文本复制检测报告单(全文标明引文)

№:ADBD2018R_20180530153127201805301548384401741538 92　　检测时间:2018-05-30 15:48:38

检测文献:　53140702_孙博阳_计算机科学与技术_基于大间隔分布机的递归特征消除算法实现与应用

作者:　孙博阳

检测范围:　中国学术期刊网络出版总库

中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库

中国重要会议论文全文数据库

中国重要报纸全文数据库

中国专利全文数据库

图书资源

优先出版文献库

大学生论文联合比对库

互联网资源(包含贴吧等论坛资源)

英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)

港澳台学术文献库

互联网文档资源

CNKI大成编客-原创作品库

个人比对库

时间范围:　1900-01-01至2018-05-30

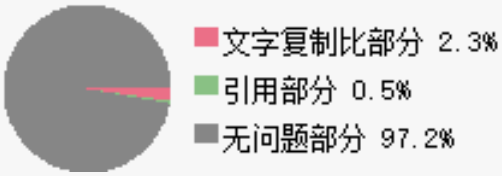
检测结果

总文字复制比：2.8%　　跨语言检测结果：0%

去除引用文献复制比：2.3%　　去除本人已发表文献复制比：2.8%

单篇最大文字复制比：0.9%

重复字数：	[788]	总段落数：	[6]
总字数：	[27751]	疑似段落数：	[2]
单篇最大重复字数：	[262]	前部重合字数：	[0]
疑似段落最大重合字数：	[485]	后部重合字数：	[788]
疑似段落最小重合字数：	[303]		



指　标：☐疑似剽窃观点　☑疑似剽窃文字表述　☐疑似自我剽窃　☐疑似整体剽窃　☐过度引用

表格：0　　公式：8　　疑似文字的图片：0　　脚注与尾注：0

0%（0）	中英文摘要等（总4642字）
9.6%（485）	第1章绪论（总5057字）
3.2%（303）	第2章数据与分析方法（总9380字）
0%（0）	第3章递归特征消除方法的实践与应用（总2893字）
0%（0）	第4章结果分析与总结（总4375字）
0%（0）	第5章总结与展望（总1404字）

（注释：无问题部分　文字复制比部分　引用部分）

1. 中英文摘要等　　总字数：4642

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

摘要

基于大间隔分布机的递归特征消除算法实现与应用

在机器学习领域，从最原始的问题——线性分类问题开始，经过多年的发掘和研究，一种名为支持向量机（ Support Vector Machine，简称 SVM）的分类器应运而生。多年以来，统计学习占有主导地位，但是随着SVM的出现，我们挣扎于的

线性可分的问题，甚至到线性不可分，SVM算法都可以顺利完成目标。 由此可见，支持向量机算法的流行是机器学习领域大步前进的重大推动力。

支持向量机(SVM)的将算法重点放在了两个“最”，即要求最大化间隔，同时保证这个间隔是最小间隔。然而，近几年研究出的新理论表示，人们发现SVM的最小间隔的尽可能大的处理已经不满足于现在的多嘈杂数据，无法获得较为广义上的应用和实践，同时表示间隔分布有着更深刻的探讨价值。

因此，大间隔分布机(Large margin Distribution Machine， LDM)应运而生，它诞生的原因正是由于将SVM的上述弊病加以剔除和更正，将关注重点创新性的放在“间隔均值”以及“间隔方差”上， 这样就恰好弥补了SVM在“间隔误差”部分的问题。

用于解决二分类问题的方法中，特征选取也是较为关键和复杂的一个部分。 本文通过递归消除特征算法来实现特征“重要性”的排行。递归消除特征算法使用一个基模型来进行多轮训练，每轮训练后，去掉最小特征得分的特征，通过剩下的少量特征进行模型的更新与许椄，重复上述过程，最后得到了每次“被去掉”的特征排序，得到特征重要性排序。

本文从理论上及实验上，通过支持向量机（ SVM ）与大间隔分布机（ LDM ）运用于相同数据集的测试准确率，用于证明LDM的可行性（与SVM预测水平相当）和在某类问题的优势。

本文采用的数据为医学领域上的到目前为止较为棘手的癌症早期诊断问题的相关数据。具体来讲，论文使用的数据集为来自 GEO数据库 GSE15471数据集，内含39对胰腺导管癌和相邻对照胰腺组织，通过支持向量机（SVM）和大间隔分布机（LDM）的递归特征消除（RFE）方法进行对照实验，对生物标志物的研究分析和鉴别来对胰腺癌的早期治疗和发展做出贡献。将实验结果进行对比和分析，用以验证大间隔分布机的优缺点。并在实际的应用中将得到的标志物进行排行，将具体的信息加以总结，为肿瘤学家提供有价值的线索，以便更好的诊断胰腺癌症。

关键字：支持向量机，最小间隔，大间隔分布机，间隔分布，递归特征消除， 胰腺癌，标志物检测

Abstract

Implementation and Application of Recursive feature elimination Algorithm Based on Large margin Distribution Machine

In the field of machine learning, a classifier called Support Vector Machine (SVM) emerged from the most primitive problem, the linear classification problem, after years of exploration and research. For many years, statistical learning has dominated, but with the advent of SVM, the linearly separable problems we are struggling with are even linearly inseparable, and SVM algorithms can successfully accomplish their goals. It can be seen that the popularity of support vector machine algorithms is a major driving force in the field of machine learning.

The support vector machine (SVM) focuses on the two "most", that is, the need to maximize the interval, while ensuring that the interval is the minimum interval. However, the new theory developed in recent years shows that people have found that the maximum possible processing of the minimum interval of SVMs is no longer satisfied with the current multi-noise data and can not obtain applications and practices that are more generalized, and that the interval distribution has more Deeply explore the value.

Therefore, a large margin distribution machine (LDM) came into being. The reason for its birth was precisely because of the elimination and correction of the above-mentioned ills of SVM, and the focus on innovation was placed on the “interval mean” and “interval”. "On the variance", this just makes up for the SVM's problem in the "interval error" section.

In the method for solving the two-classification problem, feature selection is also a key and complex part. This article implements the ranking of features "importance" through recursive elimination of feature algorithms. The recursive elimination feature algorithm uses a base model to perform multiple rounds of training. After each round of training, the feature of the minimum feature score is removed, and the model is updated with the remaining small number of features to update the model, and the above process is repeated, and finally each time is obtained. "Striped" feature sorting, to get the feature importance ranking.

This paper theoretically and experimentally uses the support vector machine (SVM) and large interval spreader (LDM) to apply the test accuracy of the same data set. It is used to prove the feasibility of LDM (comparable with the SVM prediction level) and The advantages of class problems.

The data used in this paper is related to the problem of early diagnosis of cancer that has been more difficult to date in the medical field. Specifically, the data set used in this paper is from the GSE15471 dataset in the GEO database, which contains 39 pairs of pancreatic ductal carcinomas and neighboring control pancreatic tissues. The recursive feature elimination by support vector machine (SVM) and large interval distribution machine (LDM) is used. The (RFE) method is used for conducting control experiments, analyzing and identifying biomarkers to contribute to the early treatment and development of pancreatic cancer. The experimental results were compared and analyzed to verify the advantages and disadvantages of the large interval distributor. In actual applications, the markers obtained will be ranked and specific information will be summarized to provide valuable clues for oncologists to better diagnose pancreatic cancer.

