学 校 代 码 10459
学号或申请号 201412172087
密 级

郑州大学

硕士学位论文

基于聚类和加权K近邻的烟叶分级研究

作 者 姓 名：李航
导 师 姓 名：申金媛 教授
学 科 门 类：工学
专 业 名 称：信息与通信工程
培 养 院 系：信息工程学院
完 成 时 间：2017 年 5 月

A dissertation submitted to

Zhengzhou University

for the degree of Master

The Research on Tobacco Classification Based on Clustering and Weighted KNN

By HangLi

Supervisor: Prof. Jinyuan Shen

Information and Communication Engineering

School of Information Engineering

May, 2017

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

学位论文作者：　　日期：　年　月　日

学位论文使用授权声明

本人在导师指导下完成的论文及相关的职务作品，知识产权归属郑州大学。根据郑州大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权郑州大学可以将本学位论文的全部或部分编入有关数据库进行检索，可以采用影印、缩印或者其他复制手段保存论文和汇编本学位论文。本人离校后发表、使用学位论文或与该学位论文直接相关的学术论文或成果时，第一署名单位仍然为郑州大学。保密论文在解密后应遵守此规定。

学位论文作者：　　日期：　年　月　日

摘要

烟叶收购阶段，正确客观的划分烟叶等级既可提高烟农的种植积极性，又可保障卷烟企业的经济利益。现阶段的人工分级存在主观性强、人力和物力耗费大等缺点，针对同一片烟叶，不同的专家有可能将它划分到不同的等级。因此，客观、快速、高准确率的智能分级是迫切需要的。

目前，烟叶智能分级的研究集中在基于烟叶图像和红外光谱进行分级两个方面。由于烟叶的光谱特征可以更好地反映含油量、色度、身份及成熟度等与烟叶等级密切相关的因素，所以本文基于光谱对烟叶分级进行了研究。烟叶智能分级系统的识别率和整体速度与所选择的分级模型和样本特征光谱的采集量存在很大的关系，为实现一个具有高识别率的实时烟叶智能分级系统本文主要进行了以下工作：

1. 烟叶光谱的采集、预处理和孤立样本的检测。利用型号为 UV-3600 的光谱仪采集 642(13 个等级)片烟叶的反射光谱；为降低基线漂移所带来的噪声和特征值间的差异对分级的影响，对光谱进行了归一化处理；由于可能存在错分类别的样本(孤立样本)，所以需要对构建分级模型的样本训练集进行选择。本文分别利用夹角余弦距离、欧氏距离和相关系数法并通过统计分析选择合适的阈值检测各个等级中的孤立样本和确定用于建立分级模型的样本训练集。

2. 分级模型的构建以及 K 近邻算法的改进。分别构建支持向量机（SVM）、极限学习机（ELM）、K 近邻（KNN）和加权 K 近邻等烟叶分级模型，将分级模型的识别率作为适应度函数，全光谱下 ELM 和 SVM 的测试集最优正确率分别为 85.75%和 91.02 %。加权 K 近邻方法为：一种方法是每个等级中训练集的权重相同，为该等级样本个数的倒数。另一种方法是先找出 K 个近邻，加上与距离呈负相关的权重，通过计算每个等级的权重之和为烟叶进行定级，两种方法相结合的识别率可达 90.77%。加权 K 近邻的分类效果优于传统 K 近邻，计算复杂度低于 SVM 和 ELM，本文选用加权 K 近邻作为烟叶等级判断的分类器。

3. 基于聚类思想的特征初步筛选。同时考虑相同特征的类内离散度和类间离散度，构造判别特征好与坏的鉴别函数 D，依据 D 值删除拐点右侧特征，在第 6 个拐点下取得最优分级效果，余下 326 个特征，测试集正确率由 90.77 %增加至 94.59%，既提高了识别率又降低了特征的个数。

4. 深层特征的筛选。采用粒子群（BPSO）、遗传算法（GA）、相关系数分析（CC）进一步进行特征的筛选。BPSO 取得较好的效果，特征数目由原来的 451 个减少到 143 个，这样采集光谱所耗费的时间可节省 68.3%；识别率由原来的 90.77%提高到 93.69%，提高了 2.92 个百分点。

关键词：光谱分析技术 烟叶分级 聚类 加权 K 近邻 相关系数分析

Abstract

At the acquisition stage of tobacco leaf, it can not only improve the enthusiasm of farmers but also guarantee the economic benefits of the cigarette enterprises by the correct and objective tobacco classification. At present stage, there are several problems in the artificial classification like the strong subjectivity and the waste of manpower and material resources. Meanwhile, different experts may put the same piece of leaf tobacco into different grades. Therefore, the intelligent classification which is objective, fast and high accurate is desperately needed.

At present, the reasearch on tobacco intelligent classification mostly focused on the aspect of image features and spectra features. The spectra features can reflect the characteristics which are closely associated with the grade of tobacco leaf like the oil content, chroma, identity and maturity better, so we study the tobacco grading based on the spectrum in this paper. Both the classification model and the number of samples of characteristic spectrum are related to the recognition rate and the overall speed in the tobacco intelligent classification system. In order to achieve a high recognition rate and real-time intelligent tobacco grading system, we carried out the following work:

1. Collect and pretreate the spectrum of tobacco leaf, and then text the isolated samples. We collected 642 (including 13 levels) reflectance spectrums of tobacco leaf by using the model of UV-3600 spectrometer. To reduce the noise caused by baseline drift and the influence on the grading among different characteristics, we normalized the spectrum. There may be some samples which have wrong labeled (isolated samples), so we should choose the sample traning set which was builded by the classificated models. Through the statistical analysis, we choose the appropriate threshold to test the isolated samples of each grade and determine the sample training set which was setted up in the classification model by the method of angle cosine distance, euclidean distance and correlation coefficient respectively in this paper.

2. Build the classification model and improve the K neighbor algorithm. We build the classification model of SVM, ELM, KNN and WKNN respectively, and

make the recognition rate of the classification model as the fitness function. The ELM and SVM optimal accuracy are 85.75% and 91.02% under the full spectrum. There are two different ways to rank the tobacco grades in the WKNN method: one is that all the training sets have the same weights on each grade, and the weight is the reciprocal of the number of samples; another one is to find the K neighbors first, and then plus a weight that is negative correlation with the distance. We rank the tobacco classification by calculating the sum of the weight of each level for the tobacco left. Combining these two methods, the recognition rate can reach 90.77%. The effect of the classification of the weighted K neighbor is superior to the traditional K neighbor, and the computation complexity is is lower than the SVM and ELM. So we select the weighted K neighbor as the classifier to judge the tobacco grade.

3. Preliminary screening of the characteristics based on the clustering. Considering the discrete degree within-class and between classes simultaneously. Constructing an identification function D to identify the feature is good or not, and delete characteristics which is on the right side of inflection point according to the D value. We get the optimal effect on the classification on the sixth inflexions, and also there remaines 326 features in the traning set. Using the method aboved, the accuracy in the traning test increased from 90.77% to 94.59%, meanwhile, the recognition rate increased and the number of features reduced.

4. Deep screening of the characteristics. We screen the features further by the particle swarm (PSO), genetic algorithm (GA), and analysis of correlation coefficient (CC) methods. The results show that BPSO have a better effect. By this time, the number of features is reduced from 451 to 143. The time we spent on collecting spectrum can save 68.3%, and the recognition rate has improved from 90.77% to 93.69%, increased by 2.92%.

Key Words： Spectrum analysis technology, Tobacco Grading, Clustering, Weighted K neighbor, Correlation coefficient analysis

目录

1 绪论 1
1.1 研究背景 1
1.2 研究意义 2
1.3 研究现状 4
1.4 本文研究内容 8
1.5 本文的组织结构 9
2 烟叶光谱数据的采集和预处理 10
2.1 烤烟叶样本 10
2.2 光谱仪和数据的采集 11
2.3 数据的预处理 15
2.4 本章小结 17
3 孤立样本的检测及训练集的选择 18
3.1 孤立样本的检测 18
3.1.1 夹角余弦距离 18
3.1.2 欧氏距离 20
3.1.3 相关系数 23
3.2 训练集的选择 25
3.3 本章小结 27
4 分级模型的构建及实现 29
4.1 极限学习机 29
4.2 支持向量机 31
4.3 近邻法 34
4.3.1 K 近邻 35
4.3.2 加权 K 近邻 36
4.4 本章小结 38
5 特征筛选 39

5.1 基于聚类的特征初筛选 40
5.1.1 类内离散度 41
5.1.2 类间离散度 42
5.1.3 构造鉴别函数 42
5.2 深层特征筛选 43
5.2.1 粒子群算法 44
5.2.2 遗传算法 46
5.2.3 相关系数分析法 48
5.3 本章小结 49
6 总结与展望 51
参考文献 52
致谢 56
个人简历、在学期间发表的学术论文与参与项目 57

1 绪论

1.1 研究背景

烟叶作为我国的主要经济作物，我国烟草种植面积多达 100 多万公顷，每年产量为 200 多万吨，覆盖范围多达 23 个省市，最新公布的国标（GB）将烤烟叶划分为 42 个等级[1]。烤烟叶片所处的等级与成品烟的质量密切相关，因此正确客观的给烟叶定级变得极其重要，现阶段的烟叶收购方式为：受过专业训练的人员通过视觉、嗅觉和触觉的综合因素来对烟叶进行定级，这种分级方法主观性比较强，耗费巨大的人力和财力，该方法与分级人员的经验相关[2]。针对同一片烟叶，不同的专家有可能将其划分到不同的等级，这将造成收购者与烟农之间的纠纷，结果是打消烟农的种植积极性，造成我国经济的损失。针对当前人工分级所存在的不足，科研人员基于烟叶的图像[3-5]特征、光谱[6,7]特征在智能分级方面进行了理论上的研究，光谱特征可以更好的映射与烟叶等级息息相关的厚度、含油量等因素，光谱分析技术比较成熟且在多个领域得到了广泛的应用。

光谱分析是依据物体的光谱特性对其进行鉴别及分析其化学成分的组成和相对含量，近红外光谱（780~2526nm）包含红外短波和红外长波[8]。近红外光谱所处的波段范围和物体中的含氢基团（C=O 和 X-H 等，其中 X 为：S，C，N，O）振动的各级倍频和合频的吸收区相同[9]，这些信息可以更好的反映物体的内部化学成分及其相对含量。

20 世纪 50 年代，农副产品业的快速发展对商品化仪器的技术要求越来越高，经过 Norris 等人在近红外光谱分析技术方面的努力，使得 NIR 技术在各个领域得以发展和应用[10]。60 年代中后期，由于当时的技术和设备不太完善等不足之处，当时背景下的近红外光谱分析技术具有受外界因素影响大和灵敏度低下的缺点，加之新技术的出现，使得近红外光谱技术步入低谷阶段，在新兴的领域上基本没有得以运用。1980 年以后，计算机技术使分析仪器的参数得以数字化，加之化学计量学科的发展，通过仪器提取光谱信息时的抗干扰能力得到明显的改善，使得近红外光谱技术在很多领域得以快速发展[11,12]。20 世纪 90 年代，光纤技术的迅速发展，使得近红外光谱技术在在线分析方面得到了广泛的应用，

促进工业的快速发展，带来了较好的社会效益。经过几十年的发展，具有代表性的光谱仪器生产商主要有岛津、瓦里安、Nicolet、PE、Bruker、Bio-Rad、Avantes等，不同的仪器精度和光谱区间范围，为光谱分析技术在各领域的发展提供了技术上的支持。

近红外光谱技术因具有快速、灵敏度高、高效、无破坏的特点被广泛应用于多个行业，如饲草业中草料中蛋白质、灰分、水分和氮含量的测定；药物制作过程中甘氨酸的定量测定，在线监测药物的结晶过程；乳制品等级的鉴别及脂肪和蛋白质的检测；肉类产品的等级鉴别及肉类来源的部位；烟草中的尼古丁、甲醛、铅砷、烟焦油等有害人体健康的化学成分的含量。光谱分析技术在多个领域的发展与应用，为烟叶智能分级提供了技术上的支持和保障。

随着模式识别在多个领域的广泛应用，计算速度和测试精度要求越来越高，模型构建时需要考虑特征的筛选及孤立样本检测。特征筛选不仅有利于降低模型的计算复杂度，还可以降低数据采集量。孤立样本检测不仅可以加快系统的收敛速度，而且还有利于测试精度的提高，避免出现“害群之马”的现象。由于本文烟叶数据来源于烟草公司，提供样本中可能存在错分类别的样本，这些样本对分级模型的测试精度有很大影响，在模型学习阶段时挑选出那些孤立样本可加快系统的收敛速度；有效烟叶光谱特征的筛选，既可降低光谱数据采集量又可加快分级速度；模型训练阶段训练集的选择极其重要，数目过多不利于系统的实时性，数目过少则不能保证样本的遍历性；基于此，本文对孤立点的检测、训练集的选择、分级模型的构建、有效烟叶光谱特征的筛选进行了研究。

1.2 研究意义

我国烟叶的种植面积、年产量、销售量在世界上遥遥领先，地球上每四个人当中近一人吸烟，成品烟在人们生活中普遍存在，逐渐成为了人们生活中的日常消费品，成品烟质量的好与坏和卷烟时所用的烟叶片质量的优与劣息息相关。由于烟叶生长所处的区域和气候不尽相同，生长周期也不相同。即使是同一区域的烟叶，每株烟叶上的叶片所处的位置及受光照程度差别也较大，这样复杂的生长环境加剧了分等级时的困难，最新公布的国标（GB）将烟叶划分了42 个等级。烟叶所处的等级直接影响成品烟的质量，做不好收购阶段时的分等级工作将造成国家经济的流失，快速客观的对烟叶进行等级的划分很必要。

现阶段的烟叶分级是通过受过专业训练的人员对烟叶进行眼观、鼻闻和手摸来进行等级的判定，这种方法效率低下且主观性强，为寻找高效、便捷、实用的方法，目前烟叶智能分级主要集中在获取烟叶的图像信息或者光谱信息，结合模式识别挖掘有用特征信息，筛选有用特征结合模式识别实现烟叶的自动分级。基于图像特征进行分等级时，通过计算机视觉技术提取人工划分烟叶等级时所参考的图像特征[13,14]（颜色、纹理、几何）来给烟叶进行定级，该方法不能映射与烟叶等级紧密相关的油分、气味和厚度等内在因素[15]，加上所提取的特征在数量级上相差甚大，如提取颜色特征的方差和面积特征，它们数量级的差别需要对各个特征分别进行归一化处理，这将增加机器的运算量，从而影响分级的速度，光谱分析技术因拥有独特的优势广泛应用于多个领域。

近红外光谱分析是依据光的路线经过物体的阻挡后，会产生反射光谱和部分透射光谱，由于不同物体内部所含的化学成分含量不尽相同，可以根据光谱信息值的大小对物体进行定性或者定量分析。在样品分析时，首先需要选择相应的标准库来构建训练模型，对于检测样品需通过所构建的模型进行定性分析或者模式判别[16]。之所以将近红外光谱分析技术运用于烟叶自动分级系统，因为该技术具有以下特点：

一、无破坏性、在线检测。由于近红外线的穿透能力比较强，可以直接穿透物体的外包装进行检测。这样既可以避免来回取放物体的繁琐步骤，又可以减少物品的破损概率，很大程度上减少了人工操作的时间。只需在生产线上添置相关装置，就可以实现对物品等级的快速检测并对实际工作的调度起到指引性作用。

二、信息量大、分辨率高。由于光谱波段与物体分子内部含氢基团的倍频和合频的振动区间相同，所采集的烟叶光谱特征可更好的映射与其所处等级息息相关的油分、厚度、身份等信息。经过不断的改进与完善，仪器的采样间隔可达 2nm。

三、无污染、资费低。光谱仪器的正常运行只需耗费少量的电能，不会对环境造成污染。另外，在环境复杂且对人体安全构成威胁的场所，如强辐射、高压、高温等环境下，光谱仪器可以替代人工进行操作。商用化仪器操作起来比较简单，只需简单的培训即可代替受过专业训练的人员，对物体进行定性的分析。

正所谓“金无足赤，人无完人”，近红外光谱分析的不足之处在于它测试样

本时的灵敏度低，存在基线漂移的干扰，外界因素容易引起测量精度的波动。再者就是构建模型时需要正确的标准库，标准库的选择需要满足数量尽量少、遍历性比较强且不能有错误样本的特点。光谱仪器存在一定的局限性，只适合于对那些含有含氢基团成分的物品进行测定。

NIR 技术运用于烟叶智能分级系统时需解决以下问题：其一，构建模型时训练集的选择。假设训练集的个数为无限多，基本上囊括了所有的可能性，那么系统很容易实现相当高的识别率，但是系统计算量和耗费时间也急剧增加，根本不实用，再者也难免出现不可预测的特例；如果训练集的个数特别少，就算是很好的识别系统，对于检测样本也只能是随机猜测，给出一个不一定正确的结果，这样的效果也不理想；所以训练集个数的选择及其哪一个样本入选训练集对分级模型起着至关重要的作用。其二，采集光谱特征的个数。特征个数很多的情况下，采集时间比较长，而且特征之间存在较强的相关性，不一定有利于分级；特征个数比较少时，不能更好的反应物品的特性，不易于利用这些特征实现物品的检测[17]。因此需要在采集的光谱特征中进行特征的筛选，实际应用时只需采集所保留的较优的特征，这样既有利于系统的实时性又提高了系统的测试精度。其三，分级模型的构建。以分级速度和测试集的正确识别率为基准，选取较优的模型有利于技术的推广。

基于烟叶的近红外光谱实现烟叶智能分级，可有效的弥补人工分级存在的不足之处。在实际应用当中，涉及到采集何种光谱数据，采用什么样的方法对采集的数据进行预处理以消除机器及外界因素带来的噪声，如何筛选簇中的孤立样本，哪些样本入选训练集，特征筛选的方法及构建何种模式识别模型。本文所研究的孤立样本检测、特征筛选方法及模式识别模型等就变得不可或缺。

近红外光谱分析技术为烟叶智能分级开辟了一条新的路径，该技术可以避免因专业人员经验不同导致的分级结果不同而引起的争执。客观公正的对烟叶进行定级，不仅有利于提高烟农的种烟积极性，而且保护了国家的经济利益。同时可以缩短收购烟叶阶段的周期，并且极大地减少了人员的费用，对工作人员在实际收购烟叶阶段时起到指导的作用。

1.3 研究现状

近红外光谱分析法因具有高效、便捷、无污染等优势被广泛应用于农业、

石化、食品业、制药业等多个领域，该技术在国内烟草行业的研究也颇多，本文结合光谱分析技术在烟草方面的应用与研究进展进行以下归纳分析。

一、光谱预处理方法

获取烟叶光谱数据时，考虑到仪器受外界因素（温度、湿度、光源变化等）的影响，需要对采集的数据进行相关的预处理来消除外界干扰和基线漂移所带来的噪声，预处理方法主要有均值中心化变换、归一化、平滑法去噪声、导数法、标准正态变量变换和去趋势法、多元散射校正法、傅里叶变换法、小波变换、正交信号校正法。

