南京理工大学本科生科研训练
项目结题报告

项目名称：

基于 Spark 云计算平台的高光谱分类并行算法与系统

资助类别： 校级普通项目

项目单位： 计算机科学与工程学院

负责人姓名： 蒋旭钊 **学号：** 918106840727

所在学院： 计算机科学与工程学院

所属专业： 计算机科学与技术

指导教师姓名： 吴泽彬 **职称** 教授

起止时间： 2020-04-25 至 2021-09-30

教务处　制

项目名称		基于 Spark 云计算平台的高光谱分类并行算法与系统			
资助类别		校级普通项目			
项目组成员	姓名	学号	学院/专业	手机号	角色
	蒋旭钊	918106840727	计算机科学与工程学院/计算机科学与技术	19825000890	负责人
	李硕	918106840731	计算机科学与工程学院/计算机科学与技术	13027572400	组员
	王俊玺	918106840747	计算机科学与工程学院/计算机科学与技术	15231918671	组员
指导教师	姓名	职称	学院	联系电话	电子邮箱
	吴泽彬	教授	计算机科学与工程学院		wuzb@njust.edu.cn
	姓名	职称	学院	联系电话	电子邮箱

1. 研究内容及研究方法。

高光谱遥感数据的分类识别以区分并识别图像中多个目标物为目的，是高光谱遥感技术领域的一个研究热点。高光谱数据包含丰富的空间、辐射和光谱信息，给遥感图像的分类识别带来了机遇。一方面，高光谱数据光谱分辨率高，能够获取地物精细的光谱特征曲线，可以根据需要选择或者提取特定的波段来突出目标特征；另一方面，定量化的连续光谱曲线数据为地物光谱机理模型引入图像分类提供了条件。同时，高光谱遥感数据量大，现有的单机环境很难满足这类大数据量遥感数据的高效处理需求，迫切需要采用云计算的技术进行分布式的并行优化研究。

本课题在 Spark 云计算平台上研究空谱联合的核稀疏表示的高光谱图像分类技术（Spatial-Spectral Kernel Sparse Representation for Hyperspectral Image Classification on Spark），利用云平台的分布式并行计算能力，优化设计高光谱图像分类的并行算法和系统，提升算法的处理效率和处理能力。

由于我们作为本科生，相关方面的知识储备不是特别的牢固，所以我们需要针对我们的研究内容进行比较好的划分，从中提取出几个主要的知识板块，在课余时间进行补足和增强。这里就用到了“分而治之，逐个攻破”的研究方法。首先，我们将我们的课题大致划分成了三个主要的部分：1）高光谱图像的处理与研究，2）以 Java 语言编写的空谱联合的核稀疏表示分类算法及其在 Spark 上的并行算法的优化策略，3）以空谱联合的核稀疏表示分类算法为后台的 SpringMVC 前后端系统设计。确定了我们的研究方向和目标，接下来就是对于每个方向设计出我们的学习方案

以及项目计划，在每个方向的学习和实践中，我们采用了“理论为先，反复实验”的科研方式。

在第一个方向，高光谱的图像处理与研究方面，我们首先是查找相关的书籍进行研读，《高光谱图像》让我们建立起了对这种图像格式的初步认知，了解到其在空间特征和光谱特征中蕴含的大信息量，也有利于我们理解课题中的“空谱联合”的意义，同时也了解到这庞大的信息量，正是单机处理难以跨越的瓶颈，更加明确了云计算平台并行优化的动机。理论基础的阅读还不足以让我们开展研究，我们又通过 ENVI 这类软件，通过输入高光谱图像的图片，让我们进一步实践体验了这种图像格式蕴含的信息，整体有了一个更好地理解。通过看《数字图像处理》帮助我们了解到各种数字图像的处理方式，包括均值滤波，归一化等我们之后的实验需要用的数据处理方式。