Key words: support vector machine, minimum interval, large interval distributor, interval distribution, recursive feature elimination, Pancreatic cancer, marker detection

目录

2. 第1章绪论

总字数：5057

相似文献列表　文字复制比：9.6%(485)　疑似剽窃观点：(0)

1	53130928_宋莹莹_计算机科学与技术_基于机器学习方法的海洋温跃层的深度计算 宋莹莹 - 《大学生论文联合比对库》- 2017-06-02	3.9%（197） 是否引证：否
2	若干改进的支持向量分类机 刘振丙(导师：刘小茂) - 《华中科技大学硕士论文》- 2006-04-01	3.1%（158） 是否引证：否
3	基于Ontology的个性化信息服务方法研究 刘志伟(导师：卢涛) - 《哈尔滨工业大学硕士论文》- 2006-06-01	2.4%（122） 是否引证：否
4	局域多分辨小波支持向量回归模型的研究及其应用 林兰馨(导师：孙丽莎) - 《汕头大学硕士论文》- 2009-05-01	1.9%（98） 是否引证：否
5	基于内容的图像检索若干技术研究 赵倩(导师：曹家麟) - 《上海大学博士论文》- 2012-08-01	1.7%（87） 是否引证：否
6	支持向量机性能分析及改进 - 豆丁网 -《互联网文档资源（http://www.docin.com）》- 2015	0.6%（32） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第1章绪论

1.1 机器学习的研究意义

说起机器学习，必然要提到三个问题——“什么是机器学习”，“为什么需要机器学习”，“机器学习到底能做什么事情”。

开门见山的说，机器学习这个学科的主要目的就是为了研究人类的思维模式和动作系统通过科学的模拟实现会“学习”的机

器。机器学习使得机器有了“学习”的能力，让只会进行简单或者复杂计算的机器有了“智力”。[1]

机器学习的前身或者说他需要解决的问题就是当今最为流行的科学——计算科学（computer science,缩写为CS）。无论在任何的时代，计算科学总是会对“数据”，“逻辑”，“计算力”三个核心点产生的木桶效应的场景进行研究和分析。[2]

然而，回顾机器学习的诞生，也同样是上述三个核心点互相作用的结果。

计算化时代的早期，我们有的是数据，逻辑（也可以称之为方法），但是缺乏计算力——应运而生的就是电子计算机。

信息化时代的早期，我们的计算能力逻辑都有所支撑。但是，数据量的获取却总是达不到要求。 从计算力的角度来说，我们得到了计算机这一强有力的支持，从逻辑方法而言， 在现有的计算领域内已经成熟的方法也足够这个时代的使用。 然而，数据成为了阻碍计算科学前进的一大障碍。[3]

大数据时代，也就是我们现在所处的时代，我们的计算力和数据已经得到了充分的补充和进步，随之而来的就是方法的缺失——我们没有足够的方法来处理现有的数据。[4]

人们不禁会思考，为什么我们曾经认为充足的逻辑方法到现在却出现了“匮乏”的情况——实际上，就是因为计算科学的进步，导致了现在的计算科学的领域已经不仅仅是计算机，处理处理数据这么简单了，已经扩展到了许多其他学科领域。然而其他的学科缺乏的逻辑方法问题渐渐浮上水面——变成了计算科学“分内”的工作。这也就是所谓的“能力越强，责任越大”。

谈到“机器学习”,我个人认为最经典的一句话就是——“三十年前的理论思想，三十年后的运算水平”。这句话的内容不言而喻，举个简单的例子，像认知科学和脑科学领域中，最为接近的则是传统的数理方法和冯诺依曼机体系，但是遗憾的是他们并不能有效的大幅度促进这两个领域的前进。诸如其他的人工智能领域，人体识别，图像识别领域都在等待一个新的“数理方法”来“解救”。[5]

在这个背景下，机器学习登上了历史舞台。

最后，我们可以更加笼统的说，机器学习的功能或者说最为总结归纳的说就是一种分类问题的解决方案。它可以解决当今最为流行的人工智能等诸多问题。它从数据变化的过程中选择合适的方法来自动的归纳总结逻辑，并按照这个逻辑进行与新数据的“碰撞”，达到预测的效果。

1.2 支持向量机的研究现状

机器学习的背景我们通过刚才的知识介绍已经有了初步的了解，我们来粗略的了解一下支持向量机。

在此注明，本章节仅为知识介绍，涉及计算公式和推导过程的内容可参见后面章节。

上面提到，机器学习的目的是根据经过处理过的数据样本作为“训练样本”，通过不同功能的算法求出对数据样本输入输出之间的依赖关系，通过此过程不断的修正算法得到稳定的参数，最后使得这一算法得到的学习模型有对未知的输入数据做出尽可能准确的预测的能力。

由此，诞生了两大“派别”的逻辑方法——统计学习和支持向量机。

概括的说，统计学习理论将“期望风险”和“经验风险”作为衡量算法好坏的一大重要标准，顺利解决的学习能力和推广能力上的统一这一问题。

不久后，统计学习的理论已经有了一定的进展后，发展出了一个新的通用学习的方法。 这就是我们所说的***“支持向量机（Support vector machine，简称SVM）”***。支持向量机在解决“小样本”，“非线性”，“高维模式识别”这几个问题上有较好的效果。

从理论上来讲，支持向量机的目的是求解“最有分类面”的问题。在某种意义上来说，对于线性可分的情况有一定程度的“偏好”。要求尽可能做到分类间隔最大，可以通过 Lagrange乘子法转化为对偶问题，在满足库恩-塔克条件（ Karush- Kuhn-Tucker， KKT）后，对这类约束优化问题进行分析和求解。 [6]

支持向量机算法已经流行了很长一段时间，从最初版本的SVM开始，现如今已经通过改进算法和应用算法获得了长足的进步。

从解决对偶问题的工程中我们发现，需要解决的问题相当于一个存在线性约束的二次规划问题（QP），计算的具体内容为核函数矩阵，但是这个矩阵的大小和训练样本数的平方成正相关，因此在提高样本量的过程中，占用的空间内存也是极大的。由此人们提出了三大方法——“块处理算法”，“固定工作样本集算法”，“SMO算法”。[7]

加速运算的方法还有通过核函数参数的选择以及具体选择哪个核函数来实现。 这一部分是当今SVM研究的一大重点。 它决定着解决大数据环境下的效率，影响着这一领域的前进与发展。 [8]

上文中提到过的支持向量机，对二分类问题有较好的适应性，往往能得到较为满意的结果。因此，人们会不断探索如何将支持向量机推广到多分类问题上。 现如今已经有了多种方法，但是并不成熟。

“一对多（One class Versus all Others,OVO）”,“一对一（One class Versus Another class,OVA）”,“SVM决策树（Decision Tree Method,DTM）”,“多类支持向量机（Multi-class SVM）”,这几大方法已经较为常用，但是设计和发掘更合适的算法仍然是现在支持向量机领域内最为严峻的课题和挑战。[9]

1.3 大间隔分布机的研究现状

上一小节提到的支持向量机现如今已经有了多种改进的算法，大间隔分布机(Large margin Distribution Machine, LDM)就是其中之一。

支持向量机的目的是找到最大分类面，这个分类面是需要解决二次规划问题（QP）。支持向量机的核心思想可以概括为“最大化”“最小间隔”，明确的说，就是将样本和分类面之间的距离最大限度的的缩小，以达到最终所需要的效果。

在上述思想中，“间隔理论”是其中的重中之重，值得一提的是，间隔理论不仅仅用于支持向量机，还可以用于多种学习算法。例如集成学习中的AdaBoost算法，但是在支持向量机领域，这部分仍然存在着一定的的缺陷。

直到2014年，发表在 ACM的论文——《大间隔分布学习机》，张腾、周志华两位教授[10]正式提出了大间隔分布机(Large margin Distribution Machine， LDM)，经过研究得出，间隔分布可以通过优化来获得，并且也能通过上述方法来得到更好的普适性。

大间隔分布机的核心思想——与支持向量机最大的不同——就是尝试同时最大化间隔均值和最小化间隔方差， 对于解决不同问题的 LDM算法，相应的我们会选取不同的方法求解。 例如，使用坐标下降法对核大间隔分布学习机进行求解，使用平均随机梯度下降法对大规模线性核大间隔分布学习机进行求解。

大间隔分布学习机这一概念提出到使用的时间周期并不长，因此在这方面的研究成果和论文并不丰富，本文仅按照基础的大间隔分布机的核心思想和运算理念进行探讨和实践，若有不足，欢迎学者、老师指正。

1.4 胰腺癌的研究意义

本文研究的重点是机器学习领域较为前沿的算法——支持向量机和大间隔分布机，对于理论知识的研究将在下一章节详细指出，本文证明两个算法严谨性的策略是使用胰腺癌数据作为训练样本和测试样本， 对于这类医学领域的现状和情况在本小结加以简单介绍。

据国家癌症研究所报道，胰腺癌作为癌症相关死亡的主要原因之一，2006年至2012年的5年总体生存率为7.7%。 其预后不佳的原因之一是非典型症状使早期诊断具有挑战性。 该疾病在大约80%的患者诊断时被认为是局部侵袭和转移扩散到其他部位。 90%的胰腺癌是胰腺导管癌，不能在早期诊断，因为它通常是无症状的。此外，胰腺癌的高异质性也阻碍提取高度特异性的生物标志物进行早期检测。 即使是常规化疗也不能有效的延长终末期病人的生活时间。[11] 因此，对胰腺癌的早期检测和预后的准确策略对完整的医学方面治疗有着十分关键的影响。

