SFS: A novel feature selection method for support vector machines. Pattern Recognition, 2006, 39(7): 1333-1345.

[108] 任江涛, 赵少东, 许盛灿, 印鉴. 基于二进制 PSO 算法的特征选择及 SVM 参数同步优化. 计算机科学, 2007, 34(6):179-182.

[109] 李伟红, 陈伟民, 杨利平, 龚卫国. 基于不同 Margin 的人脸特征选择及识别方法. 电子与信息学报, 2007, 29(7): 1744-1748.

[110] Ponsa D, López. Feature selection based on a new formulation of the Minimal-Redundancy-Maximal-Relevance criterion. Martí J, et al (Eds): IbPRIA 2007, Part I, LNCS 4477, pp. 47-54, 2007.

[111] 任双桥, 傅耀文, 黎湘, 庄钊文. 基于分类间隔的特征选择算法. 软件学报, 2008, 19(4): 842-850.

[112] 吴永辉, 计科峰, 李禹, 郁文贤. 利用 SVM 的极化 SAR 图像特征选择与分类. 电子与信息学报, 2008, 30(10): 2347-2351.

[113] 孙刚, 王志平, 王明新. 一种提高支持向量机性能的特征选择新方法. 计算机工程与应用, 2008, 44(3): 183-185.

[114] 杨立才, 李金亮, 姚玉翠, 吴晓晴. 基于 F-score 特征选择和支持向量机的 P300 识别算法. 生物医学工程学杂志, 2008, 25(1): 23-27.

[115] 郝艳友, 迟忠先, 李克秋, 张永. 基于 IGA 的支持向量机特征子集选择和参数优化. 计算机工程与应用, 2008, 44(22): 35-38.

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

[117] 吕世聘, 王秀坤, 孙岩, 唐一源. 改进的支持向量机特征选择算法. 计算机工程, 2009, 35(1): 171-172.

[118]计智伟, 吴耿锋, 胡珉. 基于自适应遗传算法和 SVM 的特征选择. 计算机工程, 2009, 35(14): 200-203.
[119]谢娟英, 王春霞, 蒋帅, 张琰. 基于改进的 F-score 与支持向量机的特征选择方法. 计算机应用, 2010, 30(4) : 993-996.
[120]谢娟英, 雷金虎, 谢维信, 高新波. 基于 D-score 与支持向量机的混合特征选择方法, 计算机应用, 2011, 31(12): 3292-3296.
[121]谢娟英, 雷金虎, 谢维信. 一种新的特征评价方法在红斑鳞状皮肤病诊断中的应用, 中国生物医学工程学报, 2012, 131(1):33-41.
[122]Xie Juanying, Xie Weixin, Wang Chunxia, Gao Xinbo. A novel hybrid feature selection method based on IFSFFS and SVM for the diagnosis of erythemato-squamons diseases. Journal of Machine Learning Research: Workshop and Conference Proceedings, 2010, Vol. 11: 142-151.
[123]Xie Juanying, Wang Chunxia. 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.
[124]Xie Juanying, Lei Jinhu, Xie Weixin, Gao Xinbo, Shi Yong, Liu Xiaohui. Novel hybrid feature selection algorithms for diagnosing erythemato-squamous diseases, LNCS 2012, 7231, pp. 173-185. Springer, Heidelberg. J. He et al. (Eds.): HIS 2012.
[125]戴平, 李宁. 一种基于 SVM 的快速特征选择算法. 山东大学学报（工学版）, 2010, 40(5): 60-65.
[126]刘董, 郑宁, 杨杰, 徐明. 基于遗传算法的 SVM 特征选择和模型参数优化. 计算机应用与软件, 2009, 26(1): 85-87,117.
[127]杜卓明, 冯静. 改进遗传算法和支持向量机的特征选择算法. 计算机工程与应用, 2009, 45(29): 28-30.
[128]赵明渊, 唐勇,傅翀, 周明天. 基于带特征染色体遗传算法的支持向量机特征选择和参数优化. 控制与决策, 2010, 25(8): 1133-1138.
[129]易超群, 李建平, 朱成文. 一种基于分类精度的特征选择支持向量机. 山东大学学报（理学版）, 2010, 45(7): 199-122.
[130]Xia H, Hu B Q. Feature selection using fuzzy support vector machines. Fuzzy Optim Decis Making, 2006, 5(2): 187-192.
[131]孟范静, 刘毅慧, 王洪国, 成金勇. SVM 在基因微阵列癌症数据分类中的应用. 计算机工程与应用, 2007, 43(34):246-248.
[132]张焕萍, 宋晓峰, 王惠南. 基于离散粒子群和支持向量机的特征基因选择算法. 计算机与应用化学, 2007, 24(9): 1159-1162.
[133]于华龙, 顾国昌, 刘海波, 沈晶, 朱长明. 改进的离散 PSO 和 SVM 的特征基因选择算法. 哈尔滨工程大学学报, 2009,12(30): 1339-1404.
[134]游伟, 李树涛, 谭明奎. 基于 SVM-RFE-SFS 的基因选择算法.中国生物医学学报, 2010, 29(1): 93-99.
[135]Lee C P, Leu Y. A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing, 2011, 11(1): 208-213.
[136]Luo L K, Huang D F, Ye L J, Zhou Q F, Shao G F, Peng H. Improving the computational efficiency of recursive cluster elimination for gene selection. IEEE-ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(1): 122-129.
[137]Choi H, Yeo D, Kwon S, Kin Y. Gene selection and prediction for cancer classification using