在此理论的基础上，我们才进一步开展了空谱联合的高光谱图像分类的 matlab 代码的研究。首先我们观察到，已有的 matlab 代码用到的高光谱图像数据集 Indian Pines 是以.mat 的文件进行存储的，包括二维和三维的数据格式，其中的数据信息存储是无法直观的看出的。这就首先需要我们能够从已有的 matlab 代码的处理过程中，发现高光谱图像的主要数据存储格式以及如何从三维转换为二维的，最后才能完成高光谱图像的读取工作。通过研究，我们发现，三维的高光谱图像的前两个维度存的是图像的空间信息，第三个维度存的是图像的光谱信息，如果我们要进一步直观地处理，需要将其读取为二维的形式，这样子才有利于我们之后按索引取得相应的数据。于是我们通过协调光谱图像空间维度的行和列的关系，将二维度的空间坐标转换为一个维度，然后再将光谱信息作为另一个维度，这样子就形成了行是光谱信息，列是空间信息的二维图像，完成了我们的数据格式的约定。在此基础上，通过运用归一化的预处理方法，我们得到了一个完备的训练与测试数据集，初步完成了第一个模块的任务。

在第二个方向，空谱联合的核稀疏表示分类算法的编写和 Spark 平台上的并行优化这个方面，我们也可以分为两条路线进行，1）了解空谱联合的核稀疏表示分类算法的机理以及完成它的串行代码的编写，2）以串行代码为基础，以 scala 语言编写 Spark 上的并行优化程序。

首先我们仍然是查阅和分类算法有关的资料，作为一项最新的技术，稀疏表示被广泛应用于许多领域，如图像去噪，图像超分辨，人脸识别，光谱解混和高光谱图像分类，我们的项目就用它来做高光谱图像分类。稀疏表示分类（sparse representation based classification, SRC）假设相同类别的特征包含在同一个低维空间中，因而未知类别的样本可以由结构化字典中的训练样本的线性组合稀疏表示。虽然 SRC 在高光谱图像分类中取得了好的分类结果，但是它很难区分那些线性不可分的数据。为了克服这个缺陷，核稀疏表示分类(Kernel SRC，KSRC）被用来描述特征的非线性相似性。KSRC 采用核技巧将数据投影到特征空间，从而数据变得线性可分离。KSRC 将高光谱图像作为一组无序的像元处理，并没有考虑它的图像特征。现有研究表明，利用空间信息进行高光谱图像分类能够有效提供分类精度。本实验在核特征空间中执行空间滤波，设计了一种邻域滤波核（Neighboring Filtering，NF）来描述邻域像元间的空间相似性。与组合核（Composite Kernel CK）方法只融入空间信息不同的是，项目中提出的方法仅需建立核 NF，就可以同时包含光谱和空间相似性。在 KSRC 中，核矩阵可以预先计算，但是这个核矩阵计算包含巨大的计算量，因此我们需要考虑它的优化算法，于是便尝试用 Spark 这个分布式计算框架来提升运算速度,利用云平台的分布式并行计算能力，优化设计高光谱图像分类的并行算法和系统，提升算法的处理效率和处理能力。

然后就是如何将已有的 matlab 代码优化成 Spark 云平台上的并行分类代码。我们就此产生了一系列思考，Spark 云平台是以 scala 为原生语言的，而且支持 java 并且 scala 和 java 共通，那么接下来就是如何从 matlab 代码中了解到整个的编写逻辑，然后按照此逻辑编写出相应的 java 串行代码。此外就是如何并行优化的问题，我们可以先编写出串行的 java 代码，然后在此基础上，观察出耗时最长的处理部分，然后在此基础上进行 scala 并行代码的编写。首先便是理解 matlab 中的光谱联合的高光谱图像分类代码，主要采用的是稀疏字典的分类方法，通过以一定量的训练样本作为字典，然后测试样本以字典的数据可以进行线性表示和预测，在预测的基础上分别算出测试样本在不同类上的损失，于是测试样本所处的类别就是损失值最小的那个类别。整体