然而，通过生物化学实验的经典方法难以发现癌症的血清或尿标记物。考虑到转录组数据提供了相对完整的背景，并且在描述癌症的生物学特征方面具有明显的优势，越来越多的研究人员开始采用基于机器学习的特征选择方法来寻找癌症的制造者基因。两种相对简单的特征选择方法是折叠法和T检验假设检验。此外，支持向量机递归特征消除（即SVM-RFE）非常受欢迎。

在本文中，我们提出了一种通过支持向量机递归特征消除(SVM-RFE)和大间隔分布机递归特征消除（即LDM-RFE）来找到可用于胰腺癌的临床生物标志物的方法，其不仅具有高精度，而且还将具有更好的泛化表现。这项工作的目的是提高特征选择方法的性能，并确定一组基因，其表达模式可以准确区分胰腺导管腺癌和正常胰腺组织。[12]

在我们的实验中，我们加强了RFE的过程，以实现更好的性能。

本文采用的数据集GSE15471来自GEO数据库，39对胰腺导管癌和相邻对照胰腺组织。

通过算法和相关医学领域的寻找生物标志物的方法，发现六种生物标志物 MMP7， MMP12， ANPEP， FOS， SFN， IL6， A2 M与胰腺癌患者的存活率密切相关，其编码的蛋白质可以分泌到尿液中，这也许会为肿瘤学家提供一些十分有用的线索。

1.5 论文的研究内容

总结绪论部分的内容，主要分为四部分。

第一部分介绍了机器学习领域的起源和意义，以及背景和前进方向。

第二部分介绍了支持向量机（SVM）的起源以及大体框架，以及未来发展的趋势。

第三部分介绍了本文的核心算法——大间隔分布机（LDM），通过支持向量机的思路框架的更改以及另一核心问题的探讨，引入了这个新的算法。简要介绍了背景以及它的适用领域。

第四部分介绍了本文所运用的数据来源——医学领域上的胰腺癌研究，验证了大间隔分布机算法的实用性和可行性。

下面将本论文的具体工作流程加以汇报：

1.理论学习过程。通过中国知网， 吉林大学图书馆，维普等等多种资源获取途径，下载了有关机器学习领域内的支持向量机和大间隔分布机的相关文献，期刊以及各种会议，了解了机器学习，支持向量机的内含功能和发展状况， 整理和总结了新的算法——大间隔分布机的相关知识， 熟悉了有关算法核心部分（递归特征消除），学习了 Matlab， Python两大软件的基础使用。

2.具体实现过程。 通过初步的理论学习，我针对具体要研究的算法的公式进行了详细的推导和证明，证明了LDM和SVM的区别，通过数据集的分析和测试总结出两种算法的区别， 最后将预测结果的准确率做出计算并给出最后的可视化图形。

3.结果分析过程，由于使用了胰腺癌数据，我们仅仅通过数据得到的预测值需要反馈到实际应用部分中。因此，在了解了相关的生物医学背景后，将对应的基因名称和数据量加以总结分析，最后得出结论——到底有哪些基因影响着胰腺癌的发病几率。

4. 总结展望过程。总结本论文研究过程中的不足之处并为接下来的工作发展给出合理的展望。

下面将总体的论文研究内容做简要的分析和说明。

本论文是基于大间隔分布机递归特征消除算法的实现和应用，论文主体是根据大间隔分布机这一算法的证明和实践。

第一章，绪论。分析机器学习，支持向量机，大间隔分布机和胰腺癌数据的研究意义和研究现状。 并在最后提出本论文

的内容框架。

第二章，数据与分析方法。对本论文所使用的数据集进行阐述，并介绍了数据的处理过程，得到的去噪后得到数据集。并将所运用的两大算法进行简单的公式推导和证明，最后简要介绍使用的语言和编译器。

第三章，递归特征消除方法的实践与应用。将两大算法加以递归特征消除这一手段进行计算和分析，相互对比，得到最后的数据预测的准确率。同时，引入随机检验，T检验两种方法加以对照，从广义角度来证明大间隔分布机的优越性。

第四章，结果分析与总结。通过有关生物领域的知识来分析最后得到的数据内含的实际意义，并为肿瘤学家的研究提供方便。

第五章，总结与展望。对论文的主要工作做一下总结，提出本论文可能存在的不足之处并对未来的研究给出合理的展望。

指　标
疑似剽窃文字表述

1. 4. 总结展望过程。总结本论文研究过程中的不足之处并为接下来的工作发展给出合理的展望。
2. 研究意义和研究现状。 并在最后提出本论文的内容框架。
   第二章，数据与分析方法。对本论文所使用的数据集进行阐述，并介绍了
3. 总结与展望。对论文的主要工作做一下总结，提出本论文可能存在的不足之处并对未来的研究给出合理的展望。

3. 第2章数据与分析方法

总字数：9380

相似文献列表　文字复制比：3.2%(303)　疑似剽窃观点：(0)

1	基于SVM的大间隔分布学习 李元锦 - 《大学生论文联合比对库》- 2017-06-01	2.6%（241） 是否引证：否
2	53130928_宋莹莹_计算机科学与技术_基于机器学习方法的海洋温跃层的深度计算 宋莹莹 - 《大学生论文联合比对库》- 2017-06-02	0.7%（65） 是否引证：否

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第2章数据与分析方法

2.1 数据介绍

正确选取本研究中所用到的数据集，有助于我们在接下来的研究过程中准确的获取计算结果。

我们从胰腺导管癌和相邻对照胰腺组织中选择39个配对数据的GEO数据集GSE15471作为训练数据，这是基于平台Affymetrix人类基因组U133 Plus 2.0阵列（GPL570）测量的。

填充空和清除探针值后，将整个基因探针ID转换为基因符号。GPL570是最常用的平台之一，大多数GEO癌症数据集是基于此而进行的。

因此，我们选择GSE15471作为培训数据的目的是确保模型的普遍性和泛化性能。由于考虑到模型的广泛性，我们的模型为了消除不同平台之间的测量误差，并且可以保持在将来用于不同癌症的更多不同数据集时仍然具有优异的性能。

同样的，对GSE15471的部分数据我们可以作为训练数据，为了保证模型的严谨性我们同样适用GSE15471的剩余部分作为测试数据。将模型通过本数据集进行检测，有助于发现模型的弊端，对于未来的调试参数和模型的改进打下基础。

下面简单介绍数据的来源（如图2-1）并加以展示。

注：.disease和.normal文件表明癌症基因数据和正常基因数据，由于没有合适软件打开，则为了查看数据，就使用notepad++软件读取打开。（如图2-2,2-3）

图2-1 .GEO数据获取界面（GSE15471）

719.63 675.4 1160.7 1141 1150.3 1180.3

117.98 113.66 75.882 69.138 98.388 102.34

713.39 673.16 150.73 168.76 268.34 355.91

图2-2 . GSE15471pancreatic tissue39.disease部分数据

273.1 273.17 748.6 780.17 726 654.85

114.51 113.57 63.681 64.163 61.096 65.592

409.93 376.65 74.61 71.628 50.145 77.483

图2-3 . GSE15471pancreatic tissue39.normal部分数据

2.2 数据处理方法

2.2.1 数据的处理

如上两张图所示，具体数据矩阵的维度均是20128*39。

机器学习算法应用的第一步，就是处理数据——将大量数据中表达量最为明显留存，将部分错误数据，噪声数据通过合理的算法进行忽略，剩下最“正确”“有用”的部分数据。

在本论文中所采用的数据为医学领域的胰腺癌基因表达数据，我们将采用鉴别差异表达基因的思路将数据进行处理。

首先，我们通过最为常见的T检验，将数据进行第一步处理——去噪处理。

T检验是使用数学领域内最常用的证明差异显著程度的方法——T分布，来检测和推导差异是否存在，和存在的程度大小。即比较两组数据的区分度。

其次，我们选择了可接受的FDR低于0.01的差异表达基因作为信息基因。此外，我们通过计算癌症和正常组织之间的倍数变化（FC）水平来测量所选择的基因，并且通过其FC是否小于阈值0.5或大于阈值2来获得下调或上调基因。

在这里，简要介绍一下上面提到的倍数变化（FC）以及FDR的数据处理方法。

差异表达分析是识别疾病基因的常用方法，最为常用流行和实用的是Fold change。 将数据做Log2处理，这样做的目的是可以更加准确的判断基因是否差异表达。使用起来十分方便简单，不需要大量的数据推导即可使用。

