|  |  |
| --- | --- |
| **分类号密级** |  |
| **UDC** |
|  |

**本科毕业论文（设计）**

基于Hadoop平台的用户查询分析

系统设计及实现

学生姓名 **尹硕** 学号 **16020031096**

指导教师 **刘洁**

院、系、中心 **信息科学与工程学院**

专业年级 **计算机科学与技术2016级**

论文答辩日期 **20xx年x 月x日**

**中国海洋大学**

基于Hadoop平台的用户查询分析系统设计及实现

完成日期：

指导教师签字：

答辩小组成员签字：

基于Hadoop平台的用户查询分析系统设计及实现

摘要

在如今的Web2.0时代，数据就是“黄金”，“得数据者得天下”，而网络用户的查询记录作为这些珍贵数据的一种，值得我们去分析和研究，去找到潜藏在其中的有价值的信息。大数据时代背景下，海量数据的存储和处理技术显得尤为重要，Hadoop就是应大数据存储和操作需求而产生的Apache开源软件框架，它能够进行不同主机的统一管理，实现数据的分布式存储和处理。

本文主要论述了通过使用Hadoop组件及相关技术，结合数据挖掘算法和传统的分析方法实现用户查询分析的过程。首先需要搭建分布式Hadoop集群，然后将待分析的数据文件导入HDFS，通过相关的jar包，使用eclipse集成开发环境远程连接Hadoop集群，并编写MapReduce Application程序，通过MR的辅助，来分别实现Apriori关联规则算法和DBSCAN聚类算法，从而挖掘出用户查询文本数据集的潜在信息。并且通过使用词频统计，基于经验和常识，以明确的分析需求为前提，来进行传统的数据分析。结合数据可视化技术，将上述过程的结果数据以图表的形式或者有向图、无向图的形式展现出来。

关键词: Hadoop;用户查询分析; Apriori算法;DBSCAN算法;数据分析;数据可视化。

Syntheticapertureradarimagechangedetectionbased on frequency-domain analysis and randommultigraphs

Abstract

With the development of earth observation programs, many multitemporalsynthetic aperture radar (SAR) images over the same geographical area are available. It is demanding to developautomaticchangedetectiontechniquestotakeadvantageoftheseimages.The change detection in synthetic aperture radar (SAR) images has some difficulty than optics, due to the fact that SAR images suffer from the presence of speckle noise. The change detection of different images is a process of analyzing image pairs. The same scene takes at different times to determine the changes that may occur between.Mostexisting techniquesdirectlyanalyzethedifferenceimage(DI),andtherefore,theyareeasilyaffectedby the speckle noise.

In this paper, a SAR image change detection method based on frequency domain analysis and random multigraphs is proposed. The method follows the analysis process from coarse to fine, and the logarithmic ratio operator is first used to generate DI in the coarsechangedregionslocalizationstage. Then the distinctive and salient regions is selected from DI by using frequency domain analysis. Therefore, the non-salient region is neglected, and the noisy unchanged regions caused by speckle noise is suppressed. Salientregions were selected as changing regional candidate areas and will be further categorized in the second stage. In the fine change regions classification stage, the representative sample pixels are selected from the changed region by FCM clustering, which are used as label samples. Image features are classified by input random multigraphs. By selecting a subset of neighborhood features to create graphics, this method can effectively utilize the nonlinear relationship between multi-time SAR images. Theexperimentalresultson two real SAR datasets and one simulated dataset have demonstrated the effectiveness of the proposed method.

**Keywords: random multigraphs; saliency map; synthetic aperture radar; change detection; frequency-domain analysis.**

目录

[第1章 绪论 8](#_Toc39667015)

[1.1 研究背景及意义 8](#_Toc39667016)

[1.2 本文研究内容 2](#_Toc39667017)

[1.3 论文结构 3](#_Toc39667018)

[第2章 相关知识和技术 4](#_Toc39667019)

[第章背景知识 4](#_Toc39667020)

[2.1 问题描述 4](#_Toc39667021)

[2.2 频率域显著性检测简介 4](#_Toc39667022)

[2.3 随机多图简介 5](#_Toc39667023)

[第3章 系统分析与设计 8](#_Toc39667024)

[第章方法框架介绍 9](#_Toc39667025)

[3.1 基于频域分析的粗变区域定位 10](#_Toc39667026)

[3.2 基于随机多图的精细变化区域分类 11](#_Toc39667027)

[第4章 系统实现 14](#_Toc39667028)

[第章实验结果 14](#_Toc39667029)

[4.1 数据集描述 14](#_Toc39667030)

[4.2 参数测试 17](#_Toc39667031)

[4.3 Bern数据集的实验结果 17](#_Toc39667032)

[4.4 旧金山数据集的实验结果 19](#_Toc39667033)

[4.5 模拟数据集的实验结果 20](#_Toc39667034)

[4.6 讨论 21](#_Toc39667035)

[第5章 系统运行和结果分析 22](#_Toc39667036)

[第章结论 22](#_Toc39667037)

[第6章 总结与展望 22](#_Toc39667038)

[参考文献 23](#_Toc39667039)

[致谢 24](#_Toc39667040)

**############### 憨憨目录 ######**

**目###录**

**摘要....................................................Ⅰ**

**Abstract................................................Ⅲ**

**1 绪论**

1.1 研究背景及意义..................................................1

1.2 本文研究内容....................................................3

1.3 论文结构........................................................4

**2**

2.1 国内外研究现状..................................................2

2.2 Hadoop生态系统.................................................10

2.2.1 Hadoop框架和组件简介.......................... 11

2.2.2 Hadoop的Java API和远程连接............................12

2.3 中文分词................................

2.3.1分词简介......................................

2.3.2 中文分词技术总览.............

2.3.3 庖丁分词.............

2.4 数据挖掘和数据分析.......................

2.4.1 数据分析总览...............

2.4.2 数据挖掘算法..................

2.4.3 Apriori算法...................

2.4.4 （any others?）..............

2.5 数据可视化...........

2.5.1 Python和可视化模块简介................

2.5.2 matplotlib模块.....................

**3**

3.1 需求分析.................................................

3.1.1 数据来源.......................................

3.1.2 数据集分析.......................................

3.1.3 系统的目标.......................................

3.2 数据处理............................................

3.2.1 数据存储.............................................

3.2.2 数据预处理.............................................

3.2.3 数据挖掘................................................

3.2.4 数据可视化分析......................................

3.3 系统整体架构............................................

**4**

4.1 系统环境........................................

4.1.1 操作系统..................................

4.1.2 分布式Hadoop集群搭建.......................................

4.1.3 eclipse与Hadoop集群的连接..............................

4.2 数据存储和预处理实现...................................

4.2.1 HDFS存储.................................

4.2.1 Hive表设计........................................

4.2.2 庖丁分词包.................................

4.3 数据挖掘实现..................................

4.3.1 Apriori数据挖掘算法...........................

4.3.2 分类分析...............................

4.3.3 变化和偏差分析..............................

4.5 数据可视化实现...........................

4.5.1 饼状图............................

4.5.2 柱状图.....................

4.5.3 折线图......................

4.5.4 词云...........................

**5**

5.1 URL主站排名.....................................

5.2 用户对词条的点击行为..............................

5.3 用户查询词的关联规则..............................

5.4 词频统计和分析.....................................

5.5 用户查询次数分析.........................................

**6**...................................................

.................................................