均值中心化是对采集的光谱数据进行减均值的处理；归一化处理可有效减小由于光程的不同而造成的光谱变化；常用的平滑法去噪声有平均平滑法和卷积平滑法；导数法有直接差分法和 Savitzky-Golay 卷积法，是基线校正中常用的方法；标准正态变量变换（SNV）主要用于消除由于物体的颗粒大小、光程差、表面散射等因素对光谱带来的影响，去趋势法可有效消除经 SNV 处理后光谱的基线漂移；多元散射校正法（MSC）常用于消除由于颗粒大小分布不均匀而造成的散射影响，在固体和浆状物的透反射光谱分析中有着广泛的应用；傅里叶变换法将光谱看成许多不同频率的正弦波之和，可实现数据的压缩和平滑法去噪声，便于提取有用信息；小波变换预处理法将光谱看成许多小波函数的叠加之和，通过设定合适的阈值来删减无用小波系数，在保留的有用信息可以正确构建原始信号情况下实现数据的压缩和噪声的滤除。

光谱预处理在实际中的运用，刘润杰[18]等运用减均值法对 9 个等级的烟叶光谱进行预处理，有效的消除基线漂移带来的影响，构建概率神经网络模型进行烟叶等级的判别，测试集的吻合率可达 91%。赵铭钦[19]等对采集云烟 87 类型的烟叶光谱进行 Savitzky-Golay 卷积法、MSC、均值归一化和 SNV 等预处理，结合 PCA 和 PLS-DA，判别烟叶香型时的训练集和测试集的识别率均达 100%。彭丹青[20]等对比不同小波变换压缩次数下的有用信息的保留程度，数据压缩 2 次时模型达到最优，次数过少时优化效果不明显，次数过多时删减了有用信息。贺立源[21]等获取云南、河南、湖北等烟叶产地的光谱，对数据进行平滑法和 MSC 预处理，联合 PCA 和 LS-SVM 对烟叶产地进行判别，测试集的相关系数为 0.9907。

二、光谱特征的筛选

考虑到所采集的光谱数据维数高且光谱之间可能存在很强的相关性，为挖掘有用光谱信息，降低数据采集量和分级模型的计算复杂度，提高整个系统的

速度变得非常重要。对采集的特征进行筛选，删除冗余的特征，保留有效特征变量，可加快识别速度并提高模式识别器的测试精度。现阶段常用的波长变量筛选的方法主要有相关系数分析法（CC）、连续投影法（SPA）、无信息变量消除法（UVE）、蒙特卡罗法（M-C）、间隔PLS法、遗传算法（GA）、粒子群算法（PSO）、模拟退火算法、蚁群算法等。

相关系数分析法通过分析变量之间的相关性进行特征筛选，周金治[22]等将相关系数法结合SVM用于脑电信号的特征选择，降低EEG信号维数的同时提高了分类器的识别率；温亚东[23]等筛选烟叶波长变量时将光谱间的相关系数之和最小作为判据，结果表明：所筛选出来的特征变量可更好的运用于定性分析。连续投影法基于循环原则筛选有效变量[24]，假定从一个波长开始循环，每次循环将与其投影向量最大的波长归入备选波长集合；孙俊[25]等结合连续投影法和SVR模型检测玉米叶片中的含水率，变量由72个降低到10个，验证集决定系数达到0.804。

无信息变量消除法[26]（UVE）基于偏最小二乘法的回归系数 b 进行波长变量的筛选，将 b 作为变量重要性的判据。田旷达[27]等将 UVE 法用于预测烟叶中总氮及总糖含量的模型，不仅降低了数据的维数而且提高了模型的稳健性；蒙特卡罗法通过从训练集中随机抽取一部分样本作为标准库进行 PLS 建模，多次抽取后选取显著 b 值所对应的波长[28,29]。

间隔 PLS[30]是筛选波长区间的一种方法，将光谱划分成若干个相同宽度的波段，各个波段均参与 PLS 回归，筛选 RMSECV 最小值所对应的波长段，接下来进行波段的扩张，直至得到最佳波段。陈晓辉[31]等结合 iPLS 和 CARS 算法筛选基于光谱分析液体时的有效波段及波长，降低变量维数的同时提高了模型的预测精度。遗传算法[32]模拟自然界生物进化流程，通过选择、交叉和变异等，淘汰不好的变量组合，留下最优的变量组合。刘建利[33]等结合 GA 和 PLS 预测烟草中的淀粉含量，预测精度达 0.9853。

粒子群算法[34]模拟自然界中鸟群的觅食过程，粒子间通过交互信息从而达到给定空间中最优解。申金媛[35,36]等结合 BPSO 和 SVM 对烟叶的图像和光谱特征进行筛选，降低特征个数的同时提高了识别模型的吻合率，有利于分级模型的推广。模拟退火算法[37]相似于金属退火，邹小波[38]等结合模拟退火算法和 PLS，有效的筛选出可以表征草莓中可溶性物质的含量，为鉴别草莓的成熟度提供了技术上的保障；蚁群算法[39]来源于蚂蚁觅食行为的研究，蚂蚁之间通过交换留

下的信息素可确定一条最优寻食路径，可用于有效波长的筛选。张树清[40]等采用多态蚁群算法有效的筛选出 Hyperion 影像中的有效波段，且分类正确率明显高于传统蚁群算法。

三、光谱分析技术在分析烟草中化学成分的应用

烟草中含有多种化学成分，如有机酸、钙、葡萄糖、烟碱、磷果胶、色素、氨基酸、钾、淀粉、总氮、还原糖、镁、糊精、氯、果糖、蛋白质、纤维、硫等。烟草中化学成分的含量不仅可以表征烟叶所处的等级，还可以体现叶片的成熟度。

1、常规化学成分的分析：张婷[41]采用主成分分析（PCA）法对比湖北省和国际性上优良烟草中的烟碱、总氮、氯、钾、总糖的含量，结果表明：烟草中的总氮和烟碱的含量比较符合国际标准，其余成分较标准有一定的距离。许家来[42]等对烟草中烟碱、还原糖、氯、钾及总氮的含量与烟叶的易烘烤性及耐烘烤性进行了相关分析，结果表明：烟叶的烘烤特性与其所含有的总氮量和烟碱量有较强的相关性，与氯和钾含量的相关性不大。申钦鹏[43]等提取 63 个化学成分指标，结合逐步回归法筛选所提取的特征，构建线性判别分析、支持向量机（SVM）等七种模型对烟叶香型进行分类，筛选后余下的 19 个指标对烟叶香型判断的正确率达 90%。张四平[44]针对烟叶烘烤过程中不同变黄时间对化学成分的影响进行分析与研究，得出延长 12h 后变黄的烟叶含有的化学成分与标准最匹配。赖炜扬[45]等基于正交优化方法验证了乙醇法可更好的提取再造烟中的致香物质。

2、无机化学元素的分析：刘晶[46]等对再造烟叶中的氯离子、挥发碱、钙离子等含量进行了预测分析，结果表明：再造烟叶的质量与化学成分含量有较强的相关性。田旷达[47]等对数据进行 SNV 预处理，构建 LS-SVM 模型测定烟叶中钙、镁元素，它们的验证集决定系数分别为 0.9755 和 0.9422。周淑平[48]等基于烟叶的漫反射光谱对无机元素的含量进一步分析，构建测试铁、钙、镁、锌、镁含量的数学模型，结果表明：钙和镁的稳健性较好，铁锰锌含量的测试精度不高。付秋娟[49]等构建数学模型预测烟草根中氮、钾、钙和镁的含量，其验证集的决定系数均超过了 0.93。刘岱松[50]等对比光谱分析法和化学分析法预测烤烟中钾的含量，基于光谱所建立的 PLS 模型和化学分析法的相关性比较强，吻合率达 90%以上，为快速测定烟叶中钾的含量提供了参考。章平泉[51]等基于烤烟的漫反射光谱构建数学模型预测磷酸根的含量，结果表明：该方法与化学计

量学所得结果的误差小于 10%，模型的鲁棒性比较好。

结合目前光谱分析技术在烟草领域的研究现状，需要对采集的光谱进行预处理和特征的筛选，及模式识别器的选择。

1.4 本文研究内容

本文的工作主要包括以下几个方面：

一、烟叶光谱数据的采集及预处理。为快速、高效、无损的识别烟叶所处的等级，考虑到 UV3600 型号的光谱仪可以实现不同间隔的采集物体的光谱。本文采集烟叶的反射光谱用来表征等级信息；为有效的减少外界噪声及基线漂移所带来的影响，采用归一化方法将数据变换到 0~1 范围，使数据更好的适应于分级模型的输入模式。

二、孤立样本的检测及训练集的选择。考虑到烟草局所提供的样本中存在错分烟叶等级的可能，模型训练阶段的规则为：训练样本尽量少、具有很好的遍历性、不能存在错误的样本，本文通过欧氏距离、相关系数、夹角余弦距离法进行孤立样本的检测和训练集的选择，将那些偏离中心的样本筛选出来，使它们不参与分级模型的训练。

三、分级模型的构建。构建支持向量机、极限学习机、加权 K 近邻分级模型，SVM 对小样本、非线性、维数高的数据具有良好的分级能力；ELM 训练过程不需要调整权值和偏置，只需给定隐含层节点的个数便得出唯一的最优解；由于各个等级烟叶的在数目上有一定的差别，为减少数据倾斜带来的影响，在 K 近邻算法中对所选中的训练集加上不同的权重。

四、基于聚类的特征初筛选。对于采集的光谱数据，光谱之间可能存在很强的相关性，考虑在保证分级吻合率不降低的情况下先去除部分特征以实现特征初筛选。本文将聚类思想运用于光谱特征的初筛选，兼顾特征在同类和不同类中的离散度，构造鉴别特征好与不好的函数，通过去除鉴别函数中所出现拐点的右侧特征，同时考虑特征个数和识别率因素来实现特征的初筛选，为下一步进行特征的深度筛选做准备。

五、基于粒子群、遗传算法和相关系数分析的特征深度筛选。对经初筛选余下的特征采用随机优化的方法进行深度筛选，结合分级模型的适应度函数选择最优的特征组合，降低特征维数可降低模型的计算复杂度和光谱采集的时间。

1.5 本文的组织结构

本文在基于烟叶光谱特征实现自动分级时，进行了孤立样本的检测、训练集的选择、分级模型的构建及特征的筛选。文章主要包括 6 个章节，各章节组织安排为：

第一章：绪论。简要分析当前烟叶分级的现状，介绍了近红外光谱技术及其在烟草领域方面的应用。针对烟叶光谱维数高，提出了筛选有效的光谱特征，包括初筛选和深度筛选；样本集中孤立样本的检测及训练集的选择；分级模型的构建。

第二章：烟叶光谱数据的采集和预处理。简要介绍了 42 个等级烟叶的组成，采用型号为 UV3600 的光谱仪获取烟叶反射光谱，简要概述光谱仪器的工作参数，通过对所采集烟叶的光谱进行归一化预处理来实现噪声的去除，预处理后的数据较符合于分级模型的输入模式。

第三章：孤立样本的检测及训练集的选择。针对样本集中孤立样本的检测，采用夹角余弦距离、欧氏距离和相关系数三种方法进行相关分析，筛选出孤立样本使其不参与模式识别器的训练阶段，同时设定合适的阈值进行训练集的选择。

第四章：分级模型的构建及实现。构建的分类器有极限学习机、支持向量机、K 近邻及加权 K 近邻，为避免由于各等级叶片数目的差异所造成的数据倾斜问题，对所选中的训练集加上与距离呈反比的权重。

第五章：特征筛选，特征筛选包括特征初筛选和特征深度筛选。基于聚类思想，同时考虑同一特征的类内离散度和类间离散度，构造鉴别特征好与坏的函数，按由小到大的顺序对各个特征的鉴别值进行排序，在排序后出现的拐点中筛选较好拐点，在保证识别率的前提下实现特征初筛选。采用粒子群算法、遗传算法及相关系数分析法对经过初筛选余下的特征进行深层筛选，为分级系统选取较优的特征组合，以较少的特征个数实现较高的测试集识别率，选取兼顾识别率和识别速度的模式识别器。

第六章：总结和展望。对本文在烟叶分级方面的研究进行总结，分析存在的不足之处。

2　烟叶光谱数据的采集和预处理

考虑到将来近红外光谱技术在烟叶等级划分时的实用性和推广性，采集的光谱特征个数越少且分级正确率越高则越有利于分级模型的推广，光谱仪的参数主要有波长采集的范围、采样间隔及灵敏性。通常情况下，间隔小、范围广的烟叶光谱可更好的反映烟叶内部化学成分信息，然而数据维数高时增加了分级模型的计算量，并且特征间的相关性有可能降低模型的识别率。为方便实际收购阶段的操作，本文在无损烟叶的情况下获取光谱，采集烟叶的反射光谱。

考虑到光谱仪采集烟叶光谱时受外界因素的影响，所产生的噪声及仪器本身的基线漂移都会对分级模型产生一定的干扰。为使数据更好的适应于分级模型的输入模式，消除基线漂移、外界温度、谱重叠等因素对光谱的影响，对于所采集的光谱需要进行相关的预处理。

2.1　烤烟叶样本

国家最新公布的标准将烤烟叶划分为 42 个等级，目前人工进行烟叶等级划分时先确定烟叶所处的部位及颜色，再进一步确定最终的等级。我国的烤烟叶等级依据烟叶所处部位的不同将其划分了三个等级，即上部叶片（代号 B）、中部叶片（代号 C）和下部叶片（代号 X）；依据烤烟叶基本颜色的不同将其划分了三个等级，即柠檬黄色（Lemon）、橘黄色（Orange）及红棕色（Red），还有一些非基本色如青黄色（Green Yellow）、杂色（K）等；烤烟分组是结合叶片所处的部位、颜色和其他因素将相关性比较大的、有密切关系的几个等级合并为一个等级，共分为主组和副组两个类别，主组包含有：BF、BL、BR、CF、CL、XF、XL、HF 共计 8 个主组，副组包含有：GY、BK、S、V、CXK 共计 5 个副组，其中 V 代表微带青叶片，S 代表光滑叶片，H 代表完熟叶片。

本文所研究的样本来源于河南省烟草公司平顶山市烟草公司，所提供烟叶样本均有油纸包裹并能较好的避免外界条件的干扰，也即采集光谱前样本所处的环境相同。光谱数据由 UV3600 型号的光谱仪采集所得，共计 642 条反射光谱（13 个等级）。

2.2　光谱仪和数据的采集

采用岛津公司 UV3600 型号光谱仪采集烟叶的反射光谱，光谱仪正常工作需要预热两个小时，光谱仪及软件使用的主要参数如表 2.1 所示：

表 2.1　光谱仪及软件的主要参数

采集波段	185nm~330nm
高分辨率	最高分辨率（0.1nm），最低杂散（0.00005%）
检测器	PMT、InGaAs、PbS
检测模式	光谱、动力学、光度检测
显示波长	最小值为 0.01nm
波长精确性	可见/紫外光：±0.2nm，近红外：±0.8nm
扫描波长的速度	UV/可见区：4500nm/min，近红外：4000nm/min
光源的转换	自动转换
光度计范围	吸光率：-6Abs to +6Abs
光度计系统	双光束系统
单色器	光栅型双重单色器
测定值的显示	可显示小数点后面 1~5 位
光源	50W 卤素灯
基线平面	±0.0001Abs（200~3000nm）
S/R 倒置	可用
环境温度	温度：15°C~30 °C
环境湿度	湿度：35%~80%
电源	AC100/120/220/230/240；50/60Hz
操作系统	Windows 2000/XP
处理数据	标准化、滤波、平方根、插补等
功率消耗	300VA
光谱仪尺寸	1020W*660D*270H，重 96kg
样品组件尺寸	150W*260D*140L（mm）
可选用的操作系统	Windows 2000/XP
数据预处理	标准化、峰/谷的选取、导数、平方根
报告生成器	支持自动打印

由采集波段可知，该型号的光谱仪可采集紫外光、可见光及近红外光；该仪器具有很高的分辨率，可运用于多种场合；工作环境要求较低，基本上都可以提供其正常工作的条件；不仅可以定性的分析物质，还可以定量的对物体进行分析；具有较低的功耗，通过设置可对采集的数据直接进行简单的预处理。

UV3600 型号光谱仪尺寸大且比较沉重，不便于携带。本文基于烟叶光谱进行分析，筛选有利于划分烟叶等级的光谱特征，实际中可根据具体情况选择相应型号的光谱仪。由于仪器具有三个检测器，各个检测器所测定波段的范围不尽相同，如图 2.1 所示；在不同波段的灵敏度也有一定的差异，如图 2.2 所示。

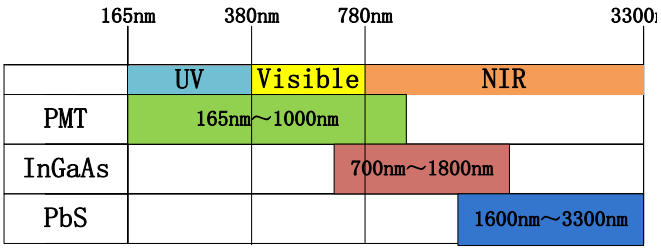

图 2.1　PMT、InGaAs 和 PbS 检测器的工作波段范围

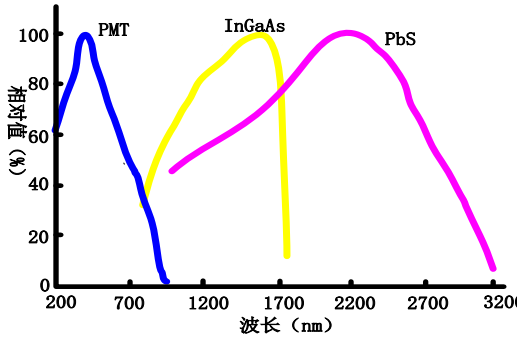

图 2.2　PMT、InGaAs 和 PbS 检测器在不同波长下的灵敏度

由 PMT、InGaAs 和 PbS 检测器所处的工作波段范围及不同波长下的灵敏度曲线可知，采集光谱时会出现检测器的切换。采集烟叶光谱的步骤如图 2.3 所示：

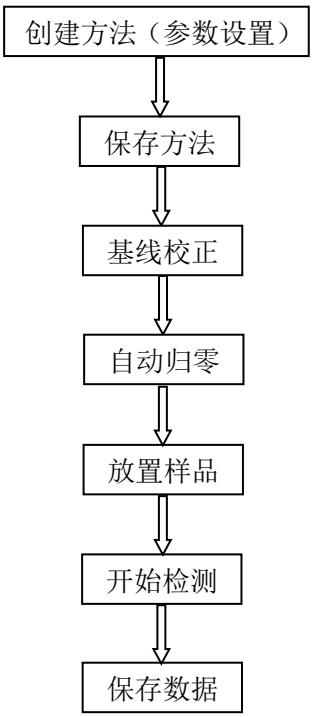

图 2.3　采集烟叶光谱流程图

按照图 2.3 流程采集烟叶的反射光谱。首先，开启开关后，光谱仪预热两个小时；放置机器自带的挡片，双击软件图标启动软件，点击 Connect 按钮实现仪器和软件的连接。其次，设置仪器的参数，采集光谱模式调到 R（Reflect），波段范围为 1500nm~2400nm，采样间隔设置为 2nm，采集波段范围内检测器的模式会进行自动切换，检测器之间的切换会使光谱的幅度出现波动，加之随仪器长时间的工作，仪器温度升高将会出现光谱的基线漂移，在烟叶分级系统中需要对数据进行相关处理以消除各种噪声。速度的模式设置为中速采集。最后，点击 Baseline 按钮实现机器的校正，待机器将测试波段校正完之后，单击归零键（Auto-zero）按钮实现归零的设置；点击 Start 键开始测试挡片的光谱，一般情况下测试值应是常数 100，保存采集的光谱数据，待仪器所测得挡板的光谱数值为 100 且比较稳定时再进行烟叶光谱的采集；每次修改仪器的参数都要对其进行校正、归零等操作。