的内容从直观上理解并不难，但是在理论公式推导求解方面，这是一个带约束的多目标分类问题，其中用到了交替乘子法 ADMM。为了能够理解相应的理论和代码，蒋旭钊通读了周志华的《机器学习》中的分类方面的部分，做了一些比较好的笔记并和组员们进行交流和讨论。然后就是串行 Java 代码的编写，李硕首先进行了部分模块的编写工作，但由于其中缺少对应的图像数据处理模块且相关的接口缺少注释不能良好地对接，不能够很好地用在我们的项目中。蒋旭钊以此为参考，对代码进行了重写，首先就是数据的选择，由于需要在编程的过程中进行反复的测试，于是我们选择了 Indian Pines 中比较小部分的数据集，通过编写 Java 的数据图像接口，能够将.mat 文件进行读入处理。然后就是空谱联合分类算法的代码，其中比较难理解的就是如何进行空谱联合的核处理的，通过研究和思考，我们发现主要是通过高斯核对领域滤波中的像素进行权重的分配，通过两两像素间的领域滤波的核函数进行相乘相加，最后得出了核矩阵的一个元素。首先对已有数据集的所有图像像素进行核方法的处理，最后得出了核矩阵的形式，在此基础上，才能进一步进行分类。有点类似于分类数据的“预处理”，因此，我们就将整个算法实现分成了空谱联合核的计算（Spatial-Spectral Kernel Computing）和用交替方向乘子法（Alternating Direction Method of Multipliers）解决图像稀疏表示的分类（Sparse Representation Classification）两大部分，在进行空谱联合核的计算上面，我们发现核矩阵中的每个元素都具有相同的计算步骤。每个元素都关联了原图像中的两个像素点，规定一个合适的窗口大小，两图像像素在对应的窗口下，用高斯核计算权重，最后权重中隐含了空间距离信息；然后通过点乘的形式蕴含光谱信息，配合前面窗口中框出的图像具有的空间信息，最后重复这样的方式计算出了光谱联合核矩阵中的每个元素。我们将训练数据和测试数据以索引的形式组合，最后只需要计算出一个大的核矩阵然后以索引找出对应的训练数据产生的核矩阵和测试数据产生的核矩阵就能够进行下一步的分类。在得到了训练样本和测试样本分别对应的核矩阵以后，我们可以使用交替乘子法来进行分类的计算，当迭代到达一定的轮次或者迭代的损失变化小于某一个范围以后，我们停止迭代。最后得出的是所有测试样本在字典数据上的损失，通过对这些损失进行求和计算，可以得出最终每个测试数据所处的类别。以上就完成了空谱联合的核稀疏表示的串行 java 代码的实现。

通过输出比对空谱联合核的计算和图像稀疏表示的分类各自的执行时间，发现前者复杂度和计算量最高，于是将并行优化重心集中在了空谱联合核的计算上面。蒋旭钊通过研究发现，Spark 并行优化主要是将原来计算量大的数据分发到不同结点上进行处理，合理利用 mapreduce 方法，最后可以在 driver 端得到最终的实验结果。其中需要对已有的数据进行合理的组织和分片，这样子能保证分发到 worker node 上的结点可以并行工作，最后通过 driver 的 collect 聚集操作进行汇总。