..............................................................66

# 第1章 绪论

本部分简要介绍了基于Hadoop的用户查询分析的研究背景和意义，并对本课题的研究做概要性的陈述，最后展示本文的结构。

## 1.1 研究背景及意义

随着硬件技术、软件技术的发展，各行各业各领域都变得与计算机应用密不可分，银行、电信、企业、传感器、社交网络每天都在产生各种形式的数据，或字符文本，或音频，或图像，或视频。数据不断地爆炸式增长，海量数据的产生使我们迎来了大数据时代，这既是计算机领域的一次飞跃，又是处在信息时代的整个当今世界一个具有里程碑意义的发展。

然而，数据量变的积累使其发生了质变，数据的存储、处理，数据的应用，数据的价值都发生了改变，并且数据不单单是数据本身了——某个大数据集背后一定隐藏着对人们有用的潜在信息，这些信息往往能够揭示某种客观规律，能够反映用户的心理诉求，能够用来进行预测，驱动决策，指导发展，指示前进方向。

传统数据存储和处理方法或技术不能满足大数据的存储和处理需求，各种与大数据相关的技术应运而生，例如分布式存储、云计算、NoSQL、数据仓库等等，本文涉及的Hadoop就是一个大数据存储和计算的框架。Hadoop是Apache下的开源软件，易扩展，支持大文件存储，能够快速应对节点故障的发生，能够兼容廉价硬件机器，并且拥有成熟的生态圈，诸多优点使它为国内外许多公司所使用。

同样地，大数据的产生也促成了数据分析方法的变革，数据库（或者数据仓库、分布式文件系统的文本文件）结合机器学习算法，产生了区别于传统数据分析方法的新方法——数据挖掘。通过编写、运行数据挖掘应用，数据拥有者能够“挖掘”出“表面数据背后的数据”，发现隐藏在数据中的规律或者信息，这种规律或信息的价值可能比人们想象的要大得多。

大数据时代下的数据分析的应用已经取得了很多的成功案例，农夫山泉通过对运营过程中产生的数据进行分析处理，解决生产和销售的平衡问题，提高物流的效率；沃尔玛对消费者购物行为的分析，得出“啤酒+尿布”的促销手段；Google成功预测冬季流感的传播；奥巴马大选成功连任，等等。而在我们身边的典型例子，就是在淘宝或者亚马逊APP里的精准的商品推荐。

对搜索引擎用户查询数据的分析，是大数据技术的重要应用之一，通过对大量查询日志记录的挖掘和分析，能够得到反映用户群体的信息，比如某段时间内，什么查询词的词频最高，即什么词是hot词，它与什么人物或者事件相关，借此可以判断该人物或者事件的社会影响力有多大；再比如通过聚类等数据挖掘算法，得到词与词之间的关联关系，从而为广告的推送提供依据；类似地，搜索引擎的使用会涉及电子商务领域，例如某品牌的衣服、鞋子或者其他商品在搜索引擎用户中很“火”，或者某两类物品在查询时大概率“出双入对”等，那么电商平台可以通过融合网络用户查询的分析技术，实现其推荐系统的功能扩充或者有效改进，而不再单单是使用销量、评价等平台内部数据进行分析和推荐。它可以为包括搜索引擎公司在内的很多企业提供数据层面的依据，很大程度上辅助预测，指导决策，从而带来更大的利润。

不仅如此，除了商业应用之外，网络用户的查询分析能够为社会科学的调查和研究提供数据支持，它能够得出反映大众心理的结论，这可能比专门的社会调查或者实验得出的结论更加快速、有效和准确。

## 1.2 本文研究内容

本文阐述在Hadoop平台上进行用户查询分析的过程，结合传统分析方法和数据挖掘算法，对搜索引擎的查询日志文本数据进行处理，得出分析的结果，并绘制图表进行展示。

本文主要研究的内容如下：

（1）学习和配置分布式Hadoop集群，研习其核心组件HDFS和MapReduce，以及其相关组件Hive数据仓库，分析它们的实现原理和使用方法，为后续的应用做准备，为算法的提出奠定基础。

（2）分析搜狗实验室提供的搜狗用户查询日志，对其中记录的数据项逐一分析和解释，并提出基于此数据集的数据分析的需求，提出用户查询分析系统的实现目标和数据处理大致方向，为算法的提出提供依据。

（3）研究数据挖掘的机器学习算法，对比监督式学习和无监督式学习，对比神经网络、决策树、贝叶斯分类、聚类和集成学习等等，研究和分析它们的原理、应用场景，最终确定使用Apriori关联规则算法和DBSCAN算法进行数据挖掘，着重介绍这两种算法，并提出它们在本用户查询分析系统的使用方法或者实现途径，以及它们与Hadoop平台的HDFS和MapReduce的结合策略。使用Apriori算法查找频繁项集，设定支持度和置信度，挖掘查询词的关联规则；使用DBSCAN算法，对查询词基于密度进行聚类，把查询词的簇进行输出可视化。

（4）建立Hive数据仓库，使用外部表关联HDFS的数据文件，这样的数据表不生成或者保留数据副本，并且隐藏了MapReduce的实现细节，可以用于进行以词频查询为主要手段的传统分析方法，基于事先确定好的分析需求，对特定类别的词进行查询，对比数量关系或比例关系，进行输出可视化，然后揭示其意义。

## 1.3 论文结构

第一章绪论。首先介绍了本文的背景，在大数据时代下海量数据的存储和处理出现了区别于传统方法的新的技术，Hadoop就是其中一种；数据挖掘伴随大数据的发展而产生，是大数据结合某些机器学习算法而实现的数据分析应用。然后介绍了本文的研究意义，基于Hadoop的用户查询分析可以为广告推送、电商推荐系统改进、社会调查和大众心理的研究提供依据。最后介绍了本文的研究内容和本文的结构。

第二章相关知识和技术。

第三章系统分析与设计。

第四章系统实现。

第五章系统运行和结果分析。

第六章总结与展望。

第六章总结与展望。

# 第2章 相关知识和技术

###### waiting for adding or filling

# 第章背景知识

本部分简要介绍了SAR图像变化检测的问题, 并给出了该方法的动机。然后, 详细介绍了频域分析和随机多图。

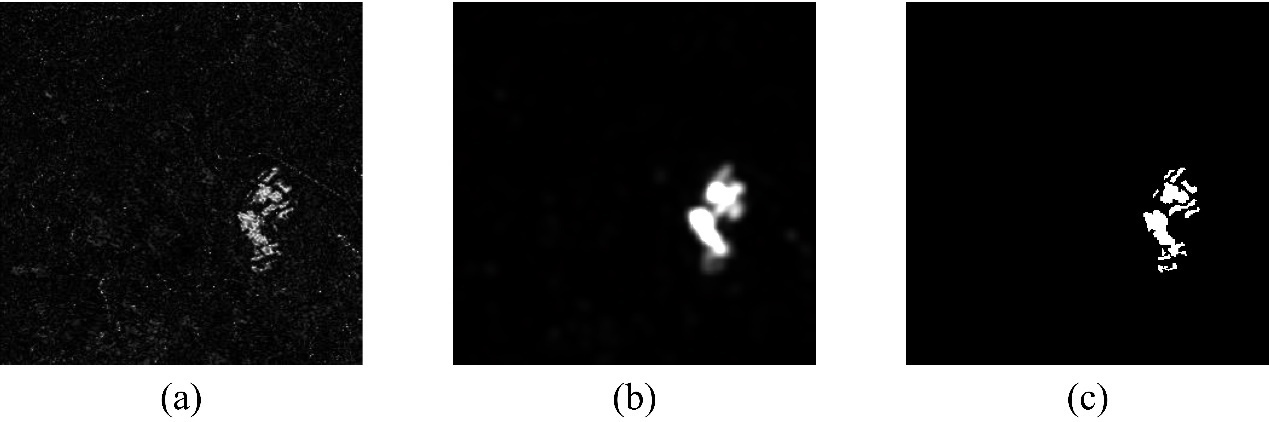
## 2.1 问题描述

考虑两个相关的多时间SAR图像和, 这两个图像都被散斑噪声污染。变化检测的目的是生成变化图, 该图表示在和,之间发生的变化。从另一个角度来看, SAR图像变化检测问题可以看作是一个二分类过程。分类后, 得到一个二值图像, 其中不变的像素标记为 0, 变化的像素标记为1。