support vector machines with a reject option. Computational Statistics and Data Analysis, 2011, 55(5): 1897-1908.

[138] Li J T, Jia Y M , Li W L. Adaptive huberized support vector machine and its application to microarray classification. Neural Computing & Applications, 2011, 20(1): 123-132.

[139] Zheng C H, Chong Y W, Wang H Q. Gene selection using independent variable group analysis for tumor classification. Neural Computing & Applications, 2011, 20(2): 161-170.

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

[141] Bonev B, Escolano F, Cazorla M A. A novel information theory method for filter feature selection. LNAI 4827, pp. 431–440, 2007. Gelbukh A and Kuri Morales A F (Eds.): MICAI 2007.

[142] Schurmann J. Pattern Classification, a Unified View of Statistical and Neural approaches. Wiley, New York, 1996.

[143] Conver W J. Practical Nonparametric Statistics. Wiley, New York, 1980.

[144] Duda R O, Hart P E, Stork P. Pattern classification and scene analysis. Wiley, New York, 2003.

[145] Chandra B, Gupta M. An efficient statistical feature selection approach for classification of gene expression data. Journal of Biomedical Informatics, 2011, 44(4): 529-535.

[146] Huang S H, Mo D, Meller J, Wagner M. Identifying a small set of marker genes using minimum expected cost of misclassification. Artificial Intelligence in Medicine, 2012, in press.

[147] Ramaswamy S, Tamato P, Rifkin R, Mukherjee S, Yeang C H, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov J P, Poggio T, Gerald W, Loda M, Lander E S, Golub T R. Multiclass cancer diagnosis using tumor gene expression signatures. PNAS, 2001, 98(26):15149-15154.

[148] Mundra P A, Rajapakse J C. Support vector based T-score for gene ranking. LNBI 5265, pp. 144-153, 2008. Chetty M, Ngom A, and Ahmad S (Eds.): PRIb 2008.

[149] Mundra P A, Rajapakse J C. SVM-RFE with MRMR filter for gene selection. IEEE Transactions on NanoBioscience, 2010, 9(1): 31-37.

[150] Tang Y, Zhang Y Q, Huang Z. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007, 4(3): 365-381.

[151] Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 2010, 26(3):392-398.

[152] 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.

[153] Zhou Q F, Hong W C, Shao G F, Cai W Y. A new SVM-RFE approach towards ranking problem. Proceedings of IEEE international conference on intelligent computing and intelligent systems, 2009, pp. 270-273.

[154] Lee M C. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications, 2009, 36(8): 10896–10904.

[155] Monirul Kabir M, Shahjahan M, Murase K. A new hybrid ant colony optimization algorithm for feature selection. Expert systems with applications, 2012, 39(3): 3474-3763.

[156] Vapnik V. The nature of statistical learning theory. New York: Springer, 2000.

[157] 边肇祺, 张学工等. 模式识别（第二版）. 北京：清华大学出版社,1999.

[158] Frank A, Asuncion A. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2010.

[159] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification. Taibei: Department of Computer Science, National Taiwan University, 2003.

[160] Chang C C, Lin C J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, 2 (3): 1-27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[161] 刘春英，贾俊平等．统计学原理．北京: 中国商务出版社, 2008.

[162] Hall M A, Correlation-based feature selection machine learning, Ph.D. Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.

[163] Han J W, Kamber Micheline. Data Mining: Concepts and Techniques, 2nd edition, 2006, 4. pp. 70-71. China Machine Press under special arrangement with Elsevier (Singapore) Pte Ltd. 2011, 7.

[164] http://genomics-pubs.princeton.edu/oncology/.