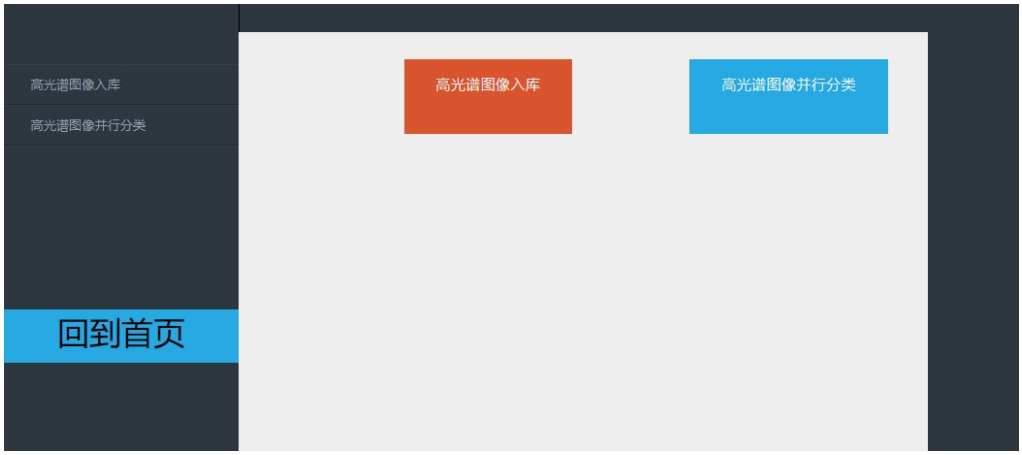
首先便是数据的分片和组织工作，和 Spark 云计算平台对应的有一套存储系统 HDFS（Hadoop File System）,首先通过配置 hdfs-site.xml 和 hdfs-core.xml 构造我们的伪分布式测试平台，然后通过 hadoop fs 指令将我们的数据以合适的分片数量上传到 hdfs 中。我们搭建的分布式平台是四台虚拟机，其中一台作为 driver 端，四台都作为 data node 进行数据处理。通过分析我们可以得到，核矩阵的每个元素都对应着一个像素，每个像素都会有窗口将其框出，理论上每个元素都会用到整幅高光谱图像的所有信息，所以我们需要将原来的高光谱图像 SSKSRCNF_img 分发到每个节点上，然后对应于我们要训练的总的数据样本，我们将其划分为 4 块，每块在各自的节点计算出核矩阵的部分，最终通过 collect 指令聚合到 driver 端，理论上可以达到 4 倍左右的加速比。至此完成了第二个模块的任务。

在第三个方向，以空谱联合的核稀疏表示分类算法为后台的 SpringMVC 前后端系统设计。我们首先是需要进行网页设计的理论知识的学习，它是目前最流行的互联网设计框架之一，主要包含了三部分。1）Model（模型）：数据模型，提供要展示的数据，因此包含数据和行为，可以认为是领域模型或 JavaBean 组件（包含数据和行为），不过现在一般都分离开来：Value Object（数据 Dao） 和 服务层（行为 Service）。也就是模型提供了模型数据查询和模型数据的状态更新等功能，包括数据和业务。2）View（视图）：负责进行模型的展示，一般就是我们见到的用户界面，客户想看到的东西。3）Controller（控制器）：接收用户请求，委托给模型进行处理（状态改变），处理完毕后把返回的模型数据返回给视图，由视图负责展示。也就是说控制器做了个

调度员的工作。它是 Spring 在网站设计方面的运用，通过依赖注入和控制反转等特性，能够实现让服务器托管大部分工作，有效降低前后端的耦合度，提升程序开发的效率。

在学习完理论基础的前提下，蒋旭钊负责了主要的系统的开发，其中在 Controller 中设计了主页 IndexController，文件上传 UploadController,调用 Spatial-Spectral Kernel Sparse Representation Classification 算法的 ClassifyController,其中算法的主要实现在 Service 层中，包括用 Hadoop API 从远程的 HDFS 库中调用分片后的数据，然后调用 Spark API 启动集群对数据进行分类处理。相关的页面如下：

主页页面：



高光谱图像上传页面：

2.主要的科学发现和创新之处，并列出具体的内容和必要的数据。

1）基于 matlab 的空谱联合的高光谱图像核分类算法，在 IDEA 中用 Java 实现了对应模块的编写，HSIhdr 实现了高光谱图像头文件的读取和处理。ByteData 结合调用 DataInputFormat 和 DataRecordReader 实现了高光谱文件从 HDFS 的读入模块，其中包括 matlab 中的 reshape3D_2D 把原始 3D 高光谱图像转换成 2D 的，同时将数据进行标准化。