本文所采集 13 个烟叶等级的级别及片数为：B2F(29 片)、B3F(34 片)、B4F(56 片)、C2F(49 片)、C3F(35 片)、X2F(52 片)、X3F(66 片)、X4F(55 片)、C2L(54 片)、C3L(49 片)、X2L(47 片)、X3L(57 片)、X4L(59 片)，共计 642 片，其中 B 代表上部，C 代表中部，X 代表下部，F 代表橘黄色，L 代表柠檬黄色。

采集不同光谱波段所花费的时间如表 2.2 所示：

表 2. 2　采集不同波段范围下的时间花销

波段范围	特征个数	耗时
300~2600	1151	4’52”
1200~2600	701	2’55”
2460~2600	71	29”

光谱仪所采取 13 个烟叶等级中部分等级的反射光谱如下所示：

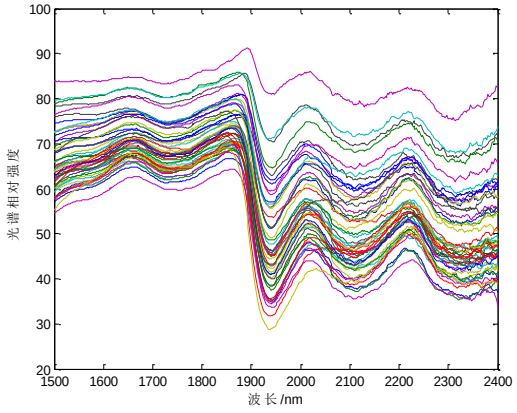

图 2. 4　C2L 等级不同波长下的反射光谱值

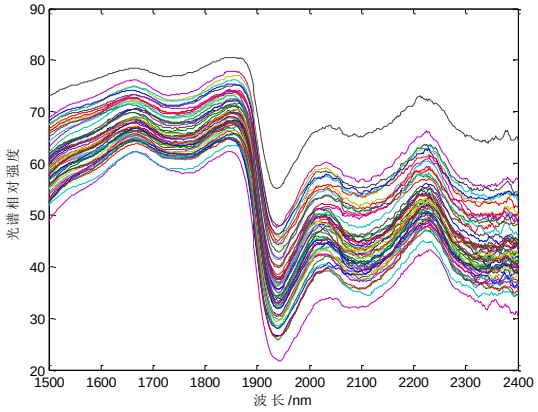

图 2. 5　C3L 等级不同波长下的反射光谱值

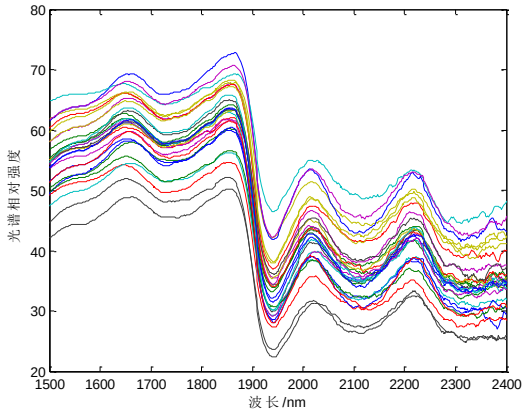

图 2.6　B2F 等级不同波长下的反射光谱值

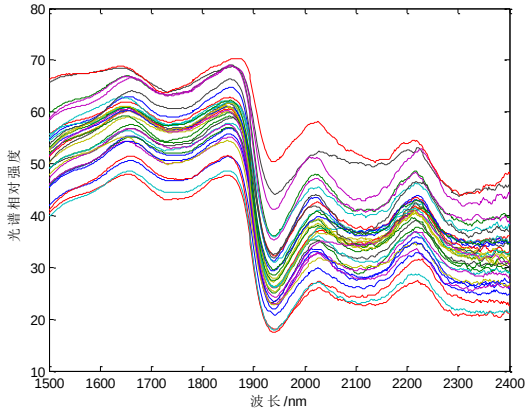

图 2.7　B3F 等级不同波长下的反射光谱值

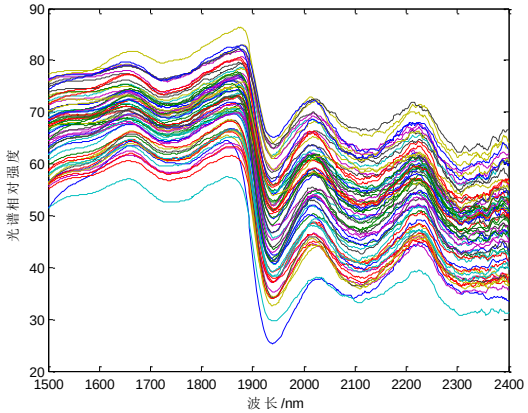

图 2.8　X3F 等级不同波长下的反射光谱值

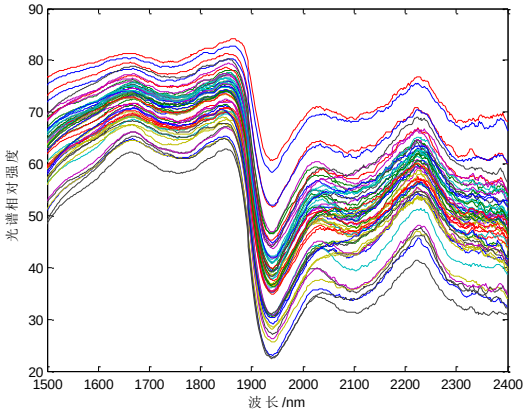

图 2.9　X3L 等级不同波长下的反射光谱值

对比烟叶的反射光谱可知：烟叶光谱在 1850nm~1950nm 波段范围内光谱值出现较大的波动，出现了较明显的吸收峰，表明在这一段光谱范围内烟叶内部的含氢基团含量变化大；从同一等级中光谱的纵向值来看，1500nm~1850nm 波段范围内光谱值波动范围小，表明该波段范围内含氢基团的含量比较一致；C2L 和 C3L、B2F 和 B3F、X3F 和 X3L，它们三种组合中的光谱取值变化趋势和光谱值基本保持一致，这正好验证了实际分组中它们处于同一主组的结论；C2L 和 C3L 级别的光谱图中存在个别样本与其同一等级差别较大的现象，结合采集光谱时，确实存在颜色差别明显的情况，构建模型时应另外考虑这些孤立样本，以保证模型的推广性。

当光线照射物体时，物体吸收一部分能量之后，会反射一部分能量，光谱相对强度即反射回来的能量占总能量的百分比。光谱在一定波段范围内的取值是连续变化的，波长之间可能存在较强的相关性，如果筛选出具有表征性的特征则可以降低光谱数据的采集量；随着机器工作时间的增加，加之各个检测器

在不同波长上的灵敏度不尽相同，会给烟叶的光谱带来一定的噪声和基线漂移，为更好的消除噪声和基线漂移所带来的影响，需要对采集的光谱进行一定的预处理。

2.3　数据的预处理

光谱数据的预处理不仅可以降低基线漂移和外界噪声对光谱的影响，还可以使数据更好的匹配于模式识别器的输入模式。对光谱进行预处理的方法主要有均值中心化、标准化、平滑法、导数法、SNV 法、MSC、WT、OSC、FT 等，烟叶等级的判定属于定性分析，定性分析即判断出物体的属性。对于烟叶光谱数据采用的预处理方法主要有减最小值、减均值和小波变换等，减最小值预处理即原始光谱值减去光谱中最小的值；减均值预处理即原始光谱值减去光谱的均值；小波变换预处理即通过把信号看成不同频率正弦函数的叠加以实现数据的压缩及信息的提取。从计算复杂度上来看，减最小值和减均值预处理因只涉及到简单的减法和除法等基本运算，所以花费时间较少；小波变换可以对光谱进行局部化分析，细划分光谱导致相关处理耗时较长。

本文将减最小值和归一化相结合，将光谱数据通过函数映射到（0,1）范围内。这样不仅有效的降低了由于基线漂移所引起的光谱波动，还可以使数据更好的匹配于模式识别模型的输入模式。具体方法如式 2.1 所示：

$$y_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{2.1}$$

公式 2.1 中，x_i 为未经处理的原始光谱，$\min(x_i)$为原始光谱数据中的最小值，$\max(x_i)$为原始光谱数据中的最大值，y_i 为经过归一化预处理之后的数据。此法既可消除仪器引起的基线漂移，又可将数据映射到(0,1)之间，同时保留了烟叶光谱值变化的趋势。除此之外，该预处理方法只涉及到减法和除法的基本运算，运算速度快。针对同一片烟叶的反射光谱，三种不同预处理方法的时间开销如表 2.3 所示：

表 2.3　不同预处理法的时间开销

方法	减最小值	减均值	WT
时间开销	$1.0*10^{-3}$s	$1.0*10^{-3}$s	36.003s

由表 2.3 可知，WT 预处理的方法所花费的时间远超于减最小值和减均值，

在实际应用中的推广性较差。

依据式 2.1 对采集的光谱(1500~2400nm，间隔 2nm)进行预处理，部分等级烟叶的反射光谱预处理前后光谱图如下所示：

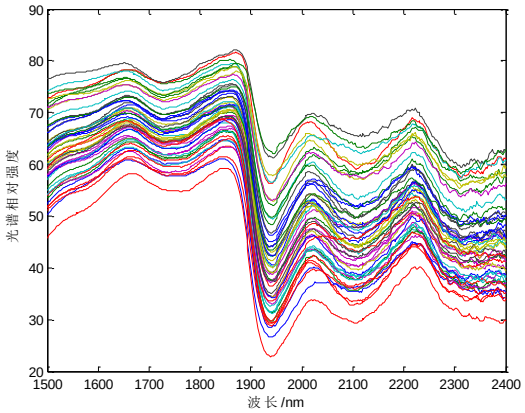

图 2. 10　X2F 等级的反射光谱

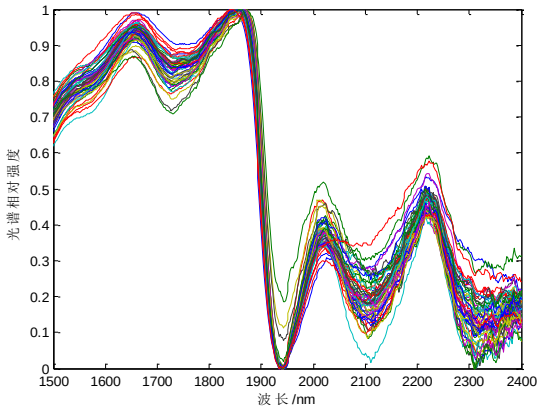

图 2. 11　X2F 等级预处理后的反射光谱

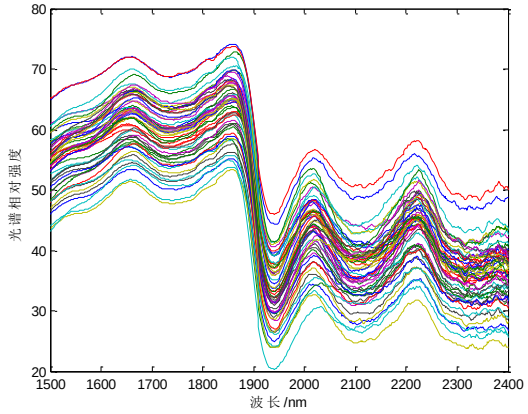

图 2. 12　B4F 等级的反射光谱

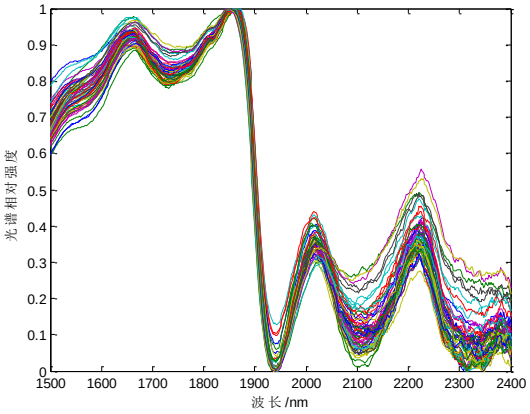

图 2. 13　B4F 等级预处理后的反射光谱

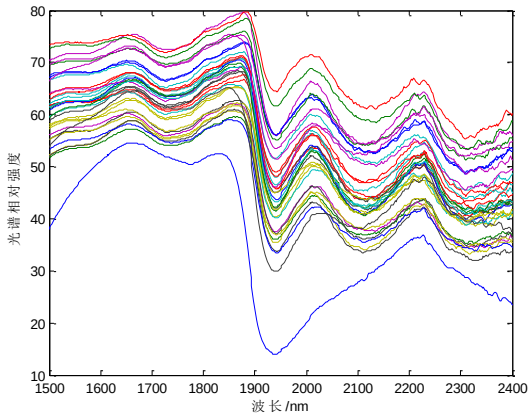

图 2. 14　C3F 等级的反射光谱

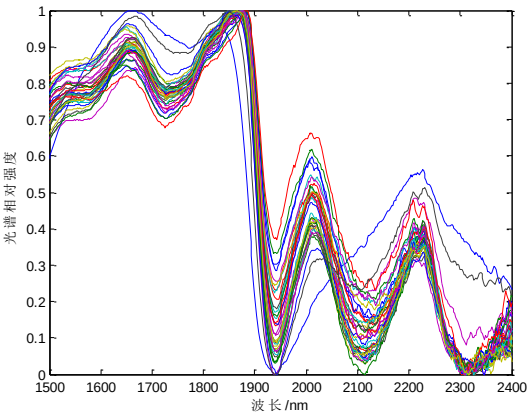

图 2. 15　C3F 等级预处理后的反射光谱

该预处理方法的不足之处为：当某一个样本的特征数值很大时，则其余的样本之间的差异性将会明显减小。孤立样本将对光谱数据的预处理造成极其不好的影响，通常会对所提供的的样本集进行孤立样本的检测。从经过预处理后的光谱图中可知，同一等级中的样本之间的差异程度明显减少，并且保证了各个特征的变化幅度处于同一级别上，从而减弱了由于数据本身的差别而带来的影响；预处理后各个特征取值范围变换到 0~1 之间，该范围的数据可以更好的适应于分级模型的输入模式。

2.4　本章小结

本章首先对当前 42 个等级的烟叶进行分组的现状进行简要的介绍，对实验所采用的样本、UV3600 型号仪器的参数进行概括；对光谱波形值变化大的波段进行了分析，得出采集的样本中可能存在错分类别的样本；然后分析了本文所采用的对采集光谱数据进行预处理的方法。

本文对采用的光谱数据预处理方法可有效的消除基线漂移和外界因素引起的噪声，将原始光谱数据的取值映射到(0,1)范围之间，这样既保留了光谱值的变化趋势，又能使数据更好的适用于分级模型的输入模式。

3　孤立样本的检测及训练集的选择

模型训练过程中，要求尽量避免错分类别的样本。由所采集烟叶的光谱图可知，某些样本与其所在同一级别中其他的样本差别很大。结合实际获取烟叶光谱时相同级别中确实存在个别样本与其他样本的颜色相差甚大的情况，可能是烟草局提供的样本集中存在错分等级的叶片。错分标签的训练样本对整个分级系统影响较大，不仅影响训练模型的收敛性，而且对所构建模型的测试集正确率也有影响。所以在训练阶段应该挑选出那些可能是错误类别的样本，使其不参与模型的训练，筛选出那些具有代表类别的样本（标准库）作为训练集[16]。

分级模型构建时所采用的训练集应满足以下几个条件：首先，训练集中尽量不要出现错分类别的样本，存在错误标签的样本将会使分级模型的预测能力大大下降。其次，训练样本要尽量的少，这样可以降低模型的冗余性，加快分级系统的收敛性。最后，要有很好的遍历性，遍历性可使训练集具有很好的完备性，可增加分级模型的预测精度。

3.1　孤立样本的检测

模型训练阶段孤立样本的检测可加快模型的收敛速度，同时有利于提高模型的测试精度，从而使模型具有良好的推广能力。训练集的正确性、遍历性和精简性对模型有着极其重要的作用。本文采用夹角余弦距离、欧氏距离和相关系数法来实现孤立样本的检测，并设定合适的阈值进行训练样本集的选择。

3.1.1　夹角余弦距离

夹角余弦距离表示样本间的相关性，通过两个向量之间的夹角余弦值表示向量之间的相似系数，其值表征空间中两个向量的在方向上的差距。通过计算样本间的夹角余弦距离可筛选出训练集中的孤立样本，样本 i 和样本 j 的夹角余弦距离计算公式为：

$$cos\alpha_{ij} = 1 - \frac{\sum_{k=1}^{m} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2}} \tag{3.1}$$

式 3.1 中 x_{ik} 和 x_{jk} 分别为同一等级中样本 i 和样本 j 的第 k 个特征变量，$cos\alpha_{ij}$

表示样本 *i* 和样本 *j* 的夹角余弦距离。由计算公式可知，当 *cosα* 取值为 0 时表示样本相同，当 *cosα* 取值为 1 时表示样本在空间上正交，完全不相同。基于夹角余弦距离可以进行孤立样本的检测。

本文计算样本间的夹角余弦距离思路为：首先选定某一样本，计算该样本与同一等级中其余样本的夹角余弦距离，求取该样本与其余样本之间的夹角余弦距离的均值来表征该样本在等级中的相似系数。依据公式 3.1 计算样本间的夹角余弦距离，部分等级中烟叶样本间的夹角余弦距离如下所示：