如前所述, 变化检测的过程包括粗糙变化区域定位和精细变化区域分类两个阶段。在第一阶段, 使用频域显著性检测来抑制DI中的未变化区域的噪声。在第二阶段, 随机多图算法为变化和不变的像素分类提供了一个很好的解决方案。

## 2.2 频率域显著性检测简介

显著性检测意味着查找具有强烈局部或全局对比度的区域。如图 2.1(a) 所示, 变化的区域是非常独特的明亮区域。与图像的其他部分相比, 这些区域吸引了人类视觉系统更多的关注。图 2.1(b) 显示了频域分析生成的显著性图, 图 2.1(c) 显示了地面真实变化图。可以看出，显著性区域的形状和位置与变化的区域相似。因此, 频域显著性检测适用于粗变区域定位。



(a)对数比运算符获得的DI；(b) 频域分析生成的显著性图；(c) 通过人工标记制作的地面真实变化图

**图2.1显著性图与地面真相变化图的相似性**

频域显著性检测采用具有适当尺度的低通高斯核进行振幅谱卷积, 得到显著性图。具体而言, 给出一个图像，首先通过傅立叶变换将其转化为频域：



通过计算振幅谱和相位谱。采用高斯核抑制图像振幅谱中的峰值，如下所示：



经过上面计算，可以得到的平滑振幅谱。然后，将和原始相位谱一起计算逆变换，进而产生显著性图：



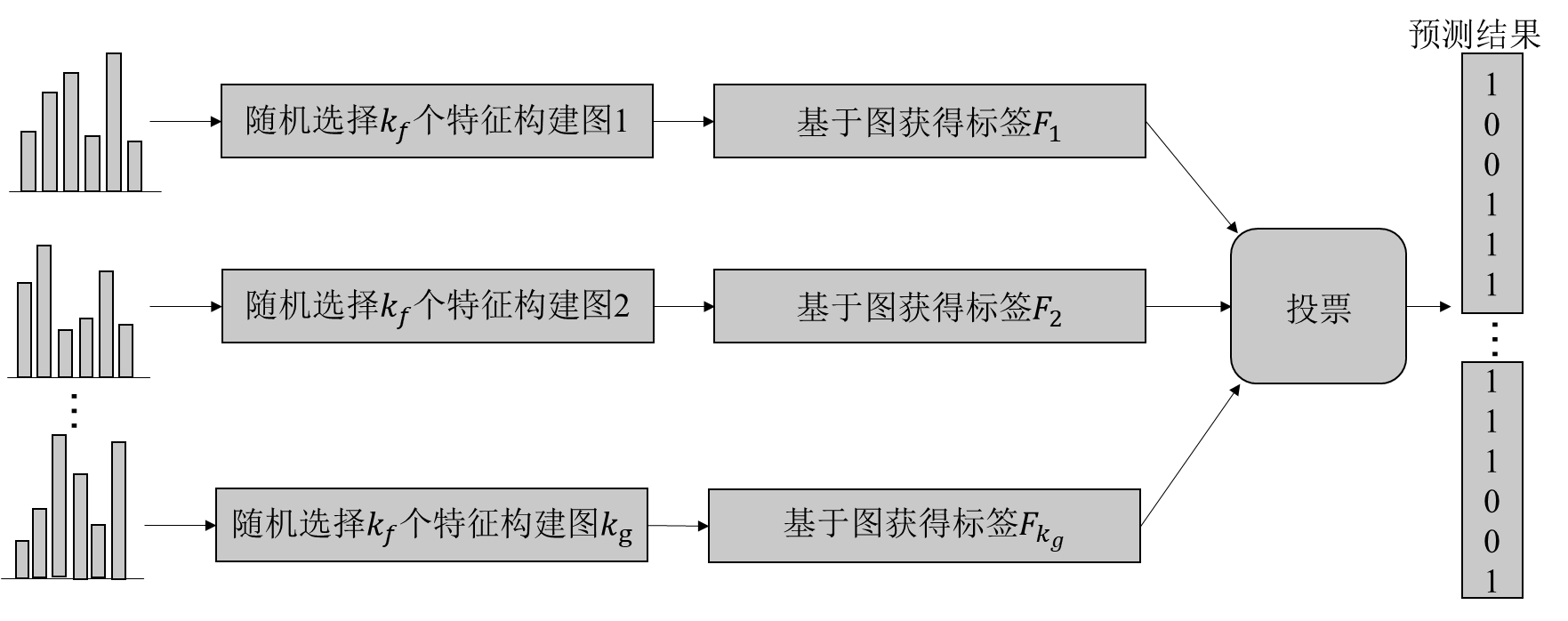
## 2.3 随机多图简介

现在有许多种为解决遥感图像应用中的分类问题而提出的方法。但在SAR图像处理中很少使用基于图形的分类方法。本文在精细变化区域分类阶段, 使用随机多图将变化区域候选项中的像素分类为变化类和不变类。它可以利用多时间SAR图像之间的非线性关系, 通过随机选择一个特征子集来创建图形。

随机多图方法使用无向图对输入数据和数据之间的关系进行建模。给出了一个数据集,是特征空间的维度。是带有标记的数据集。是未标记的集合, *u*是未标记数据的数量。

图2.2说明了随机多图算法的流程图，整个随机多图的结构描述如下：

* 步骤 1: 从每个样本的所有*d*维特征中随机选择个特征。
* 步骤 2: 选择*m* 个锚点以覆盖由锚点矩阵**A**表示的数据流形, 然后计算映射矩阵**P**, 通过选定的锚点表示其余的数据点。
* 步骤 3: 使用图形拉普拉斯正则化在此图上运行推理。
* 步骤 4: 转到步骤 1, 直到获得*kg*个图。
* 步骤 5: 选择获得的*kg*个图以获取未标记数据点的标记。



**图2.2 随机多图方法的流程图**

在本文中，使用锚图方法10来构造图。

其中每个都是一个锚点。因此, 标记预测函数可以被表示为：



其中是数据自适应权重。定义两个向量和。上述公式可以重写为：



因此,未知标签的解空间从较大的空间**f**减少到较小的空间**f**a。本文使用*k*均值聚类集中心作为锚点。因为这些聚类中心具有强大的表示能力, 可以充分覆盖整个数据集。

关于矩阵P, 本文使用局部锚点嵌入来重建任何数据点作为其最近锚点的凸组合。因此, 矩阵P可以通过以下方式确定:

其中*X*表示数据矩阵, 其中每一行都是数据样本。*A*∈Ｒ(*l+u*)*\*d*是每一行都是锚点的锚矩阵。P ∈Ｒ(*l*+*u*)\**m*是数据锚点映射矩阵。利用矩阵P, 该图可以用它的邻接矩阵来表示：

对角矩阵被定义为

在获得映射矩阵 P 后, 通过一个半监督的学习框架,未标记数据的标签可以很容易地预测。

# 第3章 系统分析与设计

本章介绍系统分析与设计，包括需求分析、数据处理和系统整体架构，参考软件生命周期逐步进行，为进一步的系统实现做准备。

## 3.1 需求分析

3.1.1 数据来源

基于Hadoop平台的用户查询分析系统，使用的数据集来自搜狗实验室（https://www.sogou.com/labs），是其语料数据栏目中的“用户查询日志”（SogouQ）。搜狗实验室提供了迷你版（mini）、精简版（reduced）、完整版三个版本的数据，都是以文本文件的形式存储，文件格式分别为.sample、.reduced、.filter，记录Sogou搜索引擎用户的查询情况，时间都是在2006年8月份，查询记录更具体的时间跟数据的版本有关。

其中，迷你版的数据集是某天大约开始的10分钟内的用户查询记录，总共10000条记录；精简版是某天从开始到结束的24小时内的记录，共170多万条；而完整版的数据则是2006年8月份共30天的查询记录，文件名从access\_log.20060801.decode.filter到access\_log.20060831.decode.filter，每天的记录组成一个数据文件，其中缺少8月25日的记录，每个文件一行一条记录，行数从几十万到一百多万不等，这样30个文件的记录数能以千万来计量，数据量总共是1.56GB。

本文涉及的用户查询分析系统所使用的数据集，无特殊说明都是完整版的Sogou用户查询日志数据。

3.1.2 数据集分析

根据Sogou实验室的数据介绍和格式说明，每条记录的数据项或者记录属性从行首到行尾依次是用户ID、查询词、URL在返回结果的排名、用户点击的顺序号、用户点击的URL，总共5项。为叙述方便，“查询词”在本文中用“查询字符串”来代替，强调它是用户输入的一个字符串，与作为它子串的“单词”区别开。各字段的具体描述见下表：

|  |  |  |
| --- | --- | --- |
| 字段名 | 代码中的变量名 | 字段的含义 |
| 用户ID | user\_id | 用户使用搜索引擎时自动赋值到Cookie的信息，同一次使用该浏览器的几次不同查询记录对应同一个ID |
| 查询字符串 | query | Sogou搜索引擎的用户在查询时向浏览器查询框中输入的查询文本字符串 |
| URL在返回结果的排名 | url\_seq | 用户提交查询后，搜索引擎返回给用户一个查询结果的页面，每个网页按照顺序从上到下排列，该字段就是用户选择的页面的从上到下的排名 |
| 用户点击的顺序号 | click\_seq | 用户点击了某个页面链接，该字段记录这次点击在这次查询中是第几次点击 |
| 用户点击的URL | url | 这次查询中，用户点击的某个网页的统一资源定位符URL |

从该数据集的字段来看，用户查询的分析方向，可以是对查询内容分析，也可以是对查询行为分析。对于查询内容，主要是对用户查询字符串的分析，使用数据挖掘算法或者传统的文本分析方法，提取单词或关键词进行关联规则或聚类分析，统计词频进行对比分析，也可以对URL来统计词频对比不同主站的访问量等等；对于查询行为，在本数据集上就是对url\_seq和click\_seq进行分析，在用户的具体查询字符串作为输入，结合Sogou搜索引擎进行搜索之后，得出查询结果，那么分析点击行为就可以估计该查询结果跟用户的真实想法之间的误差或“距离”。

3.1.3 实现目标

本系统基于Hadoop，实现的目标如下：

① 对用户的查询字符串进行分词，然后进行计数得到每个查询单词的词频，通过预处理过滤掉像标点符号这样的噪声词；

② 实现Apriori关联规则算法，根据预先设定的支持度和置信度，基于高词频的单词组成的单个词的频繁项集，分别得到双词、三个词、四个词的频繁项集，继而得到单词与单词之间的关联规则；

③ 实现DBSCAN聚类算法，预处理后得到适合聚类的单词作为样本点，设定半径、半径内最小样本数等参数，以一种非度量距离作为相似度的度量方式，进行密度聚类，得到簇作为聚类结果；

④ 基于

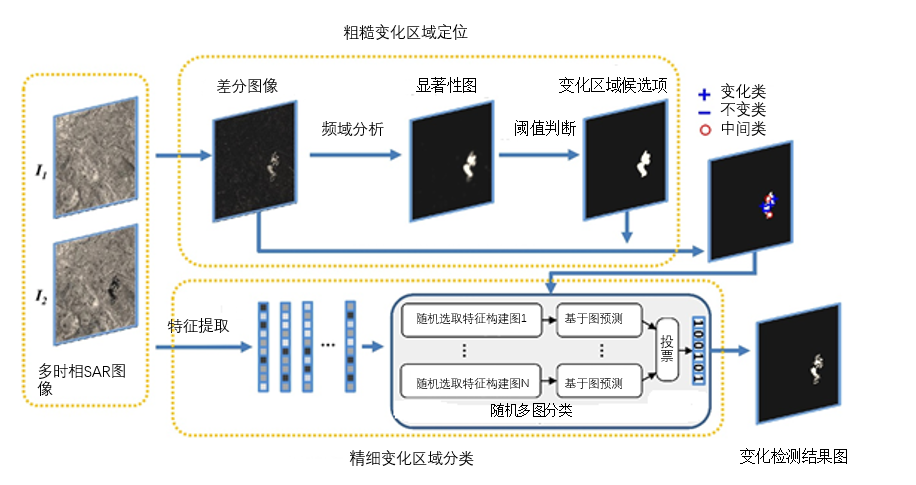
## 3.2 数据处理

## 3.3 系统整体架构

# 

# 第章方法框架介绍

本章将详细介绍本方法的实现过程。图 3.1显示了该方法的总体框架。具体而言, 所提出的变化检测方法由两个步骤组成。



**图3.1基于频域分析和随机多图的SAR图像变化检测方法流程图**

第一步：利用频域分析对粗糙变化区域进行定位。对数比率运算符首先用于生成DI。然后利用频域分析得到有显著性的区域。这些区域被选为变化的区域候选区域, 并将在第二阶段进一步分类。

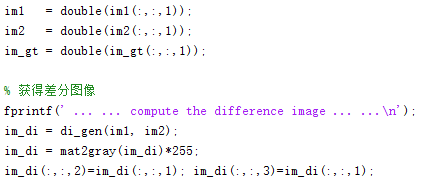
第二步：基于随机多图对精细变化区域进行分类。通过FCM聚类从变化的区域中选择代表性样本像素。这些像素作为标签样本。图像块特征是在每个像素周围生成的, 这些特征被输入随机多图进行分类。从分类结果中,可以得到变化和不变的像素, 从而得到最终的变化图。