对数据进行核函数计算的模块通过 PosCal, Totalijw2DCal, IjwWeightCal, KtotalCal 进行综合计算。

最后调用 Tools 中的分类模块对处理好的核矩阵进行分类，其中包括 stop 终止函数，soft 软阈值函数以及交替乘子法 ADMM 的实现。

2）通过完成 Java 的串行代码，输出比对空谱联合核的计算和图像稀疏表示的分类各自的执行时间，发现前者复杂度和计算量最高，于是将并行优化重心集中在了空谱联合核的计算上面。通过研究发现，Spark 并行优化主要是将原来计算量大的数据分发到不同结点上进行处理，合理利用 mapreduce 方法，最后可以在 driver 端得到最终的实验结果。其中需要对已有的数据进行合理的组织和分片，这样子能保证分发到 worker node 上的结点可以并行工作，最后通过 driver 的 collect 聚集操作进行汇总。

通过分析我们可以得到，核矩阵的每个元素都对应着一个像素，每个像素都会有窗口将其框出，理论上每个元素都会用到整幅高光谱图像的所有信息，所以我们需要将原来的高光谱图像 SSKSRCNF_img 分发到每个节点上，然后对应于我们要训练的总的数据样本，我们将其划分为 4 块，每块在各自的节点计算出核矩阵的部分，最终通过 collect 指令聚合到 driver 端形成一个最终的大的核矩阵，在此之后我们才能进行图像分类算法的编写，因为它需要用到所有整个训练数据和测试数据构成的核矩阵。该实验操作理论上可以达到 4 倍左右的加速比。通过并行实验的数据，为了能够加快算法运行速度，我们选取了图像中的一小部分作为我们的训练和测试数据，其中包括 540 个高光谱图像像素作为我们的训练数据，450 个高光谱图像像素作为我们的测试数据。在单机运行的场景下，核函数计算时间 ker_lwm time 为 10857.459s，在并行度为 4 的场景下，核函数计算时间为 4708.894s，整体运行速度有了较大的提升。但是没有达到理想的并行度要求，我们认为是多个 worker nodes 在 collect 的过程中，需要花费的交流时间拖慢了实际的并行速度。

3.成果的科学意义和应用前景（对基础研究，着重阐明其科学意义；对应用基础研究和应用研究，着重阐明其应用前景）。

我们成功实现了 Spark 平台上的高光谱遥感图像的分类，这个项目成功地实现了空谱联合的核稀疏表示的高光谱图像分类技术（Spatial-Spectral Kernel Sparse Representation for Hyperspectral Image Classification on Spark）的并行优化，可以为后来的核函数并行优化提供宝贵参考意见。

同时该项目的并行优化能够提高高光谱大数据的分类处理速度，可以促进将来的地物图像识别领域的发展，同时可以在矿产资源高光谱遥感探测领域进行应用。

通过 SpringMVC 的软件工程知识，设计基于 Spark 云计算平台的高光谱分类并行算法与系统，能够实现交互式的高光谱数据的地物识别功能，促进了该领域的软件设计发展。

4. 研究目标的达成度分析（含存在的问题与不足）。

在文献查阅方面，查阅高光谱遥感探测技术原理的各方面知识，广泛地阅读了《高光谱图像处理》、《数字图像处理》等书籍。了解高光谱遥感图像空谱联合分类的基本方法，从康泽昆学长的论文中了解了我们科研的大体步骤，搭建起了研究的整体思路框架。在图像处理方面，结合之前的基础知识，重点研究了基于空间相关性稀疏标识的高光谱分类算法，结合高光谱图像非负矩阵分解解混与稀疏表示分类梳理了图像处理的一系列算法。在并行优化方面，研读了王启聪学长的高光谱图像分类的 GPU 并行优化研究和高光谱遥感图像解混的分布式并行优化研究，加深图像分布式优化的内部细节理解。

对高光谱图像的格式理解方面，成功地了解到了高光谱图像的三种存储方式，包括 bil，bip，bsq，同时使用了 ENVI 等软件进行学习，在 Java 中基于 HDFS 编写了图像读取和保存的相关模块，熟练掌握了图像的存储格式。

在研究空谱联合的高光谱低秩表示分类方法，并基于云平台设计高光谱图像分类并行算法方面，组长蒋旭钊通过与刘倩学姐探讨分类算法的具体细节，与郑鹏、于坤学长探讨分布式并行优化的细节，成功地独立完成了 Java 串行代码以及基于 scala 的并行分类算法的编写，并通过配置虚拟机进行了分布式集群的搭建工作。