[165] Fawcett T. An introduction to ROC analysis. Pattern recognition letters, 2006, 27(8): 861-874.

[166] Swets J A. Measuring the accuracy of diagnostic systems, Science, 1988, 240(4857): 1285-1293.

[167] Notterman D A, Alon U, Sierk A J, Levine A J. Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays, Cancer research, 2001, 61(7): 3124-3130.

[168] Ben-Hur A, Horn D, Siegelmann H T, Vapnik V. Support vector clustering. Journal of Machine Learning Research, 2001, 2: 125-137.

[169] Winters-Hilt S, Merat S. SVM clustering. BMC Bioinformtics, 2007, 8(Suppl 7): S18.

攻读博士学位期间的研究成果

学术论文

[1] Xie Juanying, Xie Weixin, Wang Chunxia, Gao Xinbo. A novel hybrid feature selection method based on IFSFFS and SVM for the diagnosis of erythemato-squamons diseases. Journal of Machine Learning Research: Workshop and Conference Proceedings, 2010, Vol. 11: 142-151.

[2] Xie Juanying, Wang Chunxia. 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. (SCI 二区)

[3] Xie Juanying, Jiang Shuai, Xie Weixin, Gao Xinbo. An efficient global K-mens clustering algorithm. Journal of Computers, 2011, 6(2): 271-279. (EI 源刊)

[4] Xie Juanying, Xie Weixin, Gao Xinbo, Hone Kate, Shi Yong, Liu Xiaohui. Extending Twin Support Vector Machine Classifier for Multi-Category Classification Problems. Intelligent Data Analysis, 2011, accept. (SCI 四区)

[5] Xie Juanying, Lei Jinhu, Xie Weixin, Gao Xinbo, Shi Yong, Liu Xiaohui. Two-stage hybrid feature selection algorithms for erythemato-squamous diseases. Health Information Science and Systems, 2012, accept.

[6] 谢娟英, 谢维信, 高新波. 基于树结构的属性约简方法. 模糊逻辑与计算智能研究进展（上册）, 中国模糊逻辑与计算智能联合学术会议论文集. 中国科学技术大学出版社, 2005, pp. 360-364.

[7] Xie Juanying, Lei Jinhu, Xie Weixin, Gao Xinbo, Shi Yong, Liu Xiaohui. Novel hybrid feature selection algorithms for diagnosing erythemato-squamous diseases. LNCS, 2012, 7231, pp. 173-185. Springer, Heidelberg. J. He et al. (Eds.): HIS 2012.

[8] 谢娟英, 雷金虎, 谢维信. 一种新的特征评价方法在红斑鳞状皮肤病诊断中的应用. 中国生物医学工程学报, 2012, 131(1):33-41.

[9] 谢娟英, 王春霞, 蒋帅, 张琰. 基于改进的 F-score 与支持向量机的特征选择方法. 计算机应用, 2010, 30(4) : 993-996.

[10] 谢娟英, 雷金虎, 谢维信, 高新波. 基于 D-score 与支持向量机的混合特征选择方法. 计算机应用, 2011, 31(12): 3292-3296.

[11] 谢娟英, 马箐, 谢维信. 一种确定最佳聚类数的新算法. 陕西师范大学学报: 自然科学版, 2012, 40(1): 13-18.

[12] 谢娟英, 郭文娟, 谢维信, 高新波. 基于样本空间分布密度的改进次胜者受罚竞争学习算法. 计算机应用, 2012, 32(3): 638-642.

[13] 谢娟英, 郭文娟, 谢维信, 高新波. 基于样本空间分布密度的初始聚类中心优化 K-均值算法. 计算机应用研究, 2012, 29(3): 888-892.

[14] 谢娟英, 郭文娟, 谢维信. 基于邻域的K中心点聚类算法. 陕西师范大学学报: 自然科学版, 2012, 40(4): xx-xx.

[15] 谢娟英, 蒋帅, 王春霞, 张琰, 谢维信. 一种改进的全局 K-均值聚类算法. 陕西师范大学学报（自然科学版）, 2010, 38(2): 18-22.

[16] 谢娟英, 张琰, 谢维信, 高新波. 一种新的密度加权粗糙 K-均值聚类算法. 山东大学学报（自然科学版）, 2010, 45(7): 1-6.

参加研究的科研项目

[1] 陕西省自然科学基础研究计划项目，基于粗糙集理论的蝗灾发生预测方法研究，2011-01-01~2012-12-31，在研，主持。

基于SVM的特征选择方法研究

作者： 谢娟英
学位授予单位： 西安电子科技大学