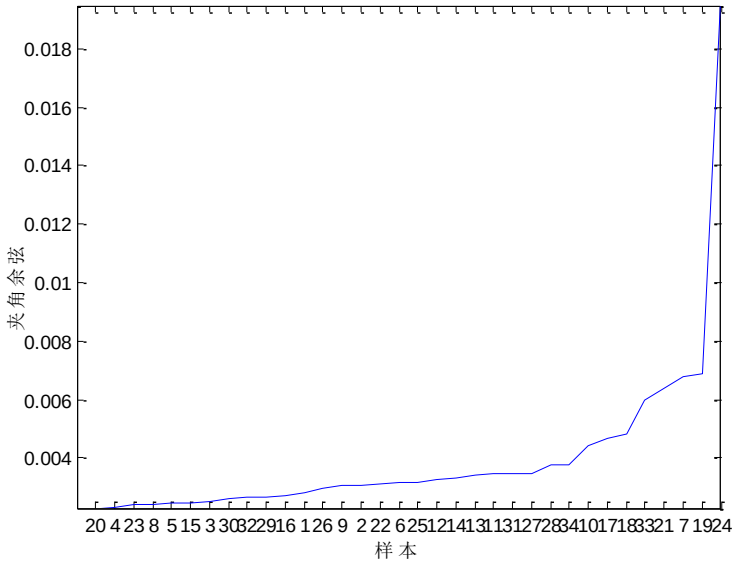

图 3. 1　B3F 等级样本间的夹角余弦距离

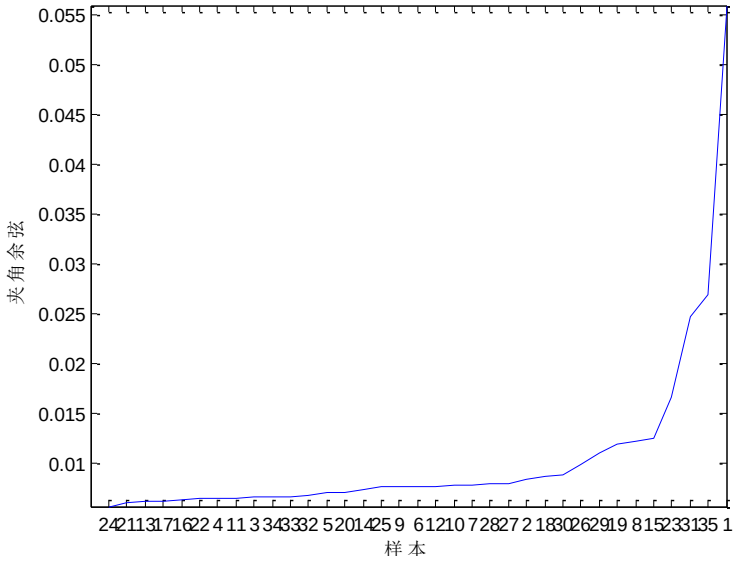

图 3. 2　C3F 等级样本间的夹角余弦距离

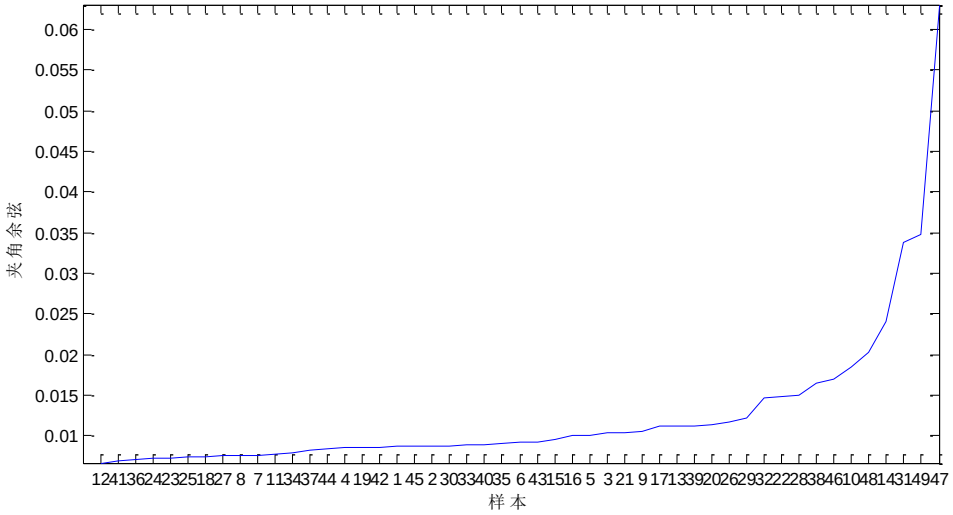

图 3.3　C3L 等级样本间的夹角余弦距离

样本间的夹角余弦距离表征它们的相似系数，夹角余弦值越小则表明该样本与其余样本越相似，越靠近等级的中心。由图 3.1 可知，标签为 20 的样本与 B2F 级别中的其它样本最相似，而标签为 24 的样本与其余样本最不相似，且相似度取值存在拐点性的变化，有可能是错分类别的样本，模型训练阶段应避免标签为 24 的样本参与其中；同理可知，图 3.2 中标签为 1 的样本和图 3.3 中标签为 47 的样本视为孤立样本;孤立样本与同一级别中其余样本的相似系数较小，所挑选出来的孤立样本应避免模型的训练。

3.1.2　欧氏距离

欧氏距离的主要思想源于欧式空间中两点之间距离的计算，是简单的统计分析方法之一。欧氏距离可表征样本间的差异性，通过计算样本间的欧氏距离，可以分析出哪些样本之间的差异性小，哪些样本与同类当中的其余样本差异性大，差异性特别大的样本可能是错分级别的样本。在模型训练阶段应避免孤立样本的参与，欧氏距离法可以筛选出那些孤立样本，在空间中计算点与点之间欧氏距离的方法如下：

二维空间中点 $a(x_1,y_1)$和点 $b(x_2,y_2)$两点之间的欧氏距离为：

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{3.2}$$

三维空间中点 $a(x_1,y_1,z_1)$和点 $b(x_2,y_2,z_2)$两点之间的欧氏距离为：

$$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2} \tag{3.3}$$

n 维空间中点 $a(x_{11},y_{12},\ldots,z_{1n})$和点 $b(x_{21},y_{22},\ldots,z_{2n})$两点之间的欧氏距离为：

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2} \tag{3.4}$$

式 3.4 中 x_{1k} 和 x_{2k} 分别表示点 a 和点 b 的第 k 个特征变量。在计算烟叶样本间欧氏距离时，x_{1k} 和 x_{2k} 可分别表示样本 1 和样本 2 的第 k 个特征。在筛选训练样本时，应该剔除那些可能错分类别的样本以保证模型的收敛性和预测能力。通过计算样本间的欧氏距离表征它们的亲疏程度，依据式 3.4 计算样本间的欧氏距离，某样本与同类中其余样本的欧氏距离超过一定的范围视为孤立样本，模型训练阶段应避免孤立样本的参与。

本文计算样本间欧氏距离的思路为：首先选定某一样本，计算该样本与同一等级中其余样本的欧氏距离，求取该样本与其余样本之间的欧氏距离的均值来表征该样本在等级中的亲疏程度。依据公式 3.4 计算样本间的欧氏距离，部分等级中烟叶样本间的欧氏距离如下所示：

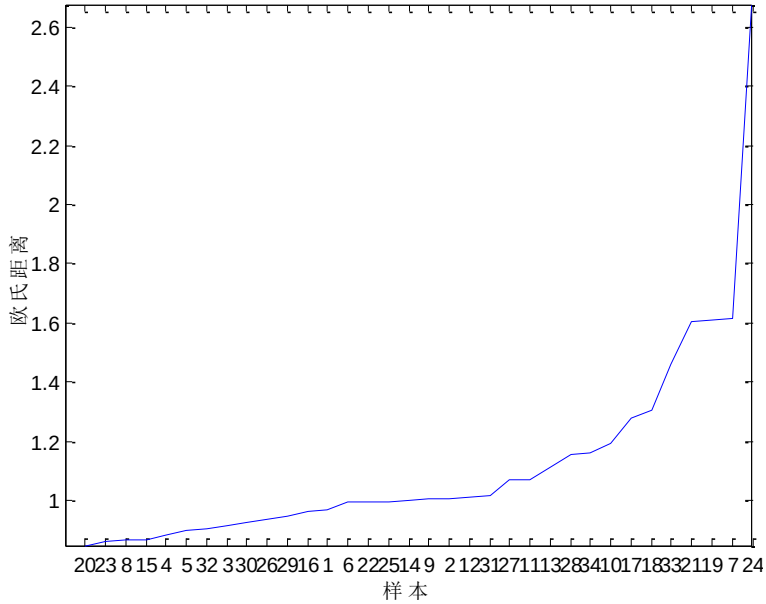

图 3. 4　B3F 等级样本间的欧氏距离

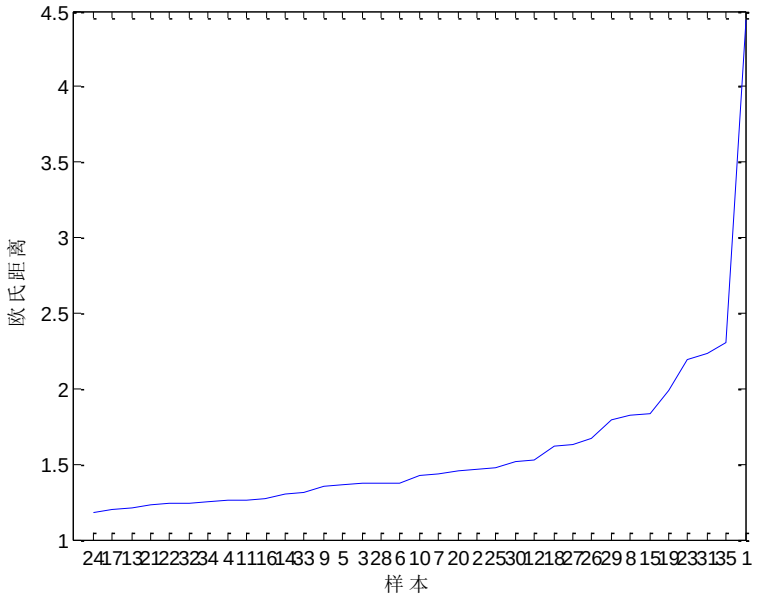

图 3. 5　C3F 等级样本间的欧氏距离

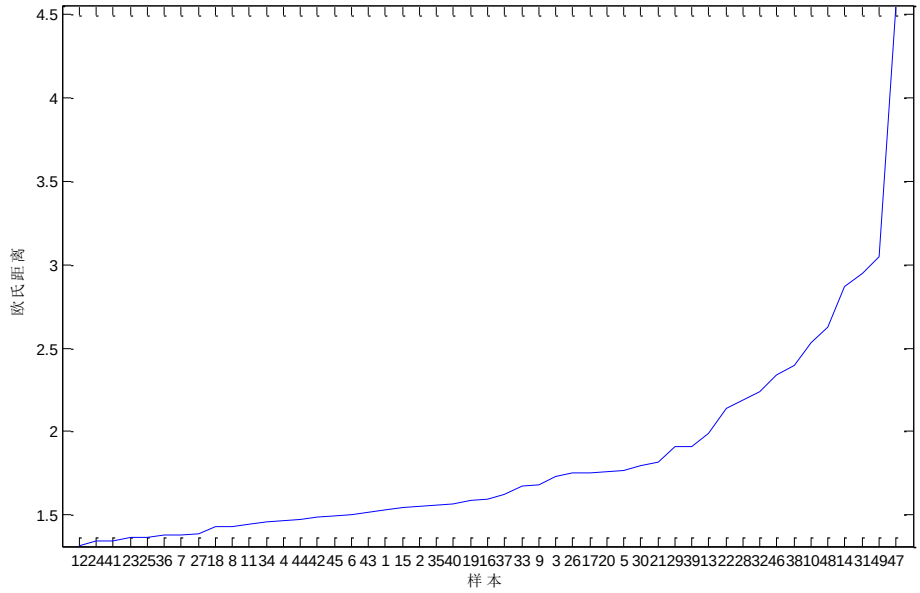

图 3. 6　C3L 等级样本间的欧氏距离

样本之间的欧氏距离表征它们的亲疏程度，欧氏距离值越小则表明该样本与其余样本的亲密性越好，距离过大偏离中心的样本可能为孤立样本。图 3.4 中标签为 24 的样本、图 3.5 中标签为 1 的样本和图 3.6 中标签为 47 的样本与同一等级中其余样本距离较大，视为孤立样本，通过欧氏距离所挑选的孤立样本和用夹角余弦距离筛选孤立样本所得出的结果相一致，应避免这些样本参与模型的训练。

3.1.3　相关系数

相关系数分析常用于化学计量学中筛选参与模型构建的有效波长，样本间的相关系数可以表示它们的紧密程度。通过计算样本间的相关系数，可分析出哪些样本比较相似，哪些是可能错分类别的样本，该方法可用于样本集中孤立样本的检测，两个样本间的相关系数计算公式为：

$$R_{xy} = 1 - \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{3.5}$$

式 3.5 中 x_i 表示样本 x 的第 i 个特征变量，y_i 表示样本 y 的第 i 个特征变量，$\bar{x}$和$\bar{y}$分别表示样本 x 和样本 y 的均值，R_{xy} 为样本 x 和样本 y 的相关系数。相关系数可以表明样本之间的相似性，取值为 0 时两个样本完全相同，计算样本间的相关系数，当某样本与同类中其余样本的相关系数超过一定的范围视为孤立样本，模型训练阶段应避免孤立样本的参与。

本文计算样本间相关系数的思路为：首先选定某一样本，计算该样本与同一等级中其余样本的相关系数，求取该样本与其余样本之间相关系数的均值来表征该样本在等级中的聚集性。依据公式 3.5 计算样本间的相关系数，部分等级中烟叶样本间的相关系数如下所示：

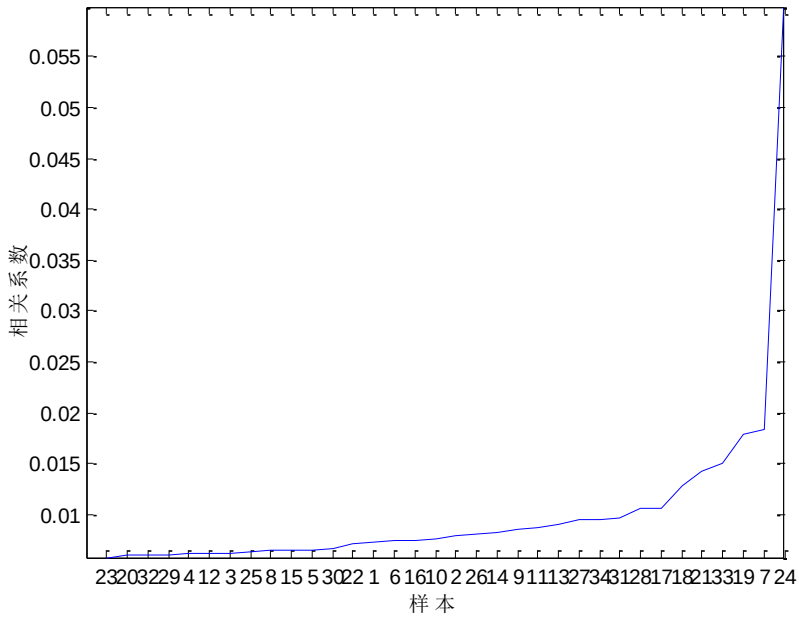

图 3.7　B3F 等级样本间的相关系数

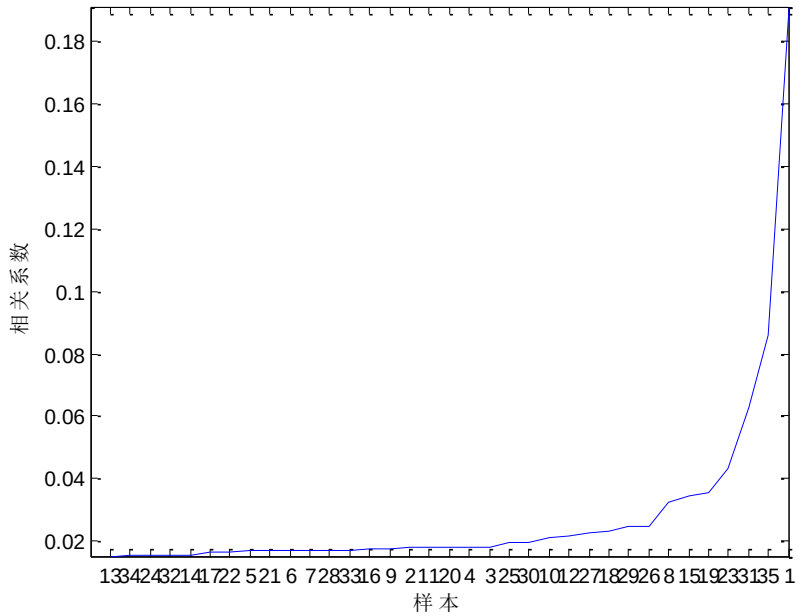

图 3. 8　C3F 等级样本间的相关系数

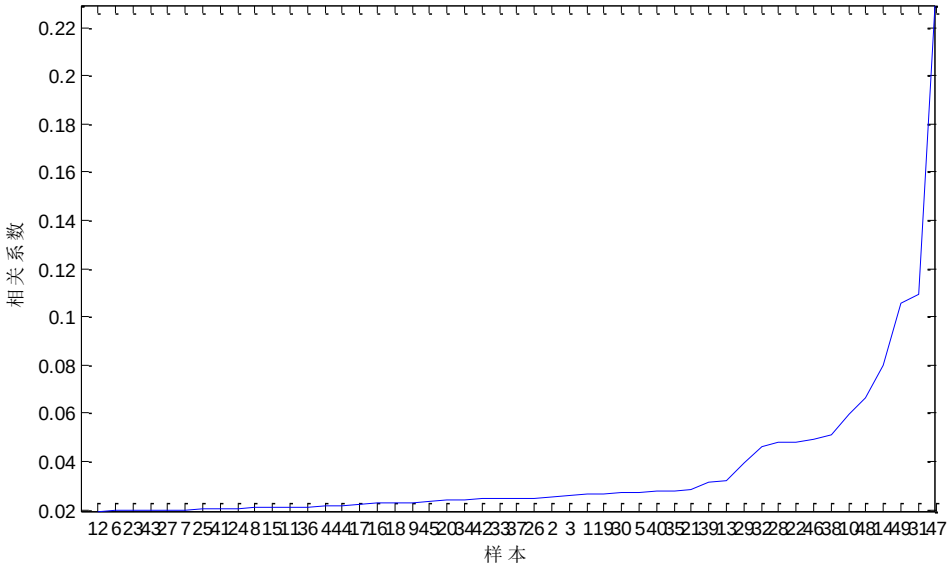

图 3. 9　C3L 等级样本间的相关系数

通过计算样本间的相关系数可知，与欧氏距离和夹角余弦距离所筛选孤立样本的结论一致， B3F 级别中标签为 24 的样本为孤立样本，C3F 级别中标签为 1 的样本为孤立样本，C3L 级别中标签为 47 的样本为孤立样本。

由计算两个样本之间夹角余弦距离、欧氏距离和相关系数的计算公式可知，当样本的特征为 m 维时，夹角余弦的计算复杂度为：3*m 次加减运算和 3*m 次乘除运算；欧氏距离的计算复杂度为：2*m 次加减运算和 m 次乘除运算；相关系数的计算复杂度为：7*m 次加减运算和 3*m 次乘除运算。由于三种方法检测

孤立样本的结果相同，考虑计算复杂度较小的方法可以节约系统内存的开销，本文采用欧氏距离方法进行孤立样本的检测。

通过对比夹角余弦距离、欧氏距离、相关系数下孤立样本检测的结果可知，三种方式下所得出的结果一致，本文采用欧氏距离下孤立样本的检测，从样本间的欧氏距离来分析，基于统计分析，将样本的平均距离大于类中心平均距离2.5倍的样本视为孤立样本，本文以此为孤立样本检测的阈值。检测结果为：B3F等级中的24，B4F等级中的2、54、55，C2F等级中的47、48、49，C2L等级中的28、47，C3F等级中的1、35，C3L等级中的31、47、49，X2F等级中的44，X4F等级中的10、13，X4L等级中的50、52，共计19片，筛选出来的样本不参与模型的训练。从图 3.7-9 中可以看出，某些样本之间的相似度比较高，为保证训练集数目尽量少、遍历性好、不存在错分类别的样本，下文对在同一等级中训练集的选择进行了研究。

3.2　训练集的选择

模式识别方法运用于烟叶智能分级时，训练集的选取和训练集的个数对识别速度和识别率有着极其重要的影响。确定训练集的个数时，其一，为保证所选择样本的完备性，选取越多的训练集则模型的预测能力越强。其二，随着训练集数量的增多，将增加系统的计算量。其三，训练集数量过少时，不能保证所选取的样本具有良好的遍历性。其四，不能有错分类别的样本入选训练集，这样的训练集将造成测试样本类别的错判。基于此，训练集过多或者过少都不利于模型的推广。

训练集的遍历性对近邻法的分类能力和实用性有着重要的影响，近邻法运用于烟叶智能分级时，由于阳光、水分及生长周期的不同，每株烟叶各个部位的叶片数量也不尽相同，这就容易出现数据的偏斜，当检测的数据出现偏斜问题时，容易出现大类吃小类的情况，即小类别的样本容易错判为大类别，这就要求所选取的训练集有较好的的遍历性。目前常用的训练集选择方法有：刘海峰等[52]根据样本的分布情况实现样本的剪裁，改善了KNN在文本分类中的速度和识别率。杨金福等[53]结合模板的约减技术去除那些距离分类边界较远的样本，减少了训练集的个数。