笼统的说，Fold change的公式可以写成

（1）

在本文中，针对于类别“癌症”和类别“正常”之间的基因倍数变化被定义为：

（2）

这里的是表达式癌症样本基因g的表达值，同理，是表达式正常样本基因g的表达值，n代表配对样本的数目。

错误发现率（ false discovery rate， FDR）简单地说就是错误拒绝。 本文中使用FDR将一开始得到的拒绝零假设的部分（p-value）使用多重假设检验来矫正错误得到（q-value）。FDR的具体公式内容就是错误的个数占整体个数的比例期望。

（3）

总结一下上述过程，我们首先下载的数据是GEO数据平台下的GSE15471数据集，通过T检验去噪，FDR ，去除奇异点，Fold change筛选最终得到差异表达量较为明显的基因数据1585*78（一半为对照组）。流程如图

图2-4 数据处理流程图

2.2.2 算法的推导与证明

1. 支持向量机算法核心推导与证明

这部分会用最精简的语言来描述支持向量机的工作原理，附加相应合适的公式推导。

D是样本集，我们的目标是找到一个超平面，使得“+”和“-”有明显区分度的分开，但是由于我们可以找到无数个划分超平面，因此我们需要找到哪一个算作最优的呢？

图2-5 两类样本点被多条分类超平面分隔开

从最直观的角度来说，“正中间”的超平面是我们所希望看到的。原因很简单，虽然我们是从人类的肉眼直观角度来决定的，但是我们的第一感觉必然是有它的道理——因为该超平面对样本的“容忍度”最好——也可以说它是最“宽宏大量”的。在诸多样本中，“扰动”往往成为一个测试“容忍度”最好的手段。

下面，我将通过核心公式推导来简单证明支持向量机的工作原理。

图2-6 超平面划分两类点

样本空间中，假设划分超平面H可将样本准确分开，表示为：

（4）

其中，w是法向量(H方向)，b是位移(H与原点距离).

这时，有两个超平面，它们同时平行于H，并且使离H最近的正负样本刚好落在上面：

（5）

对于其他样本点都可以直观的看出分布于 H1，H2的远端，满足约束：

（6）

根据空间中距离公式：

（7）

空间中原点到超平面H1的长度：

（8）

同理，空间中原点到超平面H2的长度：

（9）

因此两个超平面距离为：

（10）

它被称为间隔（margin）。

我们的任务是最大化最小间隔，即求：

（11）

等价于：

（12）

这就是支持向量机SVM的基本型。

对不等式约束的条件极值问题，可用Langrange方法求解，制造规则是：用约束方程乘以非负的Langrange系数，再从目

标函数中减去。即

（13）

其中。

我们要处理的规划问题就变成了

（14）

直接求解上式较困难，通过对偶问题来解决，

（15）

对上式求w,b的偏导，并令偏导等于0.

（16）

将公式（16）代入（13），得到：

（17）

最后得到（12）式的对偶问题：

（18）

需解出上式中。（12）式有不等式约束，因此上述过程需满足KKT条件：

（19）

既然说到了KKT条件，在这里就简单介绍下：

KKT约束的意义：

若一个样本是支持向量，则其对应的Langrange系数不为零；

若一个样本不是支持向量，则其对应的Langrange系数一定为零。

既然原问题的对偶问题公式已推导出来，下面来求解 ：

要解决的是在参数上求最大值的问题，都已知，这是个二次规划问题，解决这类问题比较高效著名的算法------SMO。

SMO基本思路：每次选两个变量并固定其他参数，根据前面约束，将用表示，并代回（18）中，求上的极值，确定下后，循环上述过程。

由上述公式推导得出，接下来求的值：由于对所有支持向量都有成立，将代入上式，有

（20），

其中为所有支持向量的下标集，可选择任一值代入（20）算出，但实际做法是求平均值：

（21），

至此，我们求出了原始公式的所有变量的值。证明完毕。

我们仅讨论关于支持向量机的训练样本线性可分的情况，样例存在线性不可分的情况往往通过特征映射到高维空间，转化为可分的情况。

总结的说，这部分只是将支持向量机SVM的大体框架的推导与证明罗列于其上，参考的书目会附在后文的参考文献部分。

通过我们的证明，我们可以发现支持向量机有几大独特的优势：

（1）创造性的提出了“支持向量”这一概念，支持向量的地位十分重要，可以说，在SVM的分类决策中有这举足轻重的地位。

（2）支持向量机是可以通过严谨的公式证明，但却可以说的很“幸运”的有了这一坚实又强大的理论背景，它对小样本学习“钟爱有加”。因为在我们的证明过程中基本都没有涉及到概率等传统的统计学计算方法。正是因为它脱离了“归纳总结”“演绎推理”这个“旧”过程，直观的感受到只是将训练数据作为输入，学习好模型后，测试数据的测量作为输出，让人们感受到过去的分类问题得到了很大程度的简化。

（3）最后结果的重点取决于这些“少数的”支持向量，而将“真理”掌握在“少数人”手中的最大好处就是可以帮助我们过滤掉大量本身就不需要我们去研究费时的部分，正是因为这个优点，让这个方法更简单，实现起来更方便。

同时，在证明过程中，我们也同样能看到有两个十分严重的缺陷：

（1）支持向量算法“钟爱”小样本数据，对于大规模的“大数据”难以广泛应用——原因之前提到过，就是通过二次规划(QP)求矩阵，矩阵越大，内存消耗越大，因此，人们广泛的了解到SVM的最大弊端就是无法处理大数据。

（2）在我们证明的开始提到过，支持向量算法解决二分类问题方面是“一把好手”，但是在处理多方面问题的时候依旧存在缺陷尽管我们提出了“SVM决策树”等等，但是由于技术的不成熟以及其他算法对于多分类问题的优势所带来的冲击，在解决多分类问题上，SVM还有很长的路要走。

2.大间隔分布机算法核心推导与证明

大间隔分布机算法在支持向量机算法的基础上进行了改进，改进的部分就是在关注重点放在了整体的间隔分布而非单个点的间隔。

下面我将通过公式推导和对比进行大间隔分布机算法的思路说明。

上一小结说到，支持向量机在对于线性的训练样本有着独特的优势，但是将样本扩展到非线性层面上的时候，从训练样本的角度来说，就已经不能实现“完美分割”了，因此，在支持向量机算法中，人们提出了“硬间隔”，“软间隔”这两个概念。

现在轮到非线性样本，解决这种类型问题的方法就是将它们的特征映射到高维空间中，令表示将映射后的特征向量，则划分出超平面的函数：

（22）

对应上述线性样本的公式：

（23）

对偶问题：

（24）

上式涉及到计算，它是样本映射到特征空间的内积。由于计算较困难，因此提出了核函数的概念：

（25）

前面一直假设训练样本在样本空间或特征空间中线性可分，实际很难确定合适的核函数，因此引入软间隔。

支持向量机要求，样本不存在划分错误的情况，将这种分类间隔称之为“硬间隔”。

相对应的，软间隔允许某些样本不满足约束条件，同时，不满足约束的样本越少越好，优化目标可写为：

（26）

其中>0是常数，是0/1损失函数：

（27）

引入松弛变量（用来表示某样本不满足约束的程度），，我们将（26）式改写为：

（28）

上面（28）式，就是我们所谓的软间隔SVM。

这一部分，我们会清楚的了解到，支持向量机（无论软间隔，硬间隔）考虑的都是单个点的间隔，但如果出现下面所述的情况，我们会发现，整体的间隔分布也对最后的最优分类面产生巨大的影响。

图2-7 大间隔分布机的优化边界分布与传统优化最小边界的区别[13]

如上图所示，显示了一个相较支持向量算法证明部分的示例图更为复杂的情况，即存在异常值或嘈杂的数据点（最靠近红色圆圈的一个蓝色三角）。

如果我们坚持优化最小余量，在图2-7中，分类器将几乎以异常值或噪声数据点为主。如果我们试图优化边际分布，那么异常值或嘈杂数据点的影响就会“自动消失”。换句话说，优化边际分布的分类器将比优化最小边界的分类器更稳健。这也就是大间隔分布学习机的起源。为了解决存在异常值或嘈杂数据点的样本分布。

再回到支持向量机算法的两大“间隔”部分，我们通过上面的公式推导已经清晰的得出结论，硬间隔和软间隔的SVM确实试图优化单个边界。本文所使用的算法的中心思想就是通过优化边界分布来取代单一优化最小边界分布，同时保持其解决方案策略的其他部分不变。