## 3.1 基于频域分析的粗变区域定位

给出两个多时相SAR 图像,和。对数比率运算符用于生成差分图像*ID*。它是由*I1*和*I2*按照以下公式逐个像素计算出来的：

*ID*=|log(*I1*) – log(*I2*)|

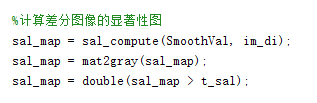
首先要从读入的两个SAR图像中获得差分图像，如图3.2。



**图3.2获得差分图像**

通过视觉感知，*ID*中变化的区域被认为是具有显著性的。因此, 本文使用显著性检测方法来定位变化的区域候选项,这里也采用了频域分析进行定位。

使用频域分析计算 DI 的显著性图，如图3.3。



**图3.3获得差分图像的显著性图**

它采用低通高斯核进行振幅谱卷积, 得到了变化图。在实现中, 输入的DI大小调整为256x256 个像素, 用于傅立叶变换。



**图3.4将输入的DI大小调整为256x256 个像素**

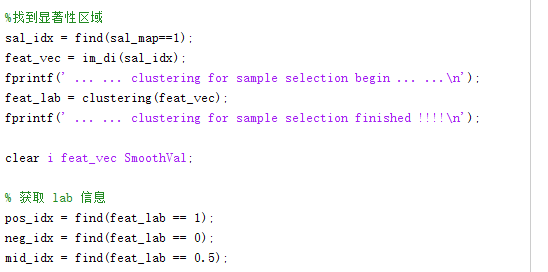
此外, 经验证, 当高斯核尺寸设置为0.05\*n (此处, n 是输入图像的宽度), 显著性检测器会获得最佳性能。故而将内核大小设置为0.05\*n。

在获得显著性图后, 利用二值化方法生成变化区域的候选项。具体来说,值大于给定阈值t的像素将作为变化区域的候选项。在阈值操作完成后, 将提取出具有显著信息的感兴趣区域。这些区域将由随机多图进行进一步分类。

## 3.2 基于随机多图的精细变化区域分类

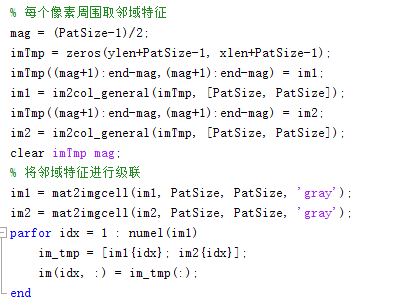
对于精细变化区域分类, 本文首先对DI中的变化区域候选对象执行FCM算法, 将像素划分为三个集群: 变化类Ωc、不变类Ωu和中间类Ωi。属于Ωc和Ωu的像素很有可能发生变化和保持不变。具体而言, Ωc和Ωu具有较高的类内相似性和较低的类间相似性, 它们可以被看作是纯粹的变化和不变的类。因此, Ωc和Ωu可以被视为随机多图的可靠标签样本。

使用FCM将DI 中的变化区域候选对象分为三类，如图3.5。



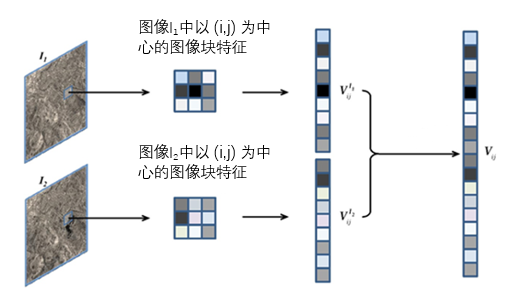
**图3.5使用FCM进行分类**

在变化的区域候选项 (属于Ωu、Ωc和Ωi) 中的所有像素邻域特征的生成如图 3.2所示。值得注意的是,由于像素的性质与其邻域高度相关,因此提取多时相SAR图像中的对应像素的周围像素并将其组合在一起作为像素的特征向量。具体来说, 在图像I1中提取以(i,j)为中心的图像块特征，也提取了在图像I2中相应的图像块特征。这两个图像块特征都转换为矢量和。这两个向量级联在一起, 形成一个新的向量。如图3.6。



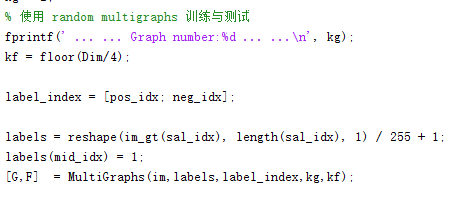
**图3.6获取每个像素的**

这个过程在图 3.7得到说明, 而表示位置(i,j)处的特征向量。



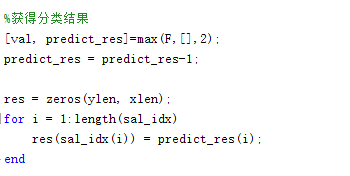
**图3.7作为随机多图输入的每个位置的邻域特征生成**

因此, 得到了变化区域候选项的像素的邻域特征。这些特征被输入到随机多图中进行分类。属于Ωc和Ωu的像素在分类中被视为标记样本。通过随机多图中的半监督推断, 可以将未标记的样本分为变化的类或不变的类。



**图3.8使用随机多图进行分类**

从分类结果中, 可以得到最终的变化检测结果。



**图3.9获得分类结果**

# 

# 第4章 系统实现

## 4.1 系统环境

## 4.2 数据存储和预处理实现

## 4.3 数据挖掘实现

## 4.4 数据可视化实现

# 

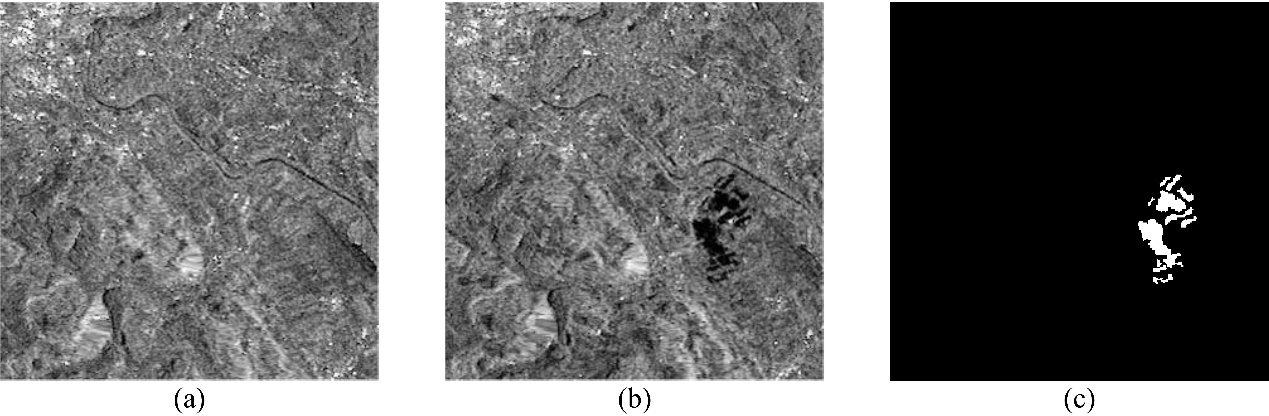
# 第章实验结果

## 4.1 数据集描述

实验中使用的第一个数据集是Bern 数据集。如图 4.1所示, 它是ERS-2 卫星拍摄的两张合成孔径雷达图像。这些图像分别于1999年4月和1999年5月拍摄。在这两个日期之间的这段时间里, 洪水淹没了Bern的部分地区。如图4.1(c) 所示, 可用的地面真值图是利用先前的知识和图像解释结合人工生成的。