由于欧氏距离统计的结果与夹角余弦距离和相关系数相同，在特征个数相

同时，该方法的计算量最小，可以节约内存的开销。所以本文选用样本间的欧氏距离来表征样本之间的距离，通过计算某一样本与其余样本之间欧氏距离的均值来表示该样本的聚集性，距离越小说明越靠近中心，根据有可能不止一个样本靠近类中心。通过设置合适的阈值，将同类样本划分为几个小区段，在各个小区段内分别寻找样本代表来作为训练集，本文将小区段内的中心样本作为训练集。

具体实现方法如下：

假设第 A 类中有 N 个 p 维矢量，A 类中第 i 个样本表示为：

$$A_i = (x_{i1}, x_{i2}, \cdots, x_{ip})' \tag{3.6}$$

A 类中样本 i 和 j 的欧氏距离表示为：

$$A_{ij} = \sqrt[2]{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{3.7}$$

遍历样本集中所有样本后所得的距离方阵为：

$$A = \begin{bmatrix} 0 & \cdots & * \\ \vdots & \ddots & \vdots \\ * & \cdots & 0 \end{bmatrix}_{N*N}$$

对矩阵 A 进行行求和运算，得到同一类别当中样本间的欧氏距离，B2F 等级中样本间的欧氏距离，按由小到大排序后的结果如图 3.10 所示：

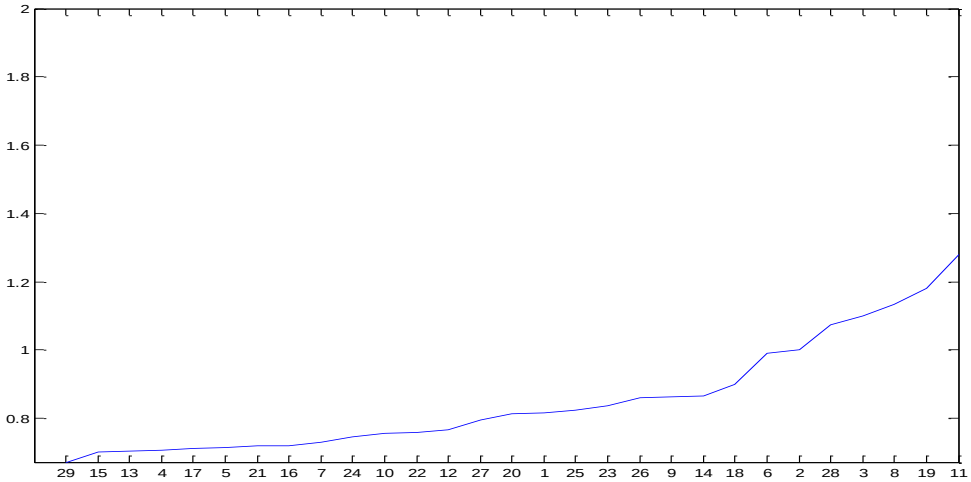

图 3. 10　B2F 等级中样本间的欧氏距离

由图 3.10 可知，标签为 29 的样本为 B2F 等级的中心。设定合适的阈值，在以样本 29 为中心的阈值范围内的样本集中，可以选取 29 作为第一训练集，其

余样本不被选中。在余下的样本中再寻找出中心样本，将小于阈值的样本用其代替，作为第二训练集，以此类推，直至样本集为空集。

训练集选取的步骤为：

(1) 设定阈值 a。

(2) 对矩阵 A 进行行求和运算，求和后各行除以矩阵 A 相应行中非零元素个数，得到矩阵 B_{N*1}。

(3) 循环开始。

(4) 如果中元素不全为 0。

(5) 寻找 B 中最小取值非零元素所在的行，记为 r。

(6) 在矩阵 A 中第 r 行寻找小于 a 的元素所在的列，记为 c。

(7) 求取矩阵 A 中 c 列的均值，均值记为$\bar{p}$。

(8) 在 c 列中寻找距离$\bar{p}$最近的样本所在的列，记为 r。

(9) 将 r 值存放在矩阵 *center* 中。

(10) 将 A 中 c 行和 c 列的元素用 0 替换。

(11) 循环结束。

(12) 输出所选取的训练集 *center*。

本文所提出的选择训练集的方法，可以避免因随机选取训练样本而不能保证遍历性的问题，考虑到样本集个数有限，且不同等级烟叶的光谱疏密程度不尽相同，相同阈值下所选择的训练集个数也有一定的差异。依据经验选取近 1/3 的样本为训练集，此时阈值取值为 0.55，训练集的选择情况为：B2F(13 片)、B3F(18 片)、B4F(21 片)、C2F(22 片)、C2L(17 片)、C3F(18 片)、C3L(19 片)、X2F(17 片)、X2L(21 片)、X3F(19 片)、X3L(16 片)、X4F(18 片)、X4L(22 片)，一共 241 片烟叶，剩余的样本作为测试集来验证分级模型的推广能力。

3.3　本章小结

采用夹角余弦距离、欧氏距离、相关系数三种方法进行孤立样本检测，三种方法的检测结果相同，欧氏距离方法的计算复杂度最小，本文基于欧氏距离方法进行孤立样本的检测和训练集的选择。

采用欧氏距离法检测孤立样本，判断样本是否为孤立样本的规则为：样本与其余样本的平均距离是否大于类中心样本与其余样本平均距离的 2.5 倍，若大

于该范围则视为孤立样本。采用此方法共计检测 19 个孤立样本，包含有：B3F 等级中的 24，B4F 等级中的 2、54、55，C2F 等级中的 47、48、49，C2L 等级中的 28、47，C3F 等级中的 1、35，C3L 等级中的 31、47、49，X2F 等级中的 44，X4F 等级中的 10、13，X4L 等级中的 50、52，构建分级模型时挑选出这些孤立样本使其不参与模型的训练。

在无孤立样本的情况下，采用样本间的欧氏距离选取阈值 0.55，在样本集中进行训练集的选择，近 1/3 的样本入选训练集，这样既保证了所选训练集具有良好的遍历性，又做到了遵循选取训练集的数量尽量少的原则。

4　分级模型的构建及实现

烟叶自动分级系统需要结合模式识别来实现等级的划分，由于所拥有的样本个数有限，需将样本划分为训练集和测试集两部分，训练集用于模型的训练，用测试集来验证模型的推广性。光谱特征的组合个数比较多，某一种光谱特征组合是否能较好的实现烟叶等级的划分需要结合分类器的适应度函数来判定。本文以测试集的正确识别率为适应度函数，正确率越高表明该种特征组合的效果越优。

本文所构建的模式识别方法有极限学习机、支持向量机、K 近邻、加权 K 近邻，通过比较各种分类器的性能，以期用一种或者几种分类器相级联来实现速度快且识别率高的分级模型。

4.1　极限学习机

极限学习机（Extreme Learing Machine）由黄广斌教授于 2006 年提出[54]，是应用范围比较广的前馈型神经网络，极限学习机的学习速度快于传统的学习算法。该方法在学习时不需要设置大量的参数，只需要给定隐含层节点的个数和核函数即可得到最优解。ELM 相比于 SVM 更容易得到最优解，可用于特征的筛选和分类。极限学习机在烟草方面的应用主要有叶片成熟度的分类和烟草病毒的预测[55,56]。

极限学习机在实际中运用时，对于复杂问题和非线性问题的求解，涉及到Moore-Penrose广义逆矩阵，对于线性方程$Ax=y$，求解最优x值时，考虑到矩阵A可能是奇异性矩阵或者不是方阵，这时可以通过求取广义逆矩阵来获得最优解，Moore-Penrose定义为：

假设矩阵$A \in R^{m*n}$，存在矩阵$G \in R^{n*m}$使得下式成立。

$$GAG=G，AGA=A，(AG)^T=AG，(GA)^T=GA \tag{4.1}$$

如果矩阵A和G满足式4.1，则称矩阵G为矩阵A的广义逆矩阵，通常把A的广义逆记为A^+。

图4.1为具有L个隐含层的单隐层神经网络，神经网络包含输入层、隐含层和输出层。

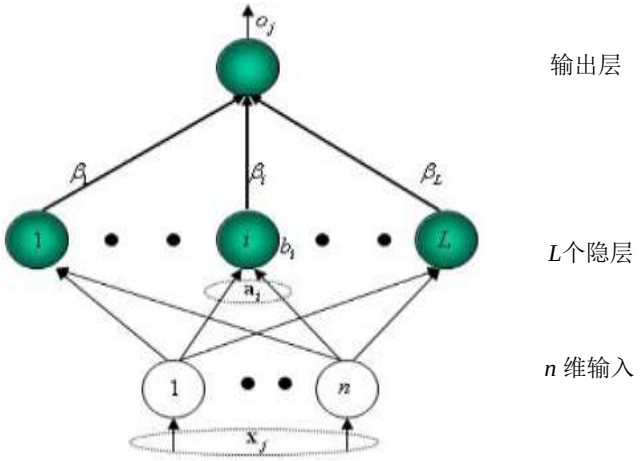

图 4.1　ELM 结构图

给定 N 个样本（x_j，t_j），其中 $x_j=[x_{j1}, x_{j2},\cdots, x_{jN}]^T$，$t_j$ 为样本的标签；a_i 为输入层与隐含层第 i 个神经元的权值，$a_i=[a_{j1}, a_{j2}, \cdots, a_{jN}]$；$b_i$ 为隐含层第 i 个神经元的偏差。网络输出为：

$$O_j = \sum_{i=1}^{L} \beta_i g(a_i \cdot x_j + b_i) \qquad j = 1,2,\cdots,N \tag{4.2}$$

其中 g(x)为激活函数，可设置成“ Sigmoid”、“Sine”或者“Hard”等函数。学习目标为：

$$\sum_{j=1}^{N} \| O_j - t_j \| = 0 \tag{4.3}$$

式 4.X 可转化为矩阵，$H\beta = T$。

$$H = \begin{bmatrix} g(a_1 \cdot x_1 + b_1) & \cdots & g(a_L \cdot x1 + b_L) \\ \vdots & \ddots & \vdots \\ g(a_1 \cdot x_N + b_1) & \cdots & g(a_L \cdot x_N + b_1) \end{bmatrix}_{N*L} \tag{4.4}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_N^T \end{bmatrix}_{L*M} \qquad T = \begin{bmatrix} T_1^T \\ \vdots \\ T_N^T \end{bmatrix}_{N*M}$$

误差函数为：

$$E = \sum_{j=1}^{N} (\sum_{i=1}^{L} \beta_i g(a_i \cdot x_j + b_j) - t_j)^2 \tag{4.5}$$

由式 4.5 可知，当输入层与隐含层的权重 a_i 和隐含层的偏差 b_i 被随机给定时，

则输出权重的凸最优解可表示为 $\hat{\beta} = H^{+}T$，其中 H^{+}为 H 的广义逆矩阵，且所得的 $\hat{\beta}$ 为最优解。

将第三章中筛选出来的训练集用于极限学习机模型的训练，其余样本作为测试集来验证模型的推广能力，不同核函数在不同间隔下测试集的正确识别率如表 4.1 所示：

表 4. 1　不同间隔下 ELM 分类器的识别率

波段间隔/核函数	Sigmoid	Sine	Hard
2 nm	85.75%	78.33%	<30%
4 nm	84.74%	77.93%	<30%
6 nm	83.31%	75.52%	<30%
8 nm	81.78%	74.39%	<30%
10 nm	80.75%	72.97%	<30%

由表 4.1 可知，相同间隔的条件下，不同的核函数识别率有较大的差别，其中 Sigmoid 函数最优；核函数为 Hard 函数时，识别率低于 30%，不适合于烟叶等级的划分。特征过多或者过少均不能取得最好的分类效果，特征较多时，识别率相对来说比较高一些，特征间可能存在较强相关性，这些特征不一定有利于等级的划分；特征过少时，留取的特征不能更好的反映烟叶等级的信息，识别率低下。基于此，对于采集的光谱特征进行筛选变得很必要。

4. 2　支持向量机

SVM（Support Vector Machine）于 1995 年由 Vapnik 等人基于统计学习理论而提出的模式识别方法。因其对小样本的模型具有良好的分类性能，被广泛应用于模式识别、控制、产品检测等多个领域。基于烟叶光谱进行烟叶等级划分时，由于烟草局所提供的样本个数有限，并且还需要将其划分为两部分，一部分作为训练集实现模型的训练，另一部分作为测试集验证模型的测试精度。考虑到样本个数有限且光谱维数高，SVM 可有效解决样本个数有限且维数高的问题。对于数据线性不可分问题，SVM 通过核函数实现数据由低维线性不可分向高维线性可分的映射；SVM 解决的是二分类问题，在解决实际多分类问题中，需要通过构造树权式或投票式分类器实现多分类。

SVM 考虑使模型的结构风险和经验风险都达到最小，在高维空间中构造最优超平面，从而达到最好的分类效果。超平面的构造如图 4.2 所示：

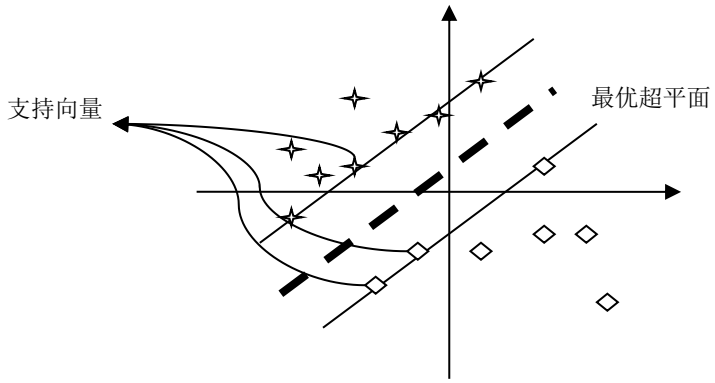

图 4.2　超平面在两分类中的结构示意图

假设 N 个训练样本为 x_i，其中 $i=1,2...N$。SVM 的输出标签值为 $y_i=\{+1,-1\}$，将 x_i 划分为两个类别，目标是在能分开类别的情况下寻找最优的分类面，即超平面 $wx+b=0$，对于 x_i 有唯一 y_i 与之相对应，能实现两分类应满足以下条件：

$$\left.\begin{array}{ll} x_i*w+b\geq+1 & for \quad y_i=+1 \\ x_i*w+b\leq-1 & for \quad y_i=-1 \end{array}\right\} \Leftrightarrow y_i(x_i*w+b)-1\geq 0 \tag{4.6}$$

进一步得出两类之间的最大间隔为 $2/\|w\|$, 在满足 4.6 的条件下转换成最小化函数：

$$\Phi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w*w) \tag{4.7}$$

引入拉格朗日函数求解最优值，函数为：

$$L=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(x_i*w+b)+\sum_{i=1}^{N}\alpha_i \tag{4.8}$$

推导过程详见文献[57], 最终得出最优的分类函数为:

$$g(\mathrm{x})=\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b \tag{4.9}$$

公式 4.9 中 $K(x_i, x)$为核函数，SVM 中实现数据低维向高维映射时常用的核函数有：

线性核函数：

$$K(x_i,x)=x_i x \tag{4.10}$$

Gauss 核函数：

$$K(x_i, x) = \exp(-\gamma \| x_i - x \|)^{\wedge 2} \tag{4.11}$$

多项式核函数：

$$K(x_i, x) = (\gamma * x_i * x + coef)^{\wedge degree} \tag{4.12}$$

Sigmoid 核函数：

$$K(x_i, x) = \tanh(\gamma * x_i * x + coef) \tag{4.13}$$

式中 *degree* 可由用户结合实际情况设置不同值，*coef* 为常数，γ 为模型最优时所求取的值，模型中的参数设置方法见文献[58]。

SVM 在解决多分类问题时采用的方式有树权式和投票式，实现流程如下图所示：

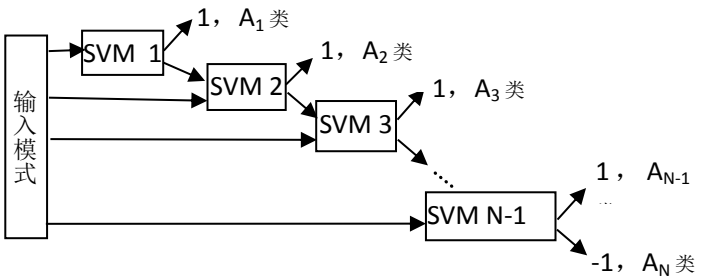

图 4.3　树权式 SVM 多分类器

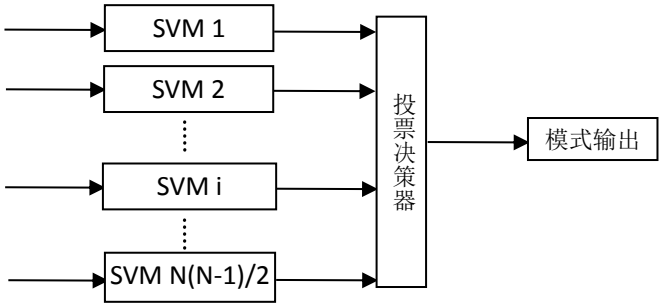

图 4.4　投票式 SVM 多分类器

图 4.3 中，对于含有 N 个等级的分类，需要构造 N-1 个分类器。如第一个分类器实现 A1 类和其余类别的区别，将其余 N-1 个类别合为一个大类；如果样本为A1类别，则只需一个分类器就可以实现样本等级的划分；如果为其它类别，再选用第二个分类器，判别样本是否属于 A2 级别，如果是则结束，否则继续采用其它分类器，直至判别出样本所属的类别。对于未知等级的样本实现其等级

的判定，最多需要 N-1 次即可实现。该方式识别的速度相比于投票式较快，但需要各个等级之间的差异性明显，否则对于有交叉混合的等级进行划分时效果不理想，易出现错判的情况；还有一种情况是，假如不止有两个分类器输出标签 1 或者输出全为 0，那么系统只能采取随机分配的原则来给样本定级。

图 4.4 中，投票式的 SVM 对于 N 个等级的分类，需构造 N(N-1)/2 个分类器。N 个等级中任意两个等级的组合构造一个分类器，共计 N(N-1)/2 种组合方式。每测试一个样本的等级时，所有的 SVM 分类器均参与识别，最终结果的判别是依据所有 SVM 分类器的输出，采用投票的形式统计各个等级所得的票数，投票决策器将获取票数最多的等级标签输出。这种方式相比树权式分类器，所构造分类器数目多，等级判定时花费一定的时间，但分类精度通常优于树权式分类器，投票式分类器在实际运用中较为广泛。

采用投票式 SVM 实现多分类，核函数为线性核函数取得最优结果，同样采用第三章中所选择出来的样本作为训练集，不同采样间隔下测试集的正确识别率如表 4.2 所示：

表 4. 2　不同采样间隔下 SVM 分类器的识别率

波段间隔	识别率				
2 nm	91.02%				
4 nm	89.49%	89.19%			
6 nm	88.02%	88.9%	88.62%		
8 nm	87.85%	87.85%	86.92%	86.49%	
10 nm	86.08%	85.05%	86.15%	85.08%	86.42%

由表 4.2 可知，SVM 分类器在采样间隔为 2nm、4nm、6nm 时，测试集的正确识别率基本保持一致，间隔大于等于 8nm 时的识别率有明显的下降。表明特征个数较多时，识别率基本上保持不变；特征个数过少时，识别率有一定的下降；特征个数过多时，波长之间的相关性比较强，并不一定有利于等级的正确划分，运用模式识别器前需要进行特征的筛选。