因此，大间隔分布机提供了一种合理的方式，通过简单的适应性来获得更强大的学习方法。

要完成大间隔分布机的学习过程，我们需要了解如何优化间隔分布。Reyzin和Schapire [14]提出最大化间隔均值或中值间隔均值，并且在间隔均值或加权组合边际的最大化方面也有一定的作用。但是，这些理论并没有的到准确的证明。Gao和Zhou[15]证明了它们的间隔定理的另一种形式，它揭示了平均值或中值平均值是不够的，并且为了表征边际分布，重要的是不仅要考虑间隔均值而且还有间隔方差。这为算法设计提供了新的方向，将均值最大化，同时将方差最小化。 最近的一些Boosting研究[16，17]证实了这一观点的正确性。

理论的提出已经完成，下面我们就开始通过公式来推导和证明大间隔分布学习机。

首先，我们先将大间隔分布学习机的两个核心统计量——间隔均值和间隔方差加以表示：

X 表示第i列为 ϕ(xi)的矩阵, X = [ϕ(x1)...ϕ(xm)], y = [y1,...,ym]⊤ 是一列向量, Y 是m × m 阶以y1,...,ym 为对角元的对角矩阵。

间隔均值：

（29）

间隔方差：

（30）

大间隔分布学习机的目的是最大化间隔均值并且最小化间隔方差。

我们同样的分成两大“间隔”进行公式对照。

线性可分的情况下，上述方法可以顺利得到硬间隔LDM：

（31）

其中λ1，λ2是参数，目的是将间隔方差和间隔均值进行相对应的平衡。当两者均为零的时候，就转变为硬间隔SVM。

对比与非线性可分的情况下，软间隔大间隔分布学习机：

（32）

同理，当两个参数均为零的时候，软间隔大间隔分布机就是软间隔支持向量机既然存在两种“间隔”的学习机，最常使用的是哪个呢？

事实证明，从支持向量机的角度来说，软间隔带给数据分类的结果一个“容忍度”，正是因为有这个“容忍度”，很多情况下

的数据嘈杂和扰动不会给最终的预测模型带来很大的影响。由此可见，软间隔的适用领域更加广泛。因此我们后续的推导优化算法就默认使用软间隔的大间隔分布学习机。

大间隔分布学习机有两种——核大间隔分布学习机和大规模线性核大间隔分布学习机。两者分别使用坐标下降法（Coordinate Descent，CD），以及平均随机梯度下降法（Average Stochastic Gradient Descent，ASGD）作为核心思路进行求解。

本文使用前者——坐标下降法解决核大间隔分布学习机。

前文提到，解决二分类问题就是一个二次规划问题（QP），我们通过上述（29）（30）代入（32）可以得到

（33）

式中的w*是我们所要求解的第一目标，求出它的最优解被认为是一大难题。

通过寻找到多方资料，我们找到一篇论文针对于这个点进行了总结，在此不加赘述，我们所需要求得的w*可以通过以下内容表示：

（34）

在经过核矩阵以及拉格朗日乘子法的处理，求偏导，实现了问题的对偶形式：

（35）

在坐标下降法中，循环的方式是将一个变量进行最小化处理，并且保证其它变量的一致性。 最后求解子问题：

（36）

令

得到

（37）

将得到的β*进行代入计算，得到最终的

（38）

最后，通过如下公式进行测试变量标签的预测

（39）

伪代码如下[10]：

Algorithm Kernel LDM

Input:Data set

Output:

Initialize

while not converge do

for do

end for

end while

输入：数据集X，，是调整间隔均值和间隔方差的平衡参数，C是约束的参数。

输出：

流程：

1. 初始化4个变量

2. 循环开始，循环条件为不收敛

3. 依次给梯度赋值，并更新每个值，其次，将更新好的值做选择运算，最后将更新。

4. 循环结束

上述部分是结合周志华教授发表的两篇论文进行精简和整合，得到的大间隔分布机的证明。上文证明只将本文所用的大间隔分布学习机的算法进行了简要证明，并不代表现有所研究的LDM算法仅此而已，其他证明部分在周志华教授的论文中也有所体现，在此不做介绍。

2.3 数据处理软件

（1）Matlab

Matlab软件为本文提供了大量的便利，不仅在图像绘制部分，在第一步处理数据的结果上也有这至关重要的作用。

该软件的功能对于大学生来说，可以简单的理解为“建模”必备软件。

因为Matlab在使用过程中，无论从官方所给出的自带函数，还是数据图形界面而言，都对于用户十分友好。可以无差别的对待刚开始接触数学物理的学生，也帮助资深的程序员进行十分方便的数据处理和图形绘制工作。

从编程语言来说，与现阶段流行的C，C++，Java，没有可比性，对于初学者而言可以说是上手十分简单，即使遇到不会使用的情况，在命令行窗口使用help命令，就会很容易的得到大量的帮助。在编译结束后，只需一个按钮就可以将程序启动运行，不仅提高的工作效率，还将大量的数据处理作为“黑箱操作”，深受广大使用者的喜爱。

同时，绘图部分也是Matlab的强项。我们经常食用的矩阵，数据量等等都可以通过各种2D,3D图来表示出来，操作指令仍然十分简单。并且，在图像的处理部分也有着相当大的长处，因此，本文使用的Matlab对我们的实验进展有了很大的影响。

[18]

（2）Python

本文采用Python语言进行大部分数据的采集，处理，分析和调用计算。Python是现在最为流行的语言之一。TIOBE 5 月编程语言榜见下图2-9。对应于前十名的发展趋势对比见图2-10。本文使用的科学计算部分就是最为实用的一部分。Python近几年发展迅速，虽说在流行度和普及度无法与Java相抗衡，但是从发展前景看，已经多年呈上涨趋势，并且随着版本的更替，细节的修复，论坛的完善等等，一切都表明着Python的未来是可期的。[19]

图2-8 TIOBE 5 月编程语言榜

图2-9 Top 10 编程语言 TIOBE 指数走势(2002-2018)

在本文，Python使用到了最大比重就是科学计算部分。[20]

Python对于其他软件的一大优势就是在程序库部分的领先。谈到科学计算，最常用的程序库就是NumPy, SciPy, Matplotlib这三个。NumPy是相当于一个微缩版的matlab(处理，存储，矩阵运算部分)，包括数组，线性代数，矩阵运算，矢量运算方面有着独特的函数处理流程，对于本文的代码部分的编写起到了至关重要的作用。SciPy是Python常用的工具包，它经常和NumPy使用，用于处理统计学问题，线性代数，傅里叶变换等，还在积分，优化，插值这方面有所涉及。Matplotlib相当于matlab的作图功能（2D部分），与matlab类似，操作者可以使用简单的几行代码实现大部分绘图，例如散点图，折线图，条形图，直方图，饼状图，箱形图等等。

最后，关于科学计算的机器学习部分，Python的Scikit-learn工具包是最优秀的典范。从交叉验证，到支持向量机的使用，只需要轻松的几个函数，就可以实现源代码上千行的内容。现在的Scikit-learn[21]已经作为先行者开创了语言自带机器学习函数的先河，例如现在google公司开发出的TensorFlow就是这个领域的佼佼者。[22]

1.4 Libsvm和LDM算法库

下面，将介绍一下本文使用的两个算法包，这两个算法包内含支持向量机和大间隔分布学习机的的计算过程。

1.1.1 Libsvm

提到SVM算法，实现算法的最简单步骤就是使用Libsvm这个软件包。

Libsvm是台湾大学林智仁教授等人设计开发出来的一款SVM多功能包，使用者只需要在调用所需要的几个函数之前处理好数据格式即可，这个软件包的内容包括解决分类问题，回归问题，分布问题等问题的诸多函数，并且内置多种核函数用于解决非线性问题的映射部分。[23]

Libsvm实现了“小”“快”“灵”三大优势。程序仅有1.82M，运算速度快，参数使用简单，最优秀的一点就是自诞生开始，就是开源，并且对于它的更新一直没有断过。

在实际应用上，Libsvm有着其强大的可移植性[24]。在下载得到的Libsvm包里面，有如下内容：

图2-11 Libsvm工具包内置6大文件夹

我们看到，这六个文件夹清晰的表明了Libsvm的完整性和扩展性。

首先，Libsvm有支持三大主流语言——Java,Matlab,Python。通过三大平台的科学计算是最多的。因此，在这三大平台拥有独立的功能包使得即使是别的语言想调用只需要写出简单的接口程序即可，并不需要大量的翻译代码工作。