在基于 Spark 云计算平台设计和开发高光谱遥感图像分类系统方面，蒋旭钊学习了当前流行的 SpringMVC 网站设计框架，并通过设计前端页面，将后端的高光谱图像分类算法进行整合，设计出了独立完整的网站。其中也涉及到了一些调用 Hadoop API 和 Spark API 的知识，相关的配置文件他也进行了比较深入地研究。

问题与不足：

在项目的开展过程中，由于大家的能力和努力方向都不一致，李硕和王俊玺同学需要花较多的时间在自己的课业和生活上，无法投入太多的时间在科研训练中。因此蒋旭钊需要承担较多的科研工作，包括高光谱图像格式的读入和研究，以 Java 语言编写的空谱联合的核稀疏表示分类算法及其在 Spark 上的并行算法的优化策略，同时包括以空谱联合的核稀疏表示分类算法为后台的 SpringMVC 前后端系统设计。最终导致项目开展的速度并没有特别高效。

（√）实现预期功能并可演示的实物作品（含数字化作品）；

（）公开发表的学术论文或录用证明；

（）专利申请受理通知书；

（）软件著作权登记证书。

实物作品的概括性文字描述：

由组长蒋旭钊进行了协同开发，组织大家学习了 Github 的相关命令，通过代码托管工具提高开发效率，以蒋旭钊为主的代码编写迭代过程公开在了相关网站上：

1）以 Java 和 Scala 开发的 Spark 云计算平台上的空谱联合的核稀疏表示的高光谱图像分类并行优化发布在：

https://github.com/codeworm111/SSKSRCNFonSPARK

2）用 SpringMVC 整合了并行优化算法的系统发布在：

https://github.com/codeworm111/SpringMVC-of-SSKSRCNF

论文题目	项目成员是否为第一作者	稿件状态	刊物名称	国内刊号（CN 号）	国际刊号（ISSN 号）	发表时间

专利名称	项目成员是否为第一发明人	申请号	专利类型

软件著作权名称	项目成员是否为第一设计人	证书编号

指导教师意见：

1．审核个人研究总结、给出教师评定成绩。

姓名	角色	教师评语	教师评定成绩
蒋旭钊	负责人	该同学在本次科研训练中表现非常突出，作为组长主要负责项目整体设计与协调，全程参与了前期调研、关键技术研究、系统整体设计、核心模块开发、集成测试、文档总结等主要工作，每项工作均能按时高质量完成，体现了该同学勤恳的科研作风。	99
李硕	组员	该同学在本次科研训练项目中，主要负责部分模块开发工作，工作过程中勤恳严谨。	80
王俊玺	组员	该同学在本次科研训练中主要承担了分类算法设计、分类模块开发、软件调试等工作，工作认真负责，实践能力强。	89

2.审核项目结题报告，明确是否同意答辩。

指导教师对项目完成情况的总体评价： 项目组基于云计算技术研究高光谱遥感数据分类的分布式并行优化方法，在分析高光谱遥感图像分类原理与核稀疏表示分类算法的基础上，充分利用云计算平台的大数据存储能力和分布式计算资源，在 Spark 云平台上设计实现了高光谱遥感图像分类分布式并行处理算法和软件功能，在保证分类精度的情况下，有效提高了高光谱图像分类的处理能力和计算效率，实现了高光谱图像的快速有效分类。 该课题属于成熟技术在具体领域的实际工程应用，工作量较饱满。项目结题报告层次清晰，语句通顺，书写规范。课题组成员在参与项目过程中，主动积极，刻苦钻研，体现了实际解决科研问题的能力。 但是，软件交互界面的美观性上还有待优化。		
是否认为项目组已经完成预期成果并同意其答辩？	（√）是	（）否
指导教师签字：　　年　　月　　日		
项目单位审核意见： **单位盖章：　　年　　月　　日**		
学校审核意见： **学校盖章：　　年　　月　　日**		