4. 3　近邻法

近邻法是模式识别中经典的算法之一，主要思想是通过计算待测样本与训练集中所有样本的距离或者相似度，主要包含有欧氏距离、马氏距离（适用于特征维数大于样本集个数的场合）、夹角余弦距离、相关距离等。当近邻个数为 1 时，为最简单的最近邻算法，将待测样本与距离最近的样本归于同一级别；当

选取 K 个近邻时，所取的 K 值一般为奇数，原则为找出与待测样本距离最近的 K 个样本，判断 K 个近邻中哪类占据的比重最大，将待测样本划入比重最大的那一类；加权 K 近邻原则上相似于 K 近邻，只是 K 个近邻中各个样本所占的比重不尽相同，一种方法是每个等级中训练的权重相同，即为 1/n。另一种方法是先找出 K 个近邻，加上与距离呈反比的权重，通过计算每个等级的权重之和，选取加权后取值最大的等级来为烟叶进行定级。

加权 K 近邻运用于烟叶等级检测时需要考虑的因素有：K 值的选取、训练集的选择、加权重的方式，这些因素将影响到烟叶等级划分时的速度和识别率。

4.3.1　K 近邻

K 近邻（KNN）是数据挖掘领域中经典的模式识别方法之一。其思想是通过计算未知类别的样本与训练集中所有样本的距离，找出 K 个与样本距离最近的训练集，统计 K 个近邻的类别标签，依据投票的原则将待判定类别的样本归入近邻中比重最大的那一类。

假定含有 C 个类别的 N 个训练集，W_i 类中含有 N_i 个样本，$1 \le i \le C$，假设 W_i 类中训练集的集合为 k_i，则判别函数可表示为：

$$g_i(\mathrm{x}) = k_i \ , i = 1,2,\cdots,C \tag{4.14}$$

待测样本所属级别的判定规则为：如果满足以下公式：

$$g_i(x) = \max_i k_i \tag{4.15}$$

则样本 x 被决策为：$x \in w_j$，K 近邻在二维平面中的示意图 4.5 所示：

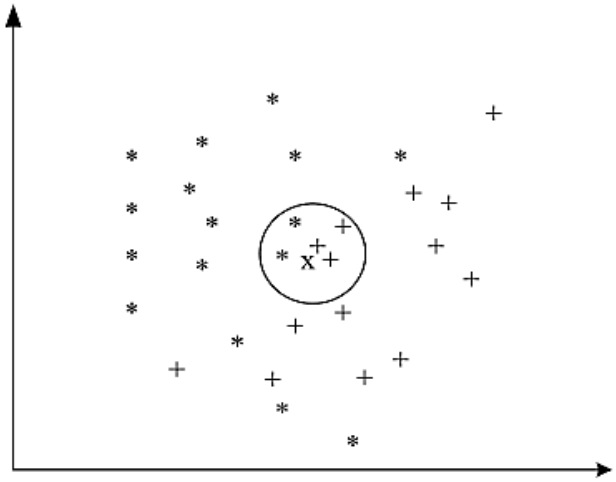

图 4.5　KNN 二维示意图

图 4.5 中 x 为等级待判别的样本，符号“*”和符号“+”代表两个不同的类别，依据 KNN 原则对样本 x 进行类别划分，当 K 取值为 5 时，所计算的 5 个近邻中包含 3 个“+”和 2 个“*”，所以待测样本的等级定为“+”类别。K 近邻法在不同间隔下的识别率如表 4.3 所示：

表 4. 3　K 值在不同间隔下的识别率(%)

波段间隔/K 值	1	3	5	7	9	11	13
2 nm	69.02	46.16	49.91	68.67	51.24	45.48	46.59
4 nm	68.58	44.61	50.80	66.56	51.24	45.70	46.81
6 nm	68.58	45.72	50.35	66.12	50.35	45.70	45.92
8 nm	68.14	44.61	50.80	66.32	51.24	46.14	46.36
10 nm	67.59	45.06	50.80	65.89	51.24	46.36	45.92

由表 4.3 可知，相同 K 值在不同间隔下测试集的正确识别率相差不大，导致这种现象的原因可能是，同一级别中的样本比较相似，较少的特征依然能表征它们之间较相似的特性；在间隔 2nm 下 K 取值为 7 时取得较好的分级效果，该方法计算复杂度较少；正确识别率较低的原因可能是不同等级训练集所筛选的个数不同，这就产生了数据倾斜的问题，需要对各个等级的训练集加上不同的权重来改善训练集个数不同的问题。

4. 3. 2　加权 K 近邻

传统 KNN 将距离测试样本最近的 K 个近邻的权重同等对待，将测试样本归入 K 个近邻中比重最大的那一类。在模式识别中，改进的方法主要有对特征进行加权，各个特征赋予不同的权重，使它们占有不同的比重[59,60]。由于测试集与各个等级之中训练集的距离或者相似度不尽相同，加之烟草公司提供的样本集存在各个等级中样本的个数有较大的差异，实际当中各个等级叶片所占的比重也不尽相同。本文考虑给训练集加上不同的权重，权重与距离或者相似度呈反比，再采取投票的形式实现进行测试样本级别的判定。通过计算待测样本与训练集间的距离或者相似度，各个训练集加上不同的权重，可有效的解决数据倾斜的问题。在训练集与测试集和传统 K 近邻完全相同的情况下，本文所提出的训练集加权的形式有以下三种：

1、同等级别中的训练集加上相同的权重，各个等级中所有的训练集加权之和均为 1。假设样本集共有 C 个等级，第 i 个等级中筛选出 n_i 个训练集，则所有的训练集组成的集合为：$X=[x_{11},x_{12},\ldots,x_{1(n1)},x_{21},x_{22}\ldots,x_{2(n2)},x_{C1},x_{C2},\ldots,x_{C(nC)}]$，其中 x_{ij}

表示第 i 个等级中含有 j 个训练样本，这样第 i 个等级中各个训练集所加的权重为 $1/n_j$。判定测试集样本等级时，遍历测试集与所有训练集的距离或者相似度，找出 K 个近邻，根据 K 个近邻所在的等级进行不同程度的加权，将测试集样本归入加权求和后取值最大的等级。采用此方法所得的结果如下表所示：

表 4. 4　改进 K 近邻方案 1 在不同间隔下的识别率(%)

波段间隔/K 值	1	3	5	7	9	11	13
2 nm	69.02	60.38	72.79	79.87	71.66	59.00	53.12
4 nm	68.58	61.04	73.02	79.42	72.33	59.45	53.78
6 nm	68.58	61.49	73.68	78.76	71.88	59.22	53.78
8 nm	68.14	60.82	74.12	78.98	72.11	59.67	53.56
10 nm	67.59	61.04	71.22	78.31	72.11	59.22	53.78

2、K 个近邻训练集所加的权重与所在的等级中训练集个数无关，和测试样本与训练集的距离或者相似度有关。判定测试样本等级时，先找出测试样本的 K 个近邻，给 K 个近邻加上与距离或者相似度呈负相关的权重，对 K 个近邻中样本所在的各个等级进行加权后求和，依据投票的原则将测试样本并入取值最大的等级。

K 个近邻中所加的权重与距离的负指数呈负相关，即距离越小所加的权重就越大，假设样本 i 距离训练集中样本 j 的距离为 d_{ij}，则训练集中样本 j 的权重为：

$$w_{ij} = e^{-\alpha * d_{ij}} \tag{4.16}$$

结合测试集正确率遍历α，在α取值为 2 时获得最优分类效果。对在训练集选中的 K 个近邻加权求和后，依据投票原则给测试集定级的结果如下表所示：

表 4. 5　改进 K 近邻方案 2 在不同间隔下的识别率(%)

波段间隔/K 值	1	3	5	7	9	11	13
2 nm	69.02	71.69	76.65	84.94	73.53	67.65	55.65
4 nm	68.58	71.91	76.10	84.37	73.19	68.54	54.99
6 nm	68.58	71.46	76.32	83.49	72.98	68.31	54.32
8 nm	68.14	71.46	75.88	82.67	71.58	66.98	54.99
10 nm	67.59	70.13	75.65	81.71	71.36	67.43	54.99

3、同时考虑各个等级中训练集个数的不同及测试样本与训练集的距离，也即方法 1 和方法 2 的结合，训练集所加的权重分为两部分，其中一部分与该等级训练集的个数有关，假设第 c 等级中含有 n_j 个样本，则训练集本身的权重为 $1/n_j$。第二部分权重与测试样本和训练集的距离呈指数关系，假设样本 i 与 c 类中训练集样本 j 的欧氏距离为 d_{ij}。则训练集所加的权重为：

$$w_{ij} = \frac{1}{n_j} e^{-\alpha * d_{ij}} \tag{4.17}$$

结合测试集正确率遍历α，在α取值为 1.9 时获得最优分类效果。对在训练集选中的 K 个近邻加权求和后，依据投票原则给测试集定级的结果如下表所示：

表 4. 6　改进 K 近邻方案 3 在不同间隔下的识别率(%)

波段间隔/K 值	1	3	5	7	9	11	13
2 nm	69.02	72.89	84.75	90.77	83.68	78.68	65.65
4 nm	68.58	72.93	84.40	90.07	83.55	78.53	64.99
6 nm	68.58	71.98	84.32	89.38	83.06	78.25	64.32
8 nm	68.14	71.46	83.68	88.67	82.48	77.16	64.99
10 nm	67.59	70.97	82.55	88.51	82.16	76.98	64.99

对比表 4.3~6 可知，改进 K 近邻法可提高烟叶等级划分的正确识别率，可见给训练集加上与距离呈负相关的权重是有效的，表明对于那些存在数据倾斜的问题，近邻法通过给训练集加上不同的权重，可以改善由于数据倾斜所造成的大类吃小类的现象。

从计算复杂度上和识别率来分析，支持向量机和极限学习机的计算量要远远大于近邻法，因为近邻法只涉及到简单的乘和除运算。加权 KNN 和支持向量机的识别率相当，分类能力都优于极限学习机。因为本文只对 13 个等级的烟叶实现智能分级，随着级别数的增加，投票式支持向量机增加的计算量要远远大于近邻法所增加的计算量，所以本文选用近邻法作为烟叶等级判定的分类器。

4. 3　本章小结

本章节构建 ELM、SVM 和近邻法等模型并实现烟叶等级的划分，SVM 和加权 K 近邻的分级效果优于 ELM，测试集的正确识别率可达 91.02%和 90.77%。给训练集加上与距离呈负相关的权重来改进近邻法，可提高测试集的正确识别率，说明给训练集加上权重后可改善数据倾斜的现象。

随着级别个数的增加，近邻法所增加的计算量要低于 SVM 所增加的计算量，所以本文将加权近邻法作为烟叶等级划分的分类器，为降低光谱数据的采集量和判定烟叶等级模型的计算复杂度，提高整个分级系统的速度，需要对采集的光谱特征进行筛选。下文将加权近邻法的识别率作为判定特征组合的好与坏，来实现特征的筛选。

5　特征筛选

模式识别中样本特征个数过多将导致分类器的识别速度变慢，而且测试精度不一定高，从而降低整个系统的实用性；选用过少的特征则将会导致等级信息表征不全，虽然有利于识别速度的提高，但整个系统的测试精度将会有所下降，因此特征筛选是模式识别中一个重要的步骤。特征选择是指从原始的 M 个特征中挑选出 m（其中 m≤M）个特征使得分类器的识别率得以改善，这样既实现了特征维数的降低，又加快了识别速度，降低了分级模型的计算复杂度。

特征选择通常包括以下几个步骤：特征子集的产生，主要有随机组合法和穷举法。假设本文所研究的烟叶有 N 维光谱特征，基于穷举法有 2^N 个特征子集组合，虽然所得的结果是最优的特征组合，因为计算量太大，现有条件基本上做不到，该法显然不实用。随机组合的方法有模拟退火法、二进制粒子群算法、遗传算法，它们按照一定的原则在特征中随机搜索，该类方法需要设置较大的迭代次数；特征子集的评价，需要构造适应度函数来判断所选特征组合的优良程度，常用的适应度函数为测试集的正确识别率；迭代终止条件，为避免算法进入无限循环，通常以最大迭代次数或者测试集的测试精度为终止条件。特征选择步骤如图 5.1 所示。

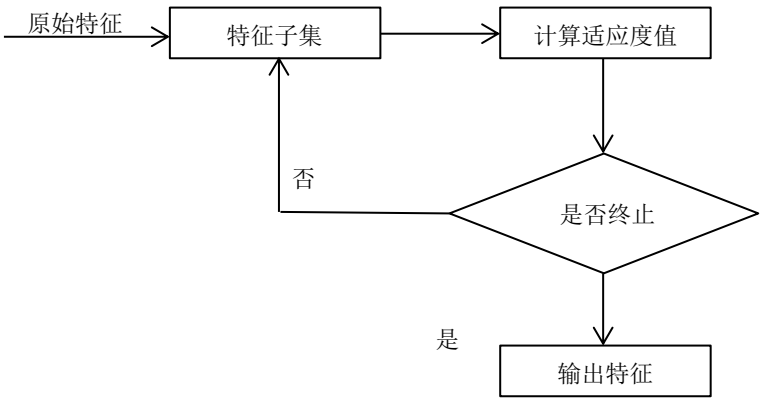

图 5.1　特征筛选的步骤

高维的光谱特征加大了模型的计算量，构建分级模型时通常会进行特征的筛选。常用的实现光谱数据降低维数的方法主要有 PCA[19]、WT[20]、SPA[24]、UVE[26]等，该类方法可降低数据的维数从而实现数据的压缩，但是不能降低需要采集的光谱特征的个数，在实际应用时采集光谱耗费的时间远大于分级所花

费的时间。另外一类降低数据维数的方法有 BPSO[35]、GA[32]、聚类[17]，这类方法结合分级模型的适应度函数可实现特征的筛选，可以降低需要采集的光谱特征的数目，极大地减少了分级系统所耗费的时间。本文分两步进行特征的筛选，包括初筛选和深度筛选。

考虑到烟叶光谱值具有连续变化的趋势，且同一级别中的各个特征在样本上的离散度值存在差异。本文结合特征的离散度，构造判别特征好与不好的判别函数 D（Discrimination），实现特征的初筛选从而降低光谱数据的维数，减少分级模型的计算量。特征深度筛选，初步进行特征筛选后，为获得更少特征数目下比较高的正确识别率，本文结合粒子群算法、遗传算法和相关系数分析法进行特征的深层筛选。

5.1 基于聚类的特征初筛选

所采集烟叶光谱的特征维数较高，可能存在冗余的特征，特征之间可能存在较强的相关性，这些因素将影响分级模型的速度和识别率。基于聚类的思想，同时考虑相同特征的类内离散度和类间离散度，构造鉴别特征好与坏的鉴别函数，在鉴别函数中寻找较优拐点，在保证识别率的前提下，实现特征的初筛选。为后期结合其他优化算法实现深层特征的选择做准备。

经过光谱仪所采集的烟叶光谱，某些特征在不同等级中的取值差异性较大。基于聚类思想，同一特征在不同等级中的差异性越大，则该特征越有利于分级。对于采集的众多光谱特征中，可采用聚类思想的方法在保证识别率的前提下对特征进行初步筛选，加快特征深度筛选优化算法的收敛速度。聚类方法可以用于光谱特征的选择，赵海东[17]等通过每个特征的类内均方差和类间均方差，分别筛选类内参数和类间参数，该法可有效的实现特征筛选，但其参数是分开筛选的，有可能漏选那些较好的特征。本文同时兼顾特征的类内离散度和特征的类间离散度，构造鉴别特征好与坏的鉴别函数 D，依据 D 值找出有效特征，具体实现方法如下：

假定含有 C 个类别的训练集，其中包含N个P维矢量；第k类中含有C_k个样本，第 k 类中第 i 个样本为：

$$X_i^{(k)} = (x_{i1}^{(k)}, x_{i2}^{(k)}, \cdots, x_{ij}^{(k)} \cdots, x_{iP}^{(k)})' \tag{5.1}$$

式5.1中，$1 \le i \le C_k$，$1 \le j \le P$，$1 \le k \le C$。k类中特征j的均值为：

$$\bar{X}_j^{(k)} = \frac{1}{C_k} \sum_i x_{ij}^{(k)} \tag{5.2}$$

离散度是指各个变量（特征）取值的差异程度，可以用变量的极差、平均差、标准差等来表征变量的稳定性。

5.1.1　类内离散度

针对不同等级的烟叶样本，同一特征在同一等级中的离散度越小，且在不同等级中的离散度越大，则该特征越能体现不同等级烟叶的特性。同一特征的类内离散度可以表明等级中样本的聚集性，类内离散度值越小越有利于等级的划分，本文将采集烟叶的光谱特征作为变量，训练集样本集中的特征j的类内离散度值计算公式为：

$$\alpha_j = \frac{1}{C}\sum_k (\frac{1}{C_k - 1}\sum_i (x_{ij}^{(k)} - \bar{X}_j^{(k)})^2) \tag{5.3}$$

依据公式5.3计算样本集中波段范围为1500nm~2400nm，采样间隔为2nm，各个特征的类内离散度如图5.2所示：

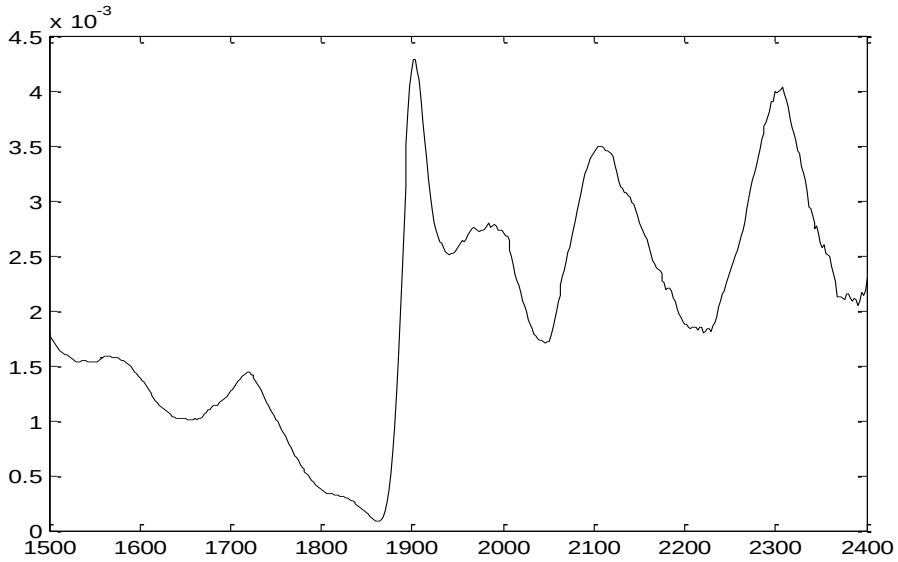

图5.2　各个特征的类内离散度

由图5.2可知，各个波长特征的类内离散度相差甚大，1500nm~1900nm波段范围内的类内离散度值相对较小，而1900nm~2400nm波段范围的类内离散度值较大。表明烟叶的光谱特征在不同波段上的差别较大，有些波长变量的聚集性较好，可以很好的表征聚类思想；某些波长变量的类内离散度值较大，需要同时考虑相同特征的类间离散度值，删减对分级不好的特征。