其次，svm-toy这个可视化工具对于初学者来说十分友好，它展示了数据是如何进行传入，操作，训练等功能，以及分类的平面，准确度都通过可视化的方式来表现出来。

最后，剩下的两个部分就是我们调用的接口部分，内含多个.dll文件和.exe文件，对于Windows系统的用户十分友好，并且内含示例数据集和文件样本等。

Libsvm对于本文的帮助巨大，只需将数据进行操作符合合适的输入格式即可，如下图所示，需要将“标签—样本序号1—数据1-样本序号2—数据2-样本序号3—数据3-……”格式套用进去即可

图2-12 Libsvm—python数据格式

调用svm_train函数进行训练，使用svm_predict进行预测。使用方便，计算准确，是不错的软件包。

1.1.2 LDM

相对于上述Libsvm工具包，LDM工具包就没有这么多的优势了。

LDM工具包里面的内容只限于Matlab使用，内含多个核函数，大部分的操作只能在Matlab上进行，包括传入数据和传出预测值。由于大间隔分布学习机算法较为“稚嫩”，因此对于LDM工具包在网络中也并不容易找到。

下图是LDM工具包的简单内容;

图2-13 LDM工具包内容

由图可见，LDM工具包里面主要包含了Matlab的调用接口函数—trainLDM.c和predictLDM.c，这两个函数调用了LDM.cpp这个主要的运算文件。

值得一提的是，在调用函数的过程中，选择核函数等参数有着较为简单实用的说明。

图2-14 LDM工具包内置使用方法

两种求解模式——坐标下降法，平均随机梯度下降法。

对应的四种核函数：线性，多项式，径向基函数以及sigmoid。

总体来说，LDM工具包由于并不成熟，使用起来较为困难，只能通过Matlab进行运行操作，对于本论文的实践部分产生

了较大阻碍。

指　标

疑似剽窃文字表述

1. 第2章数据与分析方法
   2.1 数据介绍
   正确选取本研究中所用到的数据集，有助于我们在接下来的研究过程中准确的获取计算结果。
2. 如果我们坚持优化最小余量，在图2-7中，分类器将几乎以异常值或噪声数据点为主。如果我们试图优化

4. 第3章递归特征消除方法的实践与应用

总字数：2893

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第3章递归特征消除方法的实践与应用

3.1 特征选择

本章的开始，我们先介绍特征选择的基本概念。

特征选择的核心思想是通过某种搜索算法和特征选择模式来决定如何将大量的特征简化成“识别度”最高的特征。[25]

顾名思义，特征选择就是将数据量较为庞大“特征”筛选出最能体现“个性化的”特征。引入这个概念的起因是出现了“维数灾难”——特征过多的情况下，分类器的效率出现断崖式下降。人们分析出现这个情况的原因，主要是因为在特征过多的情况下，特征的“无关”和“冗余”将分类器原本优秀的测量标准和方法产生了大量分歧，导致算法的“过拟合”。举个例子，如果测试的样本是一条狗，我们在特征学习的过程中学习到了狗特有的几种“特征”——耳朵，鼻子突出等等。我们会发现在这些特征的基础下训练的模型对于测试来讲有着不错的准确率。但是如果再引入其他的“具体的”特征——体型大小，毛发长度，眼睛颜色等等，我们会发现实际上就把“狗”这个概念过于具体化，而没有考虑到品种不同的差异，或者其他类似动物所具有的相同的特征的问题。这就是“过拟合”的危害。因此，特征选择将剔除掉“冗余特征”和“无关特征”，避免因为特征的数目过多导致的分类器运算时间过长和准确度断崖式下降。

因此特征选择的地位越发重要。

在本文算法的实践过程中，选取了特征选择算法的三大类型之一——包裹式（wrapper）。[26]

其他的两大类型在此没有用到就不再赘述。包裹式特征选择方法与其他两种的区别主要是在评价方法上面。从原始数据集导入数据，在进行搜索的过程中每训练出一个学习器，就进行测试评价，循环所有的数据，每个循环都要更新学习器，学习器的能力一旦出现下降就进行调整策略重新选择上一特征。循环结束时的学习器性能决定了这个方法的最终结果。包裹式特征选择策略相比较过滤式（filter）更注重最后的学习器性能，而不是只关注每一次的学习器性能。

图3-1 特征选择流程

上图是特征选择的主要思路和流程。

1. 通过选择的不同类型的算法进行搜索特征，为后续的函数提供第一步筛选出来的子集。
2. 通过不同方式特有的评价函数进行上述生成数据特征的水平。
3. 针对于前一步评价函数得到的结果，进行选择性反馈，如果效果较差，设置阈值，进行反馈更新，重新生成特征子集
4. 最后，经过前三步的循环迭代，得到最终的模型，用于未来数据的处理与预测。

1.2 递归特征消除思想

递归特征消除（Recursive feature elimination，RFE）其实就是一种寻求最优解的贪心算法。

在2003年，Guyou提出了较为新颖的特征选择算法——通过迭代特征删除的后项搜索方法的结合。[27]

递归特征消除算法的主要思想是先通过算法确定所需要的模型，将这个模型经过多次重复构建，每次的构建得到最好（或者最差）的特征，将这个特征放到另一个序列，然后在原特征序列中删掉这一特征，并开始重复上述过程，通过模型的重复构建和序列的多次排序，每次排序都会删除掉一个特征，直到剩余为零。同时，每次删掉的特征也构建好了一个序列，或者说，被“消除”的特征的顺序，就是最后的最优特征的排序。

在Python的语言环境中，前文提到的Scikit-learn工具包内置feature_selection库，这个库里面有RFE类，可以用于递归特征消除，同时还提供了RFEV，通过交叉验证对已有的特征操作排序。

本文并没有使用这个类库，而是将RFE的流程完整的通过代码的形式体现出来。本论文的核心思想就是使用SVM-RFE和LDM-RFE进行模型构建并得到特征序列排行，通过这两部分的排行，使用SVM分类器和LDM分类器进行比对和验证。

1.3 支持向量机和大间隔分布机的递归特征消除的实现.

这部分将讲述本文使用的核心算SVM-RFE和LDM-RFE算法。

3.3.1支持向量机的递归特征消除算法

SVM-RFE是基于支持向量机的递归特征消除算法。

在使用递归特征消除方法之前，我们要明确支持向量机的“特征”到底是什么。在SVM中，通过维度对应数据集的数据或者特征，而在SVM的超平面上的权重w，则是将其看做这个特征的排序标准——进行数据处理，将w取绝对值。由此我们得到了“特征”，我们就可以顺理成章的通过权重进行排序。在这里我们选用重要性有高到低的顺序。RFE的流程可以概括为从多个w的排序中，每次删除模型的w绝对值最小的，循环，直到w集合为空。最后得到的w排序即为特征序列排行。[28]

这里注明一下特征的评分排序准则的分数定义：

（40）

下图是SVM-RFE的算法伪代码。

图3-2 SVM-RFE算法伪代码[29]

在此简单说明一下算法流程：

输入：训练样本X，对应的标签Y——y取（-1,1）

输出：经过SVM-RFE算法得到的特征序列

算法过程：

1. 初始化训练特征集合以及最终要排序的序列
2. 开始循环体部分，循环条件为原特征集合非空
3. 首先，获取到带有结果标签的训练特征集合X
4. 其次，通过SVM-train训练分类器的值
5. 通过公式训练得到
6. 使用进行分数计算
7. 取平均结果的最小的c对应的
8. 将这个对应的特征存入集合中
9. 在原本的集合中去除刚才的最小特征
10. 开始循环，直到原特征集合为空，算法完毕

以上，就是本文所使用的SVM-RFE算法的介绍与实现。根据上述思路和对应的代码，完成了对于LDM-RFE的对照和实践。

3.3.2大间隔分布机的递归特征消除算法

对照于上述的SVM-RFE，本文所要证明的方法就是这部分内容——LDM-RFE.