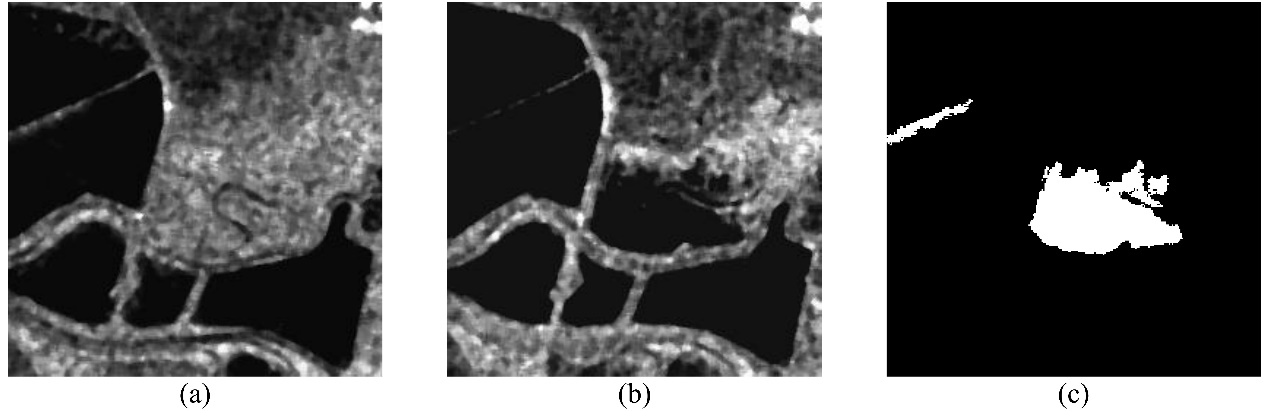
实验中使用的第二个数据集是旧金山数据集, 如图 4.2所示。它表示由ERS-2 SAR传感器在旧金山中获得的两个SAR图像的一个部分 (256x256 像素)。该数据集包含分别于2003年8月和2004年5月获得的两张图像。原始图像为7749x7713 像素, 可在Earthnet11中获得。现有的地面真值图是通过整合先前的知识和照片解释而产生的，如图 4.2(c)所示。

如图 4.3所示, 第三个数据集是模拟数据集。数据集的大小为256x256 像素。地面真相如图 4.3(c)所示。模拟数据集部分来自中国北京的全色IKONOS图像。Nakagami分布的散斑图案是由选定的参数合成的,并将这些图案作为噪声添加到图像中。图 4.3(b)中的变化区域与图 4.3(a)中的相应区域相比, 振幅幅度提高了40%。



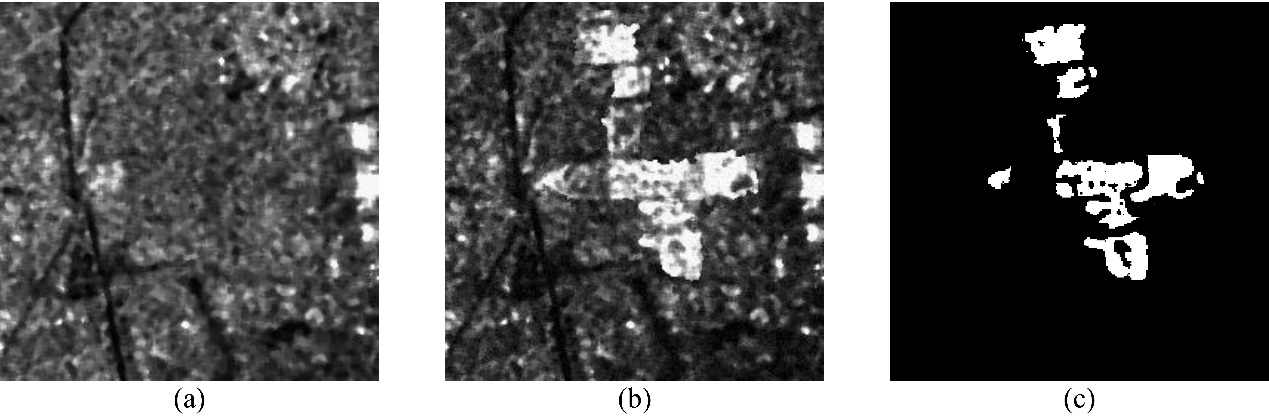
(a) 1999年4月获得的图像, (b) 1999年5月获得的图像;(c) 通过人工标记制作的地面真值图像

**图4.1 Bern数据集**



(a) 2003年8月获得的图像; (b) 2004年5月获得的图像; (c) 真值图像

**图4.2旧金山数据集**



(a) 时相一图像； (b) 时相二图像；(c) 真值图像

**图4.3模拟数据集**

变化检测结果以二进制图像的形式显示, 而白色像素表示变化的像素, 黑色像素表示不变的像素。对变化检测结果进行了定量分析, 以评价该方法的性能。首先计算错误的归为变化类(FP)和错误的归为不变类(FN)的像素数。FP是在地面真实图像中不是不变的类, 但错误地归类为已变化的类的像素数。FN是在地面真实图像中是变化类，但错误地归类为未变化类的像素数。接下来,本文计算总误差 (OE) 和进行了正确分类的百分比(PCC)。是由OE = FN + FP 计算的。PCC的计算方法如下：

PCC = (TP + TN)/(TP+TN+FP+FN)

其中TN表示正确的归为不变类的像素数, 即正确分类为未变化的像素数, TP表示正确的归为变化类的像素数, 即正确归类为变化类的像素数。最后, 计算 kappa 系数12, 给出正确分类的像素数, 并根据纯粹偶然预期的正确分类的像素数进行更正。具体来说, kappa 系数 (KC) 是由PCC和误差比例的减少 (PRE) 计算出的，公式如下：

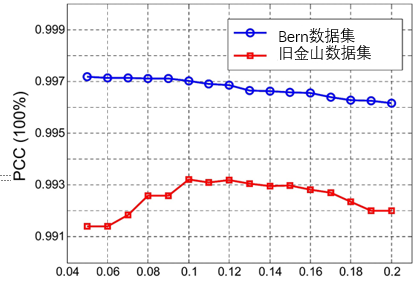
KC =

其中

PRE = +

将该方法与三个SAR数据集的密切相关方法进行了比较。用于比较的方法为主成分分析和k均值聚类(PCAKM)6,马尔可夫随机场和模糊C均值聚类 (MRFFCM)13，Gabor两级聚类 (GaborTLC)5, 和极端学习机 (NR-ELM)3。

## 4.2 参数测试



**参数*t***

**图4.4两个真实SAR数据集上 PCC 与参数*t*之间的关系**

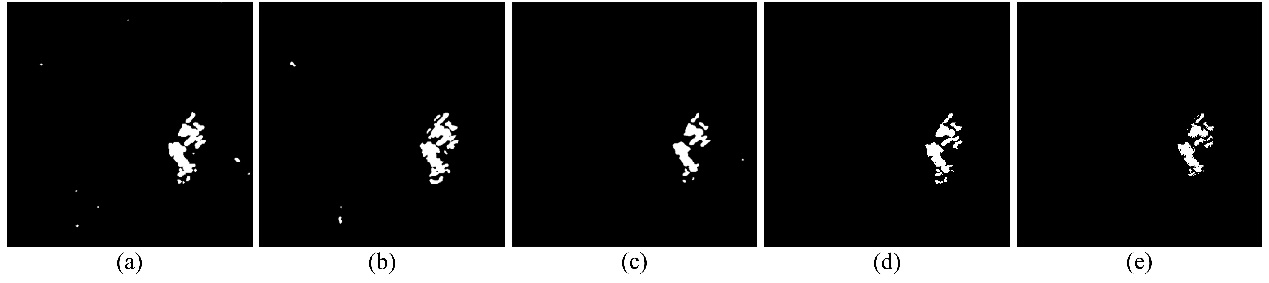
第一个实验是对参数t的测试, 该参数代表了频域分析的阈值。它是影响最终变化检测结果的重要参数。