5.1.2　类间离散度

基于聚类思想，相同特征在不同烟叶等级中的离散度越大越有利等级的划分。同一特征的类间离散度可以表征该特征在所有级别中的差异性，在烟叶等级的划分时，其值越大越好，本文将采集烟叶的光谱特征作为变量，训练样本集中特征*j*的类间离散度值计算公式为：

$$\beta_j = \frac{1}{C-1}\sum_k (\bar{X}_j^{(k)} - \frac{1}{C}\sum_k \bar{X}_j^{(k)})^2 \tag{5.4}$$

依据公式 5.4 计算样本集中波段范围为 1500nm~2400nm，采样间隔为 2nm，各个特征的类内离散度如图 5.3 所示：

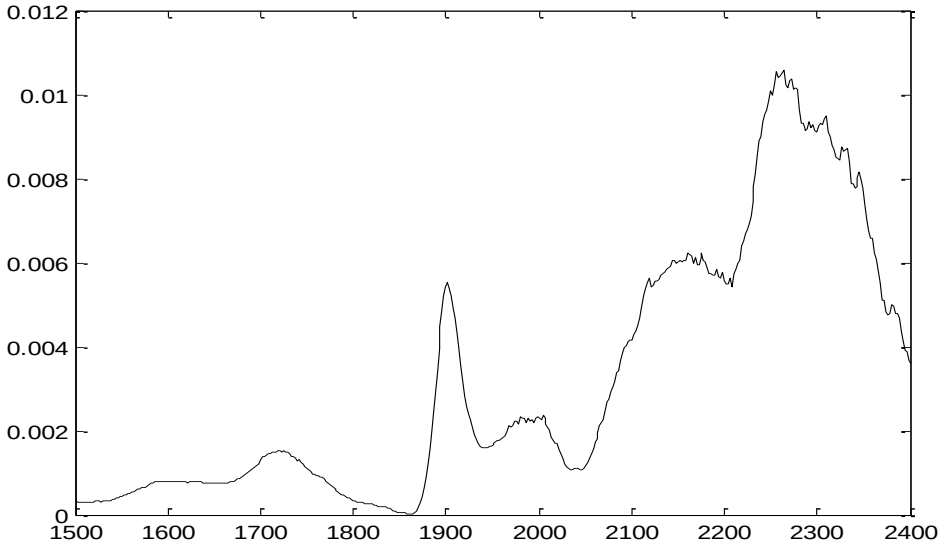

图5.3　各个特征的类间离散度

由图5.3可知，各个波长所对应特征的类间离散度相差甚大，1500nm~1900nm波段范围内的类间离散度值相对较小，而1900nm~2400nm波段范围内的类间离散度值较大，表明不同波长变量的差异性不尽相同，需要同时考虑同一特征的类内离散度来判定烟叶等级划分时是否删除该特征，以实现特征的初筛选。

5.1.3　构造鉴别函数

本文同时考虑同一特征的类内离散度和类间离散度，构造鉴别特征好与坏的函数 *D*，即相同特征的类内离散度与类间离散度比值，所构造的鉴别函数表示为：

$$D_j = \frac{\alpha_j}{\beta_j} \tag{5.5}$$

依据式 5.5 计算所采集烟叶光谱特征的鉴别值，按由小到大的顺序对光谱特征的鉴别值进行排序，排序后得到 10 个拐点，选用加权 K 近邻作为模式判别器，删除各个拐点右侧特征后的识别率如图 5.4 所示：

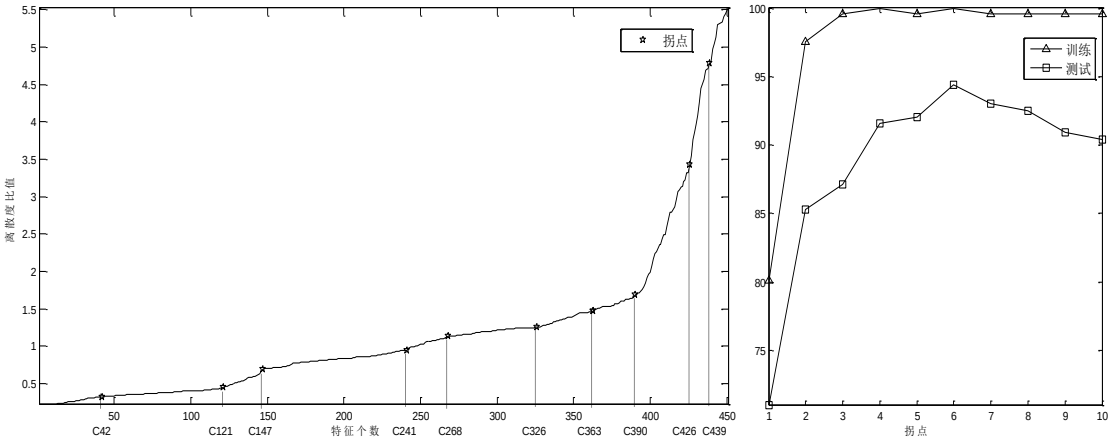

图 5.4　排序后的拐点及各拐点下的识别率

由图5.4可知，全光谱下(451个)的训练集和测试集识别率分别为100%和90.77%。依据各个特征的鉴别函数值由小到大排序后的10拐点中，从左到右10个拐点下训练集的正确率分别为：80.08%、97.51%、99.59%、100%、99.59%、100%、99.59%、99.59%、99.595、99.59%。从左到右10个拐点下测试集的正确率分别为：71.03%、85.28%、87.15%、91.59%、92.06%、94.59%、92.99%、92.52%、90.89%、90.42%。从左到右10个拐点下剩余的特征个数为：42、121、147、241、268、326、363、390、426、439。

第6个拐点下的训练集和测试集的识别率分别为100%和94.59%，分级效果明显优于其余拐点。此时余下326个特征，包含有1632：2：1822nm、1832nm、1834nm、1868：2：1928nm、1968nm、1972：2：2016nm、2020nm、2058：2：2400nm，识别率得以提高且光谱特征个数有一定的减少。本文以326个特征作为特征初筛选的结果。基于聚类思想实现特征初筛选后还需结合其他优化算法进行特征的深度筛选，以实现用更少的特征来获得更高的识别率。

5.2　深层特征筛选

为加快整个分级系统的速度，减少采集光谱特征所耗费的时间，采用 BPSO、GA 和相关系数分析对初筛选余下的特征进行深层的筛选。

5.2.1 粒子群算法

粒子群算法（PSO）基于鸟类觅食行为的思想由 Eberhart 和 Kennedy[61]提出，粒子间通过信息的交互在复杂空间中搜寻最优解，可运用于实际问题的优化。主要思路为：系统的初始化解随机给定，粒子间通过信息的交换及迭代寻找最优解。PSO 实现步骤为：

1、给定种群粒子的数目为 β，迭代次数为 M，共计产生 $M*\beta$ 个粒子，搜索空间为 α 维，所构成的种群为 $X=(x_1,...,x_i,...x_\beta)^T$，第 i 个粒子的速度和位置分别为 $v_i=(v_{i1},v_{i2},...,v_{i\alpha})^T$ 和 $x_i=(x_{i1},x_{i2},...,x_{i\alpha})^T$。

2、计算各个粒子的适应度，优化过程中适应度一般选取测试集的正确率。$p_i=(p_{i1},p_{i2},...,p_{i\alpha})^T$ 表示第 i 个粒子的最优位置，$p_g=(p_{g1},p_{g2},...,p_{g\alpha})^T$ 表示种群当前的最优位置，算法规则为所有粒子的位置和速度都趋向于个体最优和全局最优。

3、粒子速度和位置的更新公式为：

$$v_{id}^{(t+1)} = w^{(t)}v_{id}^{(t)} + c_1r_1(p_{id}^{(t)} - x_{id}^{(t)}) + c_2r_2(p_{gd}^{(t)} - x_{id}^{(t)}) \tag{5.6}$$

$$x_{id}^{(t+1)} = x_{id}^{(t)} + v_{id}^{(t+1)} \tag{5.7}$$

式 5.6 和 5.7 中 i=1, 2…β，d=1, 2… α；t 为迭代次数，w 为取值非负的惯性因子，c_1 与 c_2 为学习因子，r_1 和 r_2 的取值为 0~1 之间的随机数。

4、每迭代一次，计算各个粒子的适应度值，与其历史最优解作比较，假如效果好于历史最优解，则更新最优解。

5、比较当前迭代次数所有粒子的最优解 p_i 和种群最优解 p_g，更新 p_g。

6、判断是否满足终止条件，终止条件可以为测试集的识别率或者迭代次数，如满足终止条件则终止搜索，输出优化结果，反之返回步骤 3 继续更新。

二进制粒子群算法（BPSO）在筛选特征时，x_{id}、p_{gd} 和 p_{id} 取值为 0 或 1，各个粒子对应一组特征组合，p_i 中取值为 1 时表明特征被选中，反之表明特征没有被选中，依据 p_i 中各维的取值可判定该特征是否为有用波长。粒子位置的更新与速度的 S 型函数有关，位置更新和速度的关系为：

$$x_{id}^{(t+1)} = \begin{cases} 1, & \text{当} s(v_{id}^{(t+1)}) > p_{id}^{(t+1)} \text{时} \\ 0, & \text{当} s(v_{id}^{(t+1)}) < p_{id}^{(t+1)} \text{时} \end{cases} \tag{5.8}$$

S 型函数为：

$$s(v_{id}^{(t+1)}) = \frac{1}{1+e^{-v_{id}^{(t+1)}}} \tag{5.9}$$

式 5.8 中 $p_{id}^{(t+1)}$ 取值为 0~1 之间的随机数，BPSO 流程图如图 5.5 所示：

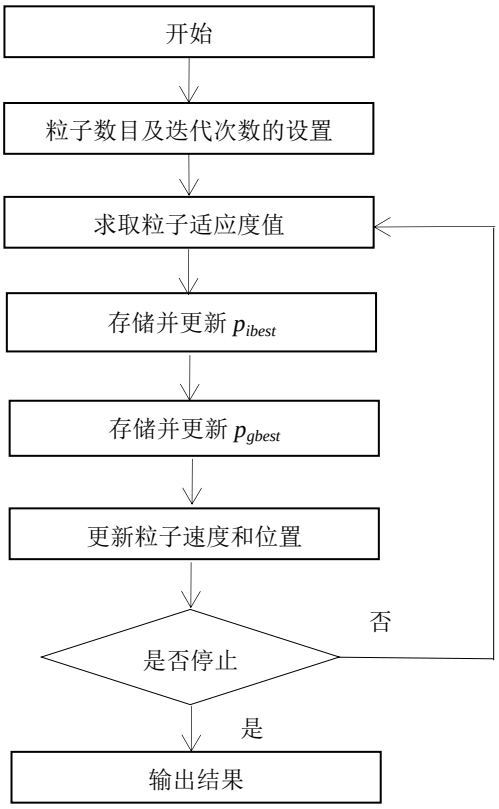

图 5.5　BPSO 流程图

本文结合 BPSO 算法筛选有效烟叶光谱特征时，粒子的数目依据经验选取为 40 个(一般取值为 20~40)，终止条件为迭代次数 1000 和测试集正确率 98%，也即迭代 1000 次或者测试集识别率达到 98%时终止特征筛选。筛选前后的特征数目、识别率及识别时间如下表所示：

表 5.1　BPSO 特征选择前后的识别情况

光谱类型	特征个数	识别率	时间
全光谱	451	90.77%	0.1315s
初筛选	326	94.59%	0.0966s
BPSO	143	93.69%	0.0519s

由表 5.1 可知，基于聚类思想实现特征初筛选后，结合 BPSO 算法进行特征的深层筛选，特征数目由原来的 451 个减少到 143 个，减少到 31.7%，采集光谱所耗费的时间可节省 68.3%；识别率由原来的 90.77%提高到 93.69%，提高了 2.92 个百分点。

5.2.2 遗传算法

遗传算法（GA）由 Holland 基于自然界生物进化机制于 1975 年提出，该算法模拟大自然中生物的选择、交叉和变异等过程[62]。优化过程中，淘汰较差的变量，保留较好的变量，最终实现问题的优化，该算法因具有高效性和较好的鲁棒性被应用于多个领域。遗传算法的主要算子有：变量的编码、随机设置群体的初始值、设定合适的适应度函数、遗传过程的操作、收敛的判据。GA 算法实现步骤为：

1、变量编码。使用 GA 时首先需要对数据进行编码处理，常用的有二进制编码、格雷码、浮点数编码等。筛选特征变量时通常选用二进制编码，用一串由 0 和 1 组成的字符串来表示染色体，染色体中取值为 1 的位置表示该变量被选中，取值为 0 时表示该变量没有被选中。

2、初始化群体。染色体中基因的值随机给定，基因的个数与烟叶光谱的维数相同，同时还需要设置染色体的个数及迭代次数。

3、适应度函数。染色体的适应度值大小可表征其生存能力，适应度值越小的染色体越容易被淘汰，GA 用于波长变量选择时适应度函数通常为测试集的正确率。

4、遗传操作，遗传机制包括有选择个体、交叉和变异。①、通常选择适应度值高的染色体作为下一代遗传的父代和母代，为提高收敛速度和避免个别基因的缺陷，染色体被选为父代和母代的概率正比于其适应度函数取值，本文选用转轮法筛选个体；②、交叉算子是遗传算法的核心，依据染色体的适应度值选择较优的染色体进行交叉组合产生新的染色体，两个染色体之间按一定的概率交换部分基因，从而实现了较好基因的被遗传到下一代，采用轮盘赌法筛选父代和母代，染色体被选中的概率为：

$$p(x_i)=\frac{f(x_i)}{\sum_{i=1}^{n} f(x_i)} \tag{5.10}$$

公式 5.10 中 n 为染色体的条数；$f(x_i)$为染色体 i 的适应度函数值；③变异是将染色体中的基因进行互补的运算，二进制编码下的运算为：将取值为 0 的位置变为 1，或者将取值为 1 的位置变为 0。变异算子可保持群体的多样性，避免出现过早收敛，同时提高了 GA 的搜索能力。选择和交叉的配合使 GA 更好的实现了

问题的最优化。

5、终止条件。为防止算法进入死循环，通常会依据实际优化问题的设置收敛判据。判据主要有迭代次数、计算时间、测试精度等，通常选用迭代次数作为算法的收敛判据。

遗传算法可实现全局搜索最优，避免了局部最优；引入与概率相关的理论来实现染色体的选择、交叉和变异，具有随机性；遗传算法的实现流程如图 5.6 所示:

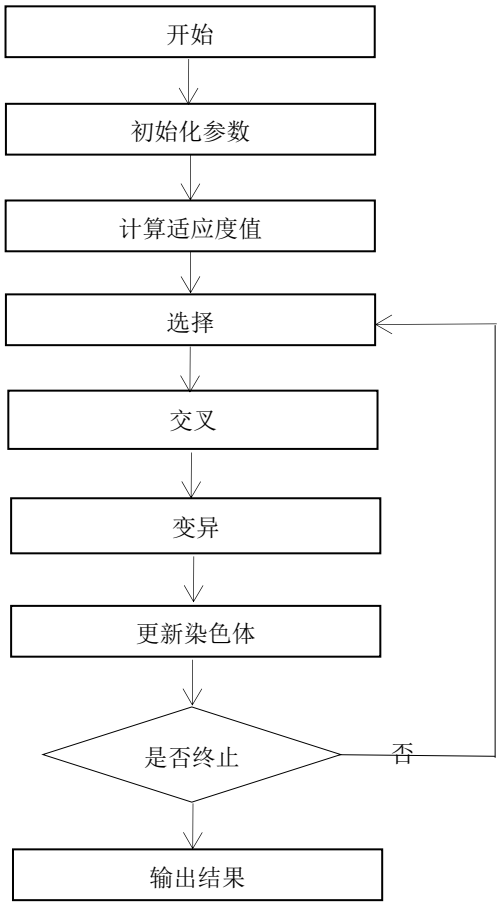

图 5.6　GA 流程图

遗传算法运用于光谱特征筛选时，采用二进制的编码形式，染色体的个数取值为 40；交叉概率为 0.7，变异概率取值为 0.05；适应度函数设计为测试集的正确识别率；终止条件为迭代次数 1000 和测试集正确率 98%，也就是迭代次数达到 1000 次或者测试集的正确识别率达到 98%时终止循环；筛选前后的特征数目、识别率及识别时间如表 5.2 所示：

表 5.2 GA 法特征选择前后的识别情况

光谱类型	特征个数	识别率	时间
全光谱	451	90.77%	0.1315s
初筛选	326	94.59%	0.0966s
GA	153	92.29%	0.0534s

由表 5.2 可知，特征初筛选后级联 GA 算法进行深层特征的筛选，筛选后余下 153 个光谱特征，相比全光谱减少了 66%；正确识别率相比全光谱提高了 1.77 个百分点，识别时间由原来的 0.1315s 减少到 0.0534s，识别率和识别速度均得以提高。但该算法的性能稍微差于二进制粒子群算法。

5.2.3 相关系数分析法

经过半监督学习进行特征初筛选后，分级模型的识别率有明显的提高，且所余下的波长数目相比全光谱有一定的减少，保留的光谱特征间可能存在很强的相关性，分析不同波长间的相关系数，可去掉相关性强的特征。

相关系数分析筛选有效光谱特征的主要思路为：在一组相关性大的波长中，筛选出与本组中其余波长相关性最大的波长，用其来表征这组波长的特性，将其余的特征滤除掉。在保证识别率的前提下选用较少的波长既可减少数据采集量，又能减少模型的计算量。筛选烟叶有效光谱特征时，波长x与波长y之间的相关系数求取公式为：

$$C_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{5.11}$$

公式 5.11 中$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}$，$\bar{y}=\frac{\sum_{i=1}^{n}y_i}{n}$。

其中 n 表示训练集的数目，x_i 表示训练集中所有样本在波长 x 上所组成的矢量，y_i 表示训练集中所有样本在波长 y 上所组成的矢量，C_{xy} 为波长变量 x 与波长变量 y 之间的相关系数，C_{xy} 值越大表明波长 x 和波长 y 之间相关性越大。基于相关系数分析法进一步筛选烟叶光谱的思路为：

假定烟叶光谱经过特征初筛选后留下m个特征变量，余下波长的类内与类间离散度的比值所组成的集合为：$u=\{u_1,u_2,\cdots,u_m\}$。设置不同的阈值，在某一确定阈值下，将 u 中最小值所对应的波长作为初选特征，在同一等级中计算该波长与 u 中其余波长的相关系数，所有的级别均做相同的运算。将在k类中与初始波长间相关系数大于所设定阈值的特征标记为C_k，遍历所有级别求取所共有

的特征，也就是特征的交集。用所选取的初始波长表征特征的交集，同时删除 u 中所出现的交集特征，至此完成了第一个波长的筛选。在经过处理后的 u 中按照上述步骤进行第二个波长的筛选，直到 u 变为空集。

特征初筛选后进行各个波长间的相关系数分析，不同阈值情况下的识别率、特征个数及分级耗时如图 5.7 所示，为寻找最优阈值，细化结果如图 5.8 所示。

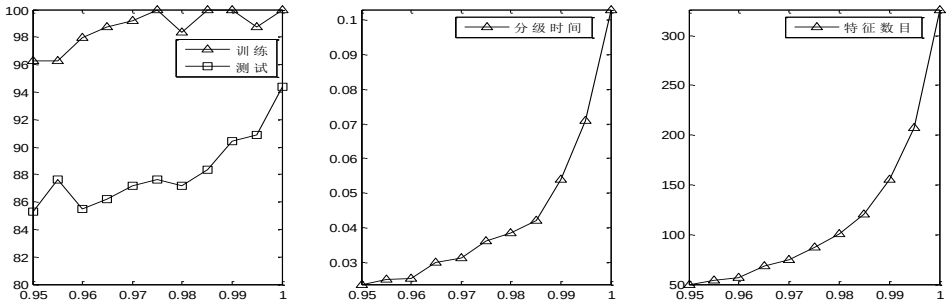

图 5.7　不同阈值下的识别率(%)、分级时间(s)、特征数目(个)