算法的思路和构架与SVM相同，使用相同的特征提取技术，只是在关注点上变为了更注重间隔均值和间隔方差。

评判策略与SVM-RFE不同，使用作为评分的关键要素。

具体的LDM-RFE伪代码如下图所示：

图3-3 LDM-RFE算法伪代码

在此简单说明一下算法流程：

输入：训练样本X，对应的标签Y——y取（-1,1）,LDM的算法传入参数 ， ，

输出：经过LDM-RFE算法得到的特征序列

算法过程：

1. 初始化训练集合以及最终要排序的序列
2. 开始循环体部分，循环条件为原特征集合非空
3. 首先，获取到带有结果标签的训练特征集合X
4. 其次，确定训练特征集合的长度
5. 之后，通过LDM-train训练分类器的值
6. 开启第二个循环体，用于计算评分，通过的相关组合进行计算，将结果放入集合中
7. 用变量存储代表每个迭代中具有相同最小特征得分的子集。
8. 最后使用代表特征选择过程后的最终排序特征列表
9. 去除原集合中评分最低的一个特征，循环操作
10. 直到原始集合为空，跳出循环，算法完毕。

由上述算法伪代码可以看出，LDM-RFE就是将大间隔分布机的特有方法的最后加入递归特征消除的部分，评分步骤较为繁琐复杂，但是思路清晰，并无偏差。

4.1 结果展示与评估

首先将总体代码的流程图表示出来

图4-1 整体算法代码流程图

通过Python和Matlab的代码运行，我们分别得到了SVM-RFE和LDM-RFE的基因序列（由于数目太多，只写出前15个基因的序号）

749945 322 295 60 4 1157 803 583 32 480 16 17 695

图4-2 SVM-RFE算法基因索引排行前15个（基因序号从1开始）

7 295 322 4 49 148 480 80 230 583 45 1009 158 12 691

图4-3 LDM-RFE算法基因索引排行前15个（基因序号从1开始）

由上图可以看到，在通过两种分类算法的递归特征消除的结果中，有一些基因同时进入了前15名，证明了这种基因在特征差异表达上的确有极大的“区分度”。

在得到基因的特征排序之后，我们使用支持向量机分类器和大间隔分布学习机分类器这两种分类器进行效果比对，将测试数据放入已训练好的SVM-RFE和LDM-RFE的模型中，计算结果，最终得到准确率。

准确率的公式我们使用如下公式来评估我们从实验中获得的基因列表的有效性：

（41）

在此注明，我们通过使用随机选取的基因排行，以及只利用T-test检验的基因排行得到了两个用于对照的数据排行，同样进行上述两大分类器的数据预测，下图可见到共四种预测排行（由于数据量过多，在此不做过多展示，因此只呈现前3个基因的表达量准确率）。

SVM分类器：

0.769230769230769 0.769230769230769 0.538461538461538 0.692307692307692 0.384615384615385
0.538461538461538 0.846153846153846 0.807692307692308 0.769230769230769 0.769230769230769
0.846153846153846 0.153846153846154 0.692307692307692 0.769230769230769 0.653846153846154

图4-4 SVM-random前200个基因排行的准确率统计（共60次）

0.923076923076923 0.923076923076923 0.923076923076923 0.923076923076923 0.923076923076923
0.961538461538462 0.961538461538462 0.923076923076923 0.884615384615385 0.923076923076923
0.923076923076923 0.923076923076923 0.807692307692308 0.923076923076923 0.961538461538462

图4-5 SVM-T-test前200个基因排行的准确率统计（共60次）

0.500000000000000 0.769230769230769 0.692307692307692 0.730769230769231 0.961538461538462
0.923076923076923 0.884615384615385 0.500000000000000 0.923076923076923 0.923076923076923
0.923076923076923 0.730769230769231 0.884615384615385 0.923076923076923 0.923076923076923

图4-6 SVM-SVMRFE前200个基因排行的准确率统计（共60次）

0.884615384615385 0.884615384615385 0.884615384615385 0.884615384615385 0.884615384615385
0.923076923076923 0.923076923076923 0.923076923076923 0.923076923076923 0.923076923076923
0.923076923076923 0.923076923076923 0.961538461538462 0.961538461538462 0.961538461538462

图4-7 SVM-LDMRFE前200个基因排行的准确率统计（共60次）

LDM分类器：

0.500000000000000 0.576923076923077 0.730769230769231 0.692307692307692 0.769230769230769
0.500000000000000 0.500000000000000 0.500000000000000 0.500000000000000 0.500000000000000
0.500000000000000 0.500000000000000 0.500000000000000 0.500000000000000 0.500000000000000

图4-8 LDM-random前200个基因排行的准确率统计（共60次）

0.500000000000000 0.846153846153846 0.769230769230769 0.807692307692308 0.807692307692308
0.500000000000000 0.692307692307692 0.769230769230769 0.769230769230769 0.807692307692308
0.500000000000000 0.653846153846154 0.769230769230769 0.730769230769231 0.769230769230769

图4-9 LDM-T-test前200个基因排行的准确率统计（共60次）

0.500000000000000 0.500000000000000 0.769230769230769 0.884615384615385 0.807692307692308
0.500000000000000 0.500000000000000 0.846153846153846 0.923076923076923 0.846153846153846
0.500000000000000 0.500000000000000 0.730769230769231 0.692307692307692 0.769230769230769

图4-10 LDM-SVMRFE前200个基因排行的准确率统计（共60次）

0.500000000000000 0.500000000000000 0.538461538461538 0.576923076923077 0.653846153846154
0.500000000000000 0.846153846153846 0.807692307692308 0.769230769230769 0.807692307692308
0.500000000000000 0.500000000000000 0.653846153846154 0.653846153846154 0.653846153846154

图4-11 LDM-LDMRFE前200个基因排行的准确率统计（共60次）

数据结果都是经过了60次的循环统计得到的准确率，包括4种基因排行，2种分类器。

上述的数据从直观的角度来说过于扁平化，我们通过折线图来表明数据的变化趋势以及方法优劣性的比较。

图4-12 SVM分类器下，4种方法的准确率曲线

由图可知，相比较而言的random(随机特征排序)和T-test(T检验排序)，很明显不如我们经过探讨得出的SVM-RFE以及LDM-RFE。

同时，在SVM和LDM对比的曲线图中，我们看到SVM在初期有些许优势，但到了中后期，LDM更趋于稳定，并逐渐赶超SVM。证明了我们的LDM-RFE算法有一一定的理论基础和实践意义。

图4-12 LDM分类器下，4种方法的准确率曲线

由图可见，在LDM分类器下，SVM和LDM的数据预测准确率不分伯仲，开始阶段LDM有优势，但到了后期，SVM的准确率较为平稳，更加有效。

由于在不同分类器上的环境不同，有些许差别同时也是可以接受的。

同时，我们对于SVM-RFE和LDM-RFE的整体正确性可以使用下图说明：

图4-13 SVM-RFE与LDM-RFE的数据特征排行得分对比

横坐标—SVM-RFE

纵坐标—LDM-RFE

内含的数据内容是通过两种方式得到的特征评分，我们看到大部分数据都是呈对角线分布，因此可以得出结论，在递归特征消除方法的使用条件下，SVM和LDM的相似度还是很高的。

4.2 生物学分析

由于本论文的实践部分是通过生物学领域的胰腺癌数据进行分析与测试的，将算法的理论和结果分析结束后，需要将对应的基因性质和基因背景简要说明，对于论文的完整性和严谨性而言十分重要。

对于我们所进行的数据特征排行的结果来看，并且结合我们算法进行准确率排行的结果加以辅佐，我们证明了数据排行前7名对区分胰腺癌和正常胰腺组织的基因表达了相差较大，因此我们可以将他们看做胰腺癌生物标志物的良好候选物。

这前7个基因的简要信息如下：

基因名称：MMP7，MMP12，ANPEP，FOS，SFN，IL6，A2M

图4-14 七大基因的基因具体数据

而具体的功能在这里也进行简要的介绍：

MMP7与胰腺癌和结肠直肠癌转移存在相关；[30]

MMP12可以分解几乎全部细胞外机制和血管壁成分；

ANPEP与上呼吸道感染有着密切的关系；

FOS基因家族编码亮氨酸拉链蛋白，此蛋白与细胞增殖、分化以及凋亡有密切联系；

SFN与良性乳腺腺上皮细胞瘤的形成有关，会抑制体外胰腺癌细胞的自我更新；

IL6编码的细胞因子可以在炎症中起关键作用；

A2M编码的蛋白可以抑制蛋白酶和炎症细胞因子

因此，在得到了实验结果后，我们通过了理论公式的证明推导和实际生物学领域的分析可知，我们得到的七个差异表达基因的组合可以作为胰腺癌的特有生物标志物。其中MMP7，FOS和A2M与患者生存率密切相关。它们编码的蛋白质可以分泌到尿液中，或许可以为临床诊断提供依据。