以Bern数据集上的结果为例，当 t 范围从0.04 到0.10时, PCC值几乎不变。当 t 大于0.11 时, PCC值略有下降。在旧金山数据集上, 当 t 从0.04 到0.11时, PCC 的值就会增加。这是因为, 当 t 相对较小时, 会保留更多的噪声区域。这些噪声区域会影响变化检测的性能。但是, 当 t 大于0.12 时, PCC 的值就会降低。当 t 值增大时, 提取的显著性区域的大小将非常小, 许多实际变化的区域将被忽略。

从两个数据集的结果中, 可以观察到, 当 t 的范围从0.10 到0.12时, PCC的值是稳定的, 是令人满意的。因此, 在本文的实现中, 将 t 的值设置为0.11。

## 4.3 Bern数据集的实验结果

实验结果以两种方式表现为: 图形形式的最终变化图和表格形式的定量分析。图 4.5显示了Bern数据集上各种方法的最终变化映射, 表1列出了用于评估的值。



(a) PCAKM6的结果 (b) MRFFCM13的结果 (c) GaborTLC5的结果 (c) NR-ELM3的结果 (e) 本文方法的结果

**图4.5Bern数据集上各种变化检测方法的可视化结果**

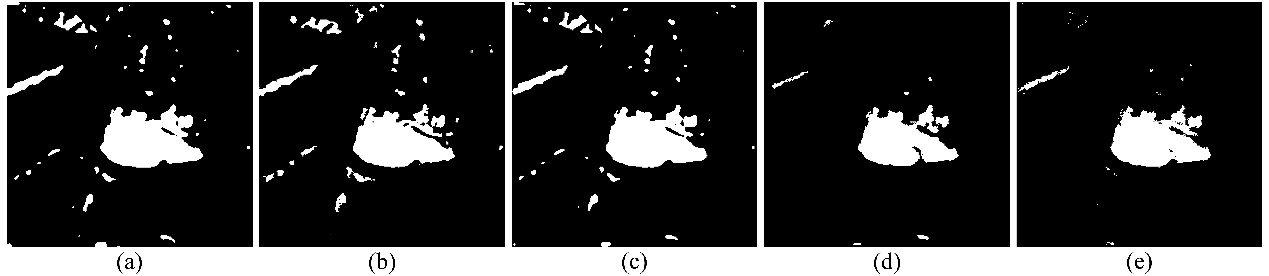
**表4.1 Bern数据集上不同方法的检测结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **方法** | **FP** | **FN** | **OE** | **PCC (%)** | **KC (%)** |
| PCAKM6 | 247 | 119 | 366 | 99.60 | 84.78 |
| MRFFCM13 | 364 | 47 | 411 | 99.55 | 84.13 |
| GaborTLC5 | 135 | 173 | 308 | 99.66 | 86.27 |
| NR-ELM3 | 147 | 146 | 293 | 99.68 | 87.16 |
| 本文方法 | 55 | 216 | 271 | 99.70 | 87.24 |

如图 4.5(a) 和4.5(b) 所示, PCAKM和MRFFCM错误地将一些不变的像素分类为变化的类, 从而生成一些噪声区域。因此, PCAKM和MRFFCM的FP值相对较高, 如表4.1所示。

如前所述, 该方法利用频域分析生成了DI的显著性图，然后, 许多嘈杂不变的区域被抑制。因此, 该方法在抑制背景噪声方面效果良好, 并且达到了最小FP值。此外, 该方法还实现了最大PCC和KC值。值得注意的是, PCC和KC值是变化检测结果分析中的最具有说服力的参数。因此, 本文可以得出结论, 该方法优于Bern数据集上的其他方法。

## 4.4 旧金山数据集的实验结果



(a) PCAKM6的结果 (b) MRFFCM13的结果 (c) GaborTLC5的结果 (c) NR-ELM3的结果和 (e) 本文方法的结果。

**图4.6旧金山数据集上各种变化检测方法的可视化结果**

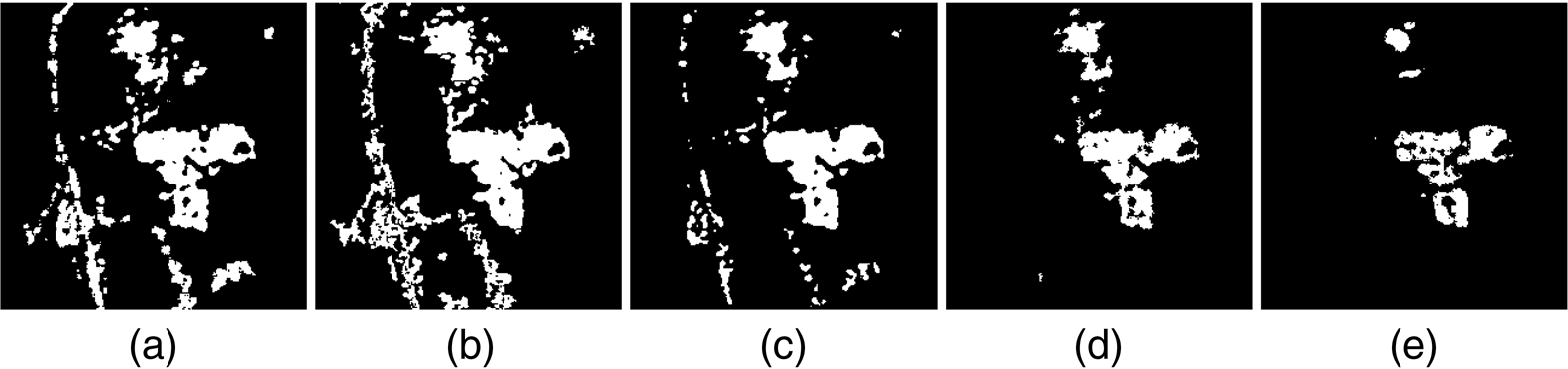
**表4.2 变化旧金山数据集上不同方法的检测结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **方法** | **FP** | **FN** | **OE** | **PCC (%)** | **KC (%)** |
| PCAKM6 | 1618 | 25 | 1643 | 97.49 | 83.68 |
| MRFFCM13 | 1511 | 191 | 1702 | 97.40 | 82.69 |
| GaborTLC5 | 1376 | 60 | 1436 | 97.81 | 85.39 |
| NR-ELM3 | 182 | 596 | 778 | 98.81 | 90.68 |
| 本文的方法 | 189 | 255 | 444 | 99.32 | 94.86 |

图4.6显示了旧金山数据集上不同方法的变化检测结果。表4.2显示了由不同方法生成的变更图上的相应定量分析。通过对这些变化图的视觉比较, 可以观察到PCAKM, MRFFCM, 和 GaborTLC产生了许多噪声区域, 在这些区域中, 许多不变的像素被错误地归类为变化的类。因此, 这些方法的FP值相对较高。高FP值会影响这些方法的整体性能。与上述方法相比, NR-ELM和本文的方法获得了类似的较低的FP值, 如表2所示。然而, 该方法的FN值远远低于NR-ELM。因此, 该方法实现了最佳的PCC和 KC值。该数据集的结果还表明, 该方法可以抑制噪声不变区域。视觉和定量分析都证明了该方法对数据集的有效性。