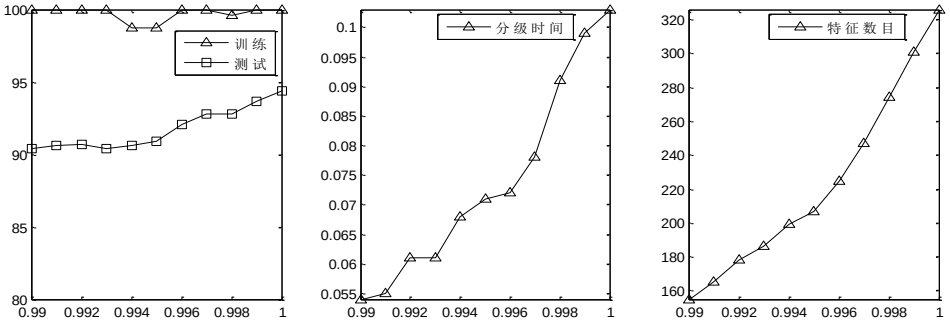

图 5.8　细化阈值后的结果

由图可知，相关系数分析运用于烟叶有效特征筛选时，测试集正确率不低于原始光谱情况下的个数为 207 个，识别率为 90.67%；在速度要求不严的时，最高识别率为 94.59%，特征个数为 326 个。相比于粒子群算法和遗传算法，该法在筛选有效波长方面有一定的不足，但通过相关系数可以分析所筛选波长之间的相关性，为定量分析打下了基础。

5.3　本章小结

基于聚类思想，同时考虑同一特征的类内离散度和类间离散度，构造鉴别特征好与坏的鉴别函数，在根据鉴别值由小到大排序后出现的拐点中筛选较好拐点，特征初筛选结果为：余下 326 个特征，测试集正确率为 94.59%。

本章节采用粒子群算法、遗传算法和相关系数分析法对经初筛选余下 326

个特征进行深层的筛选，粒子群算法取得较好的效果。经 BPSO 算法筛选特征后，烟叶光谱特征的个数由全光谱下 451 个减少到 143 个，这样采集光谱所耗费的时间可节省 68.3%；测试集的正确识别率由原来的 90.77%提高到 93.69%，提高了 2.92 个百分点。

6 总结与展望

烟草作为我国主要经济作物之一，针对当前人工进行烟叶等级的判定存在主观性强、费时且费力的缺点，本文基于近红外光谱技术对烟叶进行智能分级，包括数据的采集及预处理、孤立样本的检测、分级模型的构建、特征的初步筛选及深层筛选。本文所做工作总结如下：

1、采用型号为 UV3600 的光谱仪采集 642（13 个等级）片烟叶的反射光谱，对采集烟叶的光谱特征进行归一化预处理以消除基线漂移及外界因素所带来的噪声，使数据更好的适用于分级模型的输入模式。

2、采用夹角余弦距离、欧氏距离和相关系数法对所提供的样本集进行孤立样本的检测及训练集的选择，三种方法检测结果相同。共检测出 19 个孤立样本，这些样本不参与模型的训练，设定合适的阈值筛选将近 1/3 的样本作为训练集。

3、构建支持向量机、极限学习机和加权近邻法来实现烟叶等级的判定。支持向量机和加权近邻法的识别率相当，均高于极限学习机，加权近邻法因具有较少的计算复杂度被选用作判别器，同时考虑相同特征的类内离散度和类间离散度，构造鉴别函数实现特征的初筛选。特征个数由 451 个减少到 326，正确识别率由 90.77%增加至 94.59%，在保证识别率的前提下，特征的个数得以减少。

4、采用 BPSO、GA 和 CC 进一步进行特征的筛选。BPSO 算法取得较好的分类效果，光谱特征的个数由全光谱下的 451 个减少到 143 个，减少到原来的 31.7%，这样采集光谱所耗费的时间可节省 68.3%；测试集识别率由全光谱下的 90.77%提高到 93.69%，提高了 2.92 个百分点。

本文后续工作及展望：

首先，对烟叶进行等级划分时，仅仅考虑的烟叶的光谱特征，没有考虑人工进行等级划分时所参考的烟叶的颜色、纹理、形状等特征，考虑结合光谱特征和图像特征可能会取得较好的分级效果。

其次，改进 K 近邻法只是对训练集加上不同的权重，不同的特征并没有加权，如果对各个特征加上不同的权重，分类效果可能会更好。

最后，考虑到实验样本个数有限，本文在样本集中直接进行训练集的选择。对于数据特别多的情况下，本文所提筛选训练集的方法可以用于训练集的约减。

参考文献

[1] 国家技术监督局. GB2635-1992烤烟[S]. 北京:中国标准出版社,2006.

[2] 刘剑君，申金媛，彭丹青,等. 基于SVM的烟叶光谱分级[J]. 通信技术，2009, 42(11):197-199.

[3] 王夏，贺立源. 烤烟烟叶反射和透射图像的同步图像分割[J]. 武汉大学学报信息科学版, 2014, 39(8):998-1002.

[4] 马孝腊，申金媛，刘润杰,等. 基于密度的稀疏表示及其对烟叶分级研究[J]. 江苏农业科学, 2016, 44(9):371-373.

[5] 马建元，伍铁军. 基于图像处理和模糊识别的烟叶分级技术研究[J]. 机械制造与自动化, 2011, 40(1):90-93.

[6] 王毅，马翔，温亚东,等. 应用近红外光谱分析不同年度工业分级烟叶的特性[J]. 光谱学与光谱分析, 2012, 32(11):3014-3018.

[7] 刁航，吴永明，杨宇虹,等. 田间原位光谱的鲜烟叶成熟度判别模型的研究[J]. 光谱学与光谱分析, 2016, 36(6):1826-1830.

[8] 陆婉珍. 现代近红外光谱分析技术[M]. 中国石化出版社, 2007.

[9] 申振宇，申金媛，刘剑君,等. 基于神经网络的特征分析在烟叶分级中的应用[J]. 计算机与数字工程, 2012, 40(7):122-124.

[10] Keyworth D A. Determination of water by near-infrared spectrophotometry [J]. Talanta, 1961, 8(7):461-469.

[11] Martens M M H. Near-Infrared Reflectance Determination of Sensory Quality of Peas [J]. Applied Spectroscopy, 1986, 40(3):303-310.

[12] Kelly J J, Barlow C H, Jinguji T M, et al. Prediction of gasoline octane numbers from near-infrared spectral features in the range 660-1215 nm[J]. Anal. Chem.; (United States), 1989, 61:4(4):313-320.

[13] 刘华波，贺立源，马文杰. 基于反射与透射图像结合的烟叶自动分级研究[J]. 应用基础与工程科学学报, 2009, 17(3):343-350.

[14] 向金海，杨申，樊恒,等. 基于稀疏表示的烤烟烟叶品质分级研究[J]. 农业机械学报, 2013, 44(11):287-292.

[15] 张乐明，申金媛，刘剑君,等. 概率神经网络在烟叶自动分级中的应用[J]. 农机化研究, 2011, 33(12):32-35.

[16] 李航，申金媛，刘润杰,等. 基于聚类的烟叶标准库和特征选择[J]. 黑龙江农业科学, 2016, 38 (4):140-143.

[17] 赵海东，申金媛，刘润杰,等. 基于聚类的烟叶近红外光谱有效特征的筛选方法[J]. 红外技术, 2013, 35(10):659-664.

[18] 刘剑君，申金媛，张乐明,等. 基于红外光谱的烟叶自动分级研究[J]. 激光与红外，2011, 41(9):986-990.

[19] 王一丁，赵铭钦，付博,等. 利用可见-近红外光谱鉴定不同香型风格烤烟的方法[J]. 中

国烟草科学, 2015, 36(6):88-93.
[20] 彭丹青, 申金媛, 刘剑君,等. 基于径向基网络的烟叶光谱分级[J]. 农机化研究, 2009, 31(10):15-18.
[21] 章英, 贺立源, 叶颖泽,等. 基于LS-SVM的烤烟烟叶产地判别[J]. 湖北农业科学, 2012, 51(3):583-585.
[22] 周金治, 唐肖芳. 基于相关系数分析的脑电信号特征选择[J]. 生物医学工程学杂志, 2015, 32(4):735-739.
[23] 于晶, 温亚东, 王萝萍,等. 近红外光谱定性分析中的特征波长筛选研究[J]. 光谱学与光谱分析, 2013, 33(11):2973-2977.
[24] Moreira E D, Pontes M J, Galvão R K, et al. Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection[J]. Talanta, 2009, 79(5):1260-1264.
[25] 孙俊, 张国坤, 毛罕平,等. 基于介电特性与回归算法的玉米叶片含水率无损检测[J]. 农业机械学报, 2016, 47(4):257-264.
[26] Koshoubu J, Iwata T, Minami S. Application of the Modified UVE-PLS Method for a Mid-Infrared Absorption Spectral Data Set of Water-Ethanol Mixtures [J]. Applied Spectroscopy, 2000, 54(1):148-152.
[27] 李倩倩, 田旷达, 李祖红,等. 无信息变量消除法变量筛选优化烟草中总氮和总糖的定量模型[J]. 分析化学, 2013, 41(6):917-921.
[28] Xu H, Liu Z, Cai W, et al. A wavelength selection method based on randomization test for near-infrared spectral analysis [J]. Chemometrics & Intelligent Laboratory Systems, 2009, 97(2):189-193.
[29] Abdullah S, Sabar N R, Ahmad Nazri M Z, et al. An Exponential Monte-Carlo algorithm for feature selection problems [J]. Computers & Industrial Engineering, 2014, 67(1):160-167.
[30] Norgaard L, Saudland A, Wagner J, et al. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy[J]. Applied Spectroscopy, 2000, 54(3):413-419.
[31] 陈晓辉, 黄剑, 付云侠,等. 基于iPLS和CARS数据融合技术的波长选择算法[J]. 计算机工程与应用, 2016, 52(16):229-232.
[32] Li H, Shen J, Kong Y, et al. Screening the Effective Spectrum Features of Tobacco Leaf Based on GA and SVM[C]// International Conference on Sensor Network and Computer Engineering. 2016:201-204.
[33] 马红辉, 王中江, 袁天军,等. 烟草中淀粉近红外光谱变量的筛选及校正模型的建立[J]. 烟草科技, 2015,48(8) :37-43.
[34] Li H, Wu J, Huang K, et al. Study of Flue-Cured Tobacco Classification Model Based on the PSO-SVM [J]. Research Journal of Applied Sciences Engineering & Technology, 2013, 5(19):4671-4676.
[35] 李航, 赵海东, 申金媛,等. 基于BPSO和SVM的烟叶近红外有用特征光谱选择[J]. 物理实验, 2015, 36 (6):8-12.
[36] 杨帆, 申金媛. 基于BPSO和SVM的烤烟烟叶图像特征选择方法研究[J]. 湖北农业科学,

2015, 54(2):449-452.
[37] 王动民, 张军, 赵滨. 基于模拟退火算法的近红外光谱定标模型的简化[J]. 光谱实验室, 2006, 23(5):921-925.
[38] 石吉勇, 邹小波, 赵杰文,等. BiPLS结合模拟退火算法的近红外光谱特征波长选择研究[J]. 红外与毫米波学报, 2011, 30(5):458-462.
[39] Shamsipur M, Zareshahabadi V, Hemmateenejad B, et al. An efficient variable selection method based on the use of external memory in ant colony optimization. Application to QSAR/QSPR studies [J]. Analytica Chimica Acta, 2009, 646(1–2):39-46.
[40] 丁小辉, 李华朋, 张树清. 基于多态蚁群算法的高光谱遥感影像最优波段选择[J]. 遥感技术与应用, 2016, 31(2):275-284.
[41] 张婷, 俞飞, 肖少红. 湖北省主产烟区烟叶化学成分含量特征分析[J]. 中国烟草学报, 2010, 16(3):24-27.
[42] 徐秀红，许家来，杨永花，等. 烤烟烘烤性状与烟叶化学成分的相关性[J]. 中国烟草学报, 2014,20(6):103-106.
[43] 申钦鹏, 张霞, 张涛,等. 基于烟叶化学成分烤烟香型分类模型的建立[J]. 湖北农业科学, 2015, 54(5):1220-1226.
[44] 张四平. 不同变黄时间对烟叶叶绿素及其主要化学成分含量的影响[J]. 湖南农业科学, 2014, 43 (20):75-76.
[45] 赖炜扬, 林凯, 鹿洪亮,等. 再造烟叶正交优化提取及其化学成分和致香成分分析[J]. 厦门大学学报(自然版), 2016, 55(1):144-148.
[46] 刘晶, 向海英, 王保兴,等. 应用近红外技术快速测定再造烟叶产品的主要化学成分[J]. 昆明理工大学学报(自然科学版), 2015, 40 (5):108-113.
[47] 田旷达, 邱凯贤, 李祖红,等. 近红外光谱法结合最小二乘支持向量机测定烟叶中钙、镁元素[J]. 光谱学与光谱分析, 2014, 34(12):3262-3266.
[48] 周淑平, 程贵敏, 李卫红,等. 近红外光谱法快速测定烤烟中钙、镁、铁、锰和锌的含量[J]. 贵州农业科学, 2007, 35(1):28-31.
[49] 付秋娟, 王树声, 窦玉青,等. 烟草根中N、K、Ca、Mg的近红外光谱分析[J]. 烟草科技, 2006,34(10):35-37.
[50] 刘岱松, 金兰淑, 杨朝辉,等. 烤烟烟叶钾含量的近红外光谱法快速测定[J]. 土壤通报, 2010,41 (2):417-419.
[51] 章平泉, 杜秀敏, 王春玲,等. 近红外光谱法测定烤烟中磷酸根含量的研究[J]. 光谱实验室, 2008, 25(2):244-248.
[52] 刘海峰, 刘守生, 姚泽清. 文本分类中基于训练样本空间分布的K近邻改进算法[J]. 情报学报, 2013, 32(1):80-85.
[53] 杨金福, 宋敏, 李明爱. 一种新的基于距离加权的模板约简K近邻算法[J]. 电子与信息学报, 2011, 33(10):2378-2383.
[54] Huang G B, Zhu Q Y, Siew C K.Exreme Learning Machine: Theory and application [J]. Neurocomputing, 2006, 70(1):489-501.
[55] 王杰, 毕浩洋. 基于正则极限学习机的烟草病毒病预测[J]. 郑州大学学报理学版, 2013, 45(4):58-62.

[56] 王杰，毕浩洋. 基于极限学习机的烟叶成熟度分类[J]. 烟草科技, 2013,42(5):17-19.
[57] 高隽. 人工神经网络原理及仿真实例[M]. 机械工业出版社, 2003.
[58] 李洋. 关于SVM的那点破事[EB/OL]. http://www.matlabsky.com/thread-10966-1-1.html.
[59] 方天红，陈庆虎，廖海斌,等. 融合纹理与形状的人脸加权新特征[J]. 武汉大学学报信息科学版, 2015, 40(3):321-326.
[60] 罗会兰，杜芳芳，孔繁胜. 像素点特征加权的尺度自适应跟踪算法[J]. 通信学报，2015, 36(10):200-210.
[61] Kennedy J, Eberhart R. Particle swarm optimization: Proceedings of IEEE international conference on neural networks, 1995[C]. Perth, Australia, 1995:1942-1948.
[62] 褚小立，袁洪福，王艳斌,等. 遗传算法用于偏最小二乘方法建模中的变量筛选[J]. 分析化学, 2001, 29(4):437-442.

致谢

光阴似箭，随着论文的撰写及工作的敲定，即将离开对我人生影响极其重要的母校——郑州大学。回想在这三年研究生生涯当中，从老师及身边同学身上学到的不仅有相关专业知识，更有的是高尚的品质和为人处世的方式，及如何快速、高效的处理所遇到的问题。这些知识使我获益匪浅，在此由衷的感谢所有关心和指导过我的老师、同学、家人和朋友。

首先感谢我的父母及导师申金媛教授。父母含辛茹苦的把我养大，每天起早贪黑的努力工作挣钱供我上学，感谢你们给了我来之不易的学习的机会。申老师拥有渊博的知识、广阔的视野、宽厚仁慈的胸怀和严谨的治学态度。在学习上，申老师对我们悉心指导，对于我们不懂的问题，申老师不耐其烦的给我们以讲解，让我们遨游于知识的海洋；在生活当中，申老师和蔼可亲、平易近人、善解人意，对待我们像待自己的孩子，让我们深深的体验到浓厚的师生情。是您帮我打开了知识之门，让我感受到知识的力量。感谢您在我找工作之时给予的帮助和指导，这些指导和帮助使我获益匪浅，我将牢记您的教诲。申老师，谢谢您！

同时，感谢刘润杰副教授。感谢您每周都抽出时间来给我们开例会，为大家在学习上的沟通和交流提供了一个良好的平台，大家一起分享所学的知识，共同解决阅读文献时发现的疑点。这些经历给我的求学之路增添了一道亮丽的风采，大家在学习上互勉互励、共同进步。刘老师的正确引领，是我们学习路上的启明灯，感谢刘老师在学习和生活上给予的关心和指导。

感谢穆晓敏教授，谢谢您在项目期间在实验设备、学习和生活上给予的帮助，感谢您为我们提供了良好的科研环境，在此向您致以衷心的感谢与祝福。

感谢身边的同学：金鸽、胡敬旭、赵刘威、马宗浩、耿亚楠及 205 实验室的伙伴们等，是你们平时陪我吃饭、聊天、一起玩耍。有困难时大家相互帮忙解决，一起分享学习和生活中的乐趣。你们的陪伴使我的生活充满了乐趣，我将珍惜这些深厚的友谊之情。

最后，感谢我的对象王萍，感谢你在生活和学习上对我的关心与帮助。在生活上对我无微不至的关怀，使我内心倍感温暖；在学习上，虽说所学专业相差甚远，她给的建议都很受用，开拓了我的视野。感谢你陪我度过的生活中的点点滴滴，使我的生活更加丰富多彩。

个人简历、在学期间发表的学术论文与参与项目

个人简历

李航：男　　198906
本科：中国海洋大学青岛学院　　专业：通信工程
硕士：郑州大学　　专业：信息与通信工程

发表的论文

[1] 李航, 赵海东, 申金媛, 等. 基于BPSO和SVM的烟叶近红外有用特征光谱选择[J]. 物理实验. 2015, 35(6):8-12.

[2] 李航, 申金媛, 刘润杰, 等. 基于聚类的烟叶标准库和特征选择[J]. 黑龙江农业科学, 2016, 38(4):140-143.

[3] Hang li,Jin-yuan Shen,Zhong-ji Cheng,*et al*. Screening the effective spectrum features of tobacco leaf based on GA and SVM.International Conference on Sensor Network and Computer Engineering[C].Xian, China, 2016:201-204.(ISTP检索)

[4] 申金媛, 李航, 刘润杰, 等. 基于相关系数的有效特征光谱筛选方法[J]. 郑州大学学报(理学版)(中文核心、已录用).

参与的项目

基于数据挖掘算法的烟叶分类方法的研究（河南省烟草公司科技计划项目）