6. 第5章总结与展望

总字数：1404

相似文献列表　文字复制比：0%(0)　疑似剽窃观点：(0)

原文内容　红色文字表示存在文字复制现象的内容; 绿色文字表示其中标明了引用的内容

第5章总结与展望

5.1 工作总结

在机器学习领域内，支持向量机算法已经走过了好几个年头，它的改进与发展对于机器学习这一方向的前进有着重要的影响。

在2014年，周志华教授创造性的提出了大间隔分布机这一概念，将支持向量机的算法的一大缺陷利用改进算法核心——将单一间隔问题转化为整体分布的边际化间隔问题——将关注点重新放在了间隔均值和间隔方差。在业界产生了极大地反响。

目前，LDM算法仍然在将理论转化为实际应用的路上——从官方给出的工具包可以看出，现在的LDM仍没有走向成熟，但是它对于SVM的挑战力依旧存在。

本文就这一论点，总结了大间隔分布机的起源，诞生，直至推导公式证明部分，分析了它与SVM的联系与区别，最后给出了LDM的核心推导公式。理论不并代表一个学习器的成熟，因此本文采用了生物医学领域内的胰腺癌数据集进行验证。在验证的过程中，我们在特征选择方法的选取上采用了递归特征消除(RFE)，通过每次筛选出一个特征评分最差的进行数据特征

排序，最终得到经过SVM-RFE和LDM-RFE的数据特征排序，最后通过两种不同的分类器对已经排行好的数据进行预测分析计算准确率，用实际的用例来证明了LDM-RFE算法的正确性和合理性，并与SVM-RFE进行比较分析，发现两者效果相当，甚至在某一部分LDM更优于SVM。

论文的研究过程有如下特点：

1. 本论文采用GEO数据库——成熟的医学领域基因表达量的数据发布平台，通过准确合理的数据来源进行了科学的数据处理分析，尽可能的得到了差异表达基因，为后续的数据分析训练测试提供了强有力的保障，同时也满足了学术领域的论文规范的严谨性。

2. 本论文在使用机器学习领域较为成熟的支持向量机算法的过程中没有一味的不经思考的使用，而是把SVM作为基础，发展到近几年研究的新理论——大间隔分布机（LDM）进行研究和分析，通过合理的特征选择算法——RFE进行科学性的特征排序。总体的过程严谨科学，不存在弄虚作假。

3. 本论文在得到结论之后没有简单的结束，而是通过实验所得结论进行实际问题的深入研究，在生物学理论的背景下对实验得到的数据做出了生物学方面的分析与探讨，为生物领域的专家们做出了自己的微小贡献。

5.2 问题与展望

在研究结束的同时，本文也对自身没有讨论到的知识点做出以下的总结：

1. 大间隔分布机算法迄今为止并没有真正意义上的普及与应用，研究时间较短，因此在理论部分没有特别完整的体系，在未来的研究中需要进一步思考与总结。

2. 本文所使用的数据来源于GEO数据库，该数据库对胰腺癌的数据量并不充足，只存在两个GSE平台，本文选择了其中之一，因此数据的可说服性并不是十分强大。因此，希望在未来得到研究中能找到更加科学并大规模的数据库来对本实验进行更加准确的验证。

3. 在本文中，理论说明部分可以得出结论LDM在间隔分布较为特殊的时候优于SVM，但是本文所使用的数据集是否满足这一条件有待商榷，但是我们还是可以通过理论证明我们的结论的严谨性和科学性。

4. 在实际的生物医学领域上，本文做出的贡献仅仅是拿到数据，做出基因数据特征排行，而后续的理论研究需要储备大量的医学方面的知识才能完成，本人并未深度接触这一领域，因此对于本文的结论部分只是简单粗略的总结，难免会存在不少不足之处，希望在接下来的学习中对这一部分进行更深入的研究。

参考文献

[1] 甄盼好. 浅谈机器学习方法[J]. 网络安全技术与应用, 2014(1):176-177.

[2] 闵应骅. 计算科学的回顾与前瞻[J]. 自然科学进展, 2000, 10(10):877-883.

[3] 周宏仁. 信息化:从计算机科学到计算科学[J]. 中国科学院院刊, 2016(6):591-598.

[4] 余乐安, 李想. 大数据时代的计算科学与优化[J]. 国际学术动态, 2015(4):11-13.

[5] 刘乃文. 冯·诺依曼机原理的教学研究与应用[J]. 计算机工程与设计, 2006, 27(10):1831-1834.

[6] 曹健, 孙世宇, 段修生,等. 基于KKT条件的SVM增量学习算法[J]. 火力与指挥控制, 2014(7):139-143.

[7] 胡燕, 熊浩勇, 付香英. 线性可分文本的SVM算法研究与改进[J]. 计算机与数字工程, 2008, 36(3):18-20.

[8] 周晓剑, 马义中, 朱嘉钢. SMO算法的简化及其在非正定核条件下的应用[J]. 计算机研究与发展, 2010, 47(11):1962-1969.

[9] 史朝辉, 王晓丹, 赵士敏,等. 改进的SVM决策树分类算法[J]. 空军工程大学学报·自然科学版, 2006, 7(2):32-35.

[10] Zhang T, Zhou Z H. Large margin Distribution Machine[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM,2014:313-322

[11] 马臣, 姜永晓, 刘曙正,等. 中国胰腺癌发病趋势分析和预测[J]. 中华流行病学杂志, 2013, 34(2):160-163.

[12] 李丹, 余涛, 曾智,等. 基于Oncomine和GEO数据库分析S100P在胰腺癌中的表达及临床意义[J]. 安徽医药, 2017, 21(12):2180-2184.

[13] Zhi-Hua Zhou. Large Margin Distribution Learning. In Artificial Neural Networks inPattern Recognition, 8774: 1 11, 2014.

[14] L. Reyzin and R. E. Schapire. How boosting the margin can also boost classifier complexity. In Proceeding of 23rd International Conference on Machine Learning, pages 753–760, Pittsburgh, PA, 2006

[15] W. Gao and Z.-H. Zhou. On the doubt about margin explanation of boosting. Artificial Intelligence, 199-200:22–44, 2013. (arXiv:1009.3613, September 2010).

[16] C. Shen and H. Li. Boosting through optimization of margin distributions. IEEE Transactions on Neural Networks, 21(4):659–666, 2010.

[17] P. K. Shivaswamy and T. Jebara. Variance penalizing AdaBoost. In J. ShaweTaylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 24, pages 1908–1916. MIT Press, Cambridge, MA, 2011.

[18] 苏中义. MATLAB简介[J]. 上海电机学院学报, 2003, 6(4):82.

[19] 孟岩, 闫辉, 朱海燕. Guido谈Python的现状与发展[J]. 程序员, 2007(7):26-27.

[20] 张若愚. Python科学计算[M]. 清华大学出版社, 2012.
[21] Garreta R, Moncecchi G. Learning scikit-learn: Machine Learning in Python[M]. Packt Publishing, 2013.
[22] Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning[J]. 2016.
[23] Chang C C, Lin C J. LIBSVM: A library for support vector machines[J]. 2011, 2(3):1-27.
[24] 刘天祥, 包腾飞, 宋锦焘,等. 基于遗传算法的LIBSVM模型大坝扬压力预测研究[C]// 全国大坝安全监测技术与应用学术交流会. 2016.
[25] 王娟, 慈林林, 姚康泽. 特征选择方法综述[J]. 计算机工程与科学, 2005, 27(12):68-71.
[26] 董小国, 丁冉. IDS自适应特征选择算法——进化包装(Wrapper)算法分析[J]. 微计算机信息, 2006, 22(33):46-48.
[27] Guyon I, Weston J, Barnhill S, et al. Gene selection for cancer classification using Support Vector Machines.[J]. Machine Learning, 2002, 46(1-3):389-422.
[28] 王君. 基于SVM-RFE的特征选择方法研究[D]. 大连理工大学, 2015.
[29] 王俭臣, 单甘霖, 张岐龙,等. 基于改进SVM-RFE的特征选择方法研究[J]. 网络新媒体技术, 2011, 32(2):70-74.
[30] 胡育新, 李平, 刘宝姝,等. 胰腺癌组织中基质金属蛋白酶-7(MMP7)的表达[J]. Chinese Journal of Cancer, 2000, 19(6):521-523.