## 4.5 模拟数据集的实验结果

至于模拟数据集, 结果如图 4.7所示, 并在表3中列出。数据集上的散斑噪声的影响相当强烈。从图 4.7, 可以观察到, 在使用 PCAKM, MRFFCM, GaborTLC 和NR-ELM算法时，许多不变像素被错误地归类为变化的类。因此, 表4.3中这些方法的FP值相对较高。然而, 本文的方法能有效抑制散斑噪声, 达到最低的FP值。此外, KC值明显高于其他方法, 这意味着该方法在强散斑噪声情况下是有效的。



(a) PCAKM6的结果 (b) MRFFCM13的结果 (c) GaborTLC5的结果 (c) NR-ELM3的结果 (e) 本文方法的结果

**图4.7模拟数据集上各种变化检测方法的可视化结果**

**表4.3变化模拟数据集上不同方法的检测结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方法 | FP | FN | OE | PCC (%) | KC (%) |
| PCAKM6 | 5588 | 544 | 6132 | 90.94 | 50.81 |
| MRFFCM13 | 6883 | 232 | 7115 | 89.14 | 48.67 |
| GaborTLC5 | 2866 | 570 | 3436 | 94.76 | 65.91 |
| NR-ELM3 | 1090 | 897 | 1987 | 96.97 | 87.40 |
| 本文的方法 | 154 | 1364 | 1518 | 97.68 | 89.26 |

## 4.6 讨论

该方法是检测多时间合成孔径雷达图像变化区域的一种很有前途的工具。如表4.1-4.3 所示, 与四种密切相关的方法相比, 该方法具有更好的性能。用于比较的方法有PCAKM, MRFFCM, GaborTLC, 和 NR-ELM.。

从表4.1-4.3 中列出的定量分析中, 该方法实现了最小的OE值和最大KC值。散斑噪声产生的 FP是由DI中的非显著区域的噪声产生的, 这些区域在频域分析过程中被忽略。因此, 该方法在三个数据集上实现了FP值的最佳性能。此外, 随机多图随机选择邻域特征的子集来创建图形;它可以利用多时间SAR图像之间的非线性关系, 从而在一定程度上抑制散斑噪声。很明显, 该方法在保持细节的同时, 有利于抑制不变区域中的噪声。

# 第5章 系统运行和结果分析

## 5.1 URL主站排名

## 5.2 用户对词条的点击行为

## 5.3 用户查询词的关联规则

## 5.4 词频统计和分析

## 5.5 用户查询次数分析

5.1 URL主站排名.....................................

5.2 用户对词条的点击行为..............................

5.3 用户查询词的关联规则..............................

5.4 词频统计和分析.....................................

5.5 用户查询次数分析.........................................

# 第章结论

本文提出了一种基于频域分析和随机多图的SAR图像变化检测方法。该方法遵循从粗糙到精细的过程。在粗糙变化区域定位阶段，首先使用对数比率运算符生成DI。然后利用频域分析得到变化的区域候选项。在精细变化区域分类阶段, 使用FCM从变化的区域候选项中选择具有代表性的样本。然后, 在每个像素周围生成图像块特征, 并将这些特征输入随机多图进行分类。从分类结果中, 可以得到最终的变化图。通过视觉和定量分析的实验结果, 验证了该方法的有效性。

# 第6章 总结与展望

########### waiting to be filled...

# 参考文献

########### waiting to be filled

1. B. Hou 等人, “基于高斯-日志比图像融合和压缩投影的 SAR 图像无监督变化检测” IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7(8),3297–3317 (2014).
2. J. Lu 等人, “利用堆叠限制玻尔兹曼机器进行SAR图像变化检测的差异学习” Soft Comput. 20(12), 4645–4657 (2016).
3. F. Gao等人, “基于近邻的比率和极端学习机器的合成孔径雷达图像的变化检测” J. Appl. Remote Sens. 10(4), 046019(2016).
4. Y. Zheng 等人, “组合差分图像与K均值聚类技术在 SAR 图像变化检测中的应用” IEEE Geosci. Remote Sens. Lett. 11(3), 691–695 (2014).
5. H. Li et al., “Gabor feature based unsupervised change detection of multitemporal SAR images based on tow-level clustering,” IEEE Geosci. Remote Sens. Lett. 12(12), 2458–2462 (2015).
6. T. Celik, “Unsupervised change detection in satellite images using principal componentanalysis and k-means clustering,” IEEE Geosci. Remote Sens. Lett. 6(4), 772–776 (2009).
7. J. Li 等人, “基于频域标度空间分析的视觉显著性” IEEETrans. Pattern Anal. Mach. Intell. 35(4), 996–1010 (2013).
8. Q. Zhang 等人, “随机多图: 用于高维数据分类的半监督学习框架” Image Vision Comput. 60, 30–37 (2017).
9. L. Breiman, “随机林” Mach. Learn. 45(1), 5–32 (2001).
10. W. Liu, J. He, and S. F. Chang, “Large graph construction for scalable semi-supervisedgraph-based method,” in Proc. Int. Conf. on Machine Learning, pp. 679–686, (2010).
11. “Earthnetonline,”<http://earth.esa.int/ers/ers_action/SanFrancisco_SAR_IM_Orbit_47426_>

20040516.html (11 May 2017).

1. R. L. Brennan and D. J. Prediger, “Coefficient kappa: some uses, misuses, and alternatives,”Educ. Psychol. Meas. 41, 687–699 (1981).
2. M. Gong 等人, “具有改进 MRF 能量函数的模糊聚类分析在合成孔径雷达图像中的变化检测” IEEE Trans. Fuzzy Syst. 22(1), 98–109 (2014).

# 致谢

########## waiting to be filled.

岁月如梭，三年的大学生活即将结束，在此瑾向大学生涯中的所有老师和同学表示深深的谢意。

感谢我的毕业设计的指导老师高峰老师对我的照顾和指导。不管是在生活上还是在学习中，高峰老师都给我很多的帮助，给我解决困惑。在准备毕业的过程中，高峰老师给我提供建议和帮助，我的毕业准备才得以顺利进行。在此次毕业论文写作中，从论文选题到整个毕业论文完成，每当我遇到问题，高峰老师都是引导我、帮助我和鼓励我完成任务。高峰老师为人友善、亲和，高峰老师的处事风格和治学态度，让我在以后的人生道路上受益匪浅。

因为我是申请了国外的院校，我也荣幸的请到高峰老师做我的推荐人，高峰老师为我写了推荐信，令我非常感动。在此之前高峰老师还带领我们的小组进行竞赛，并得到了奖项。在那场竞赛中，高峰老师给我们的项目提供了很多帮助，非常感谢。

同时也要感谢大学生涯中的每一位老师，唐瑞春老师、秦勃老师、黄磊老师和魏振钢老师等等，在我的本科三年中，教会了我了丰富的想法和传授了最新的知识，给了我很大的启发。

感谢所有的同学给我的帮助，每当我有问题时，是你们给我支持和莫大的安慰，帮助我一起解决生活上和工作上的困惑。我的大学生活充满了欢乐和感动。

最后，我要感谢我的父母，是你们对我关怀使我在遇到困难时才不退却，感恩与你们的爱将永远铭记于心。