**实验十 机器学习之聚类实验**

**一、实验目的：**

（1）理解Kmeans、Meanshift、层次聚类算法等算法的特点及主要的应用情境，能够利用sklearn中相关库解决现实世界中的各类聚类问题；

（2）掌握利用matplotlib对相关数据、图像、模型等信息进行可视化的方法；

（3）掌握对聚类模型结果进行分析和解释的能力。

**二、实验仪器及实验环境**

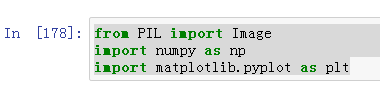
（1）硬件：PC机；

（2）软件：Anaconda Jupyter Notebook，Spyder，Pandas，Sklearn

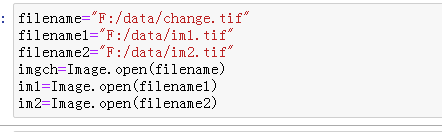
**三、实验内容：**

（一）基于聚类算法的遥感影像变化检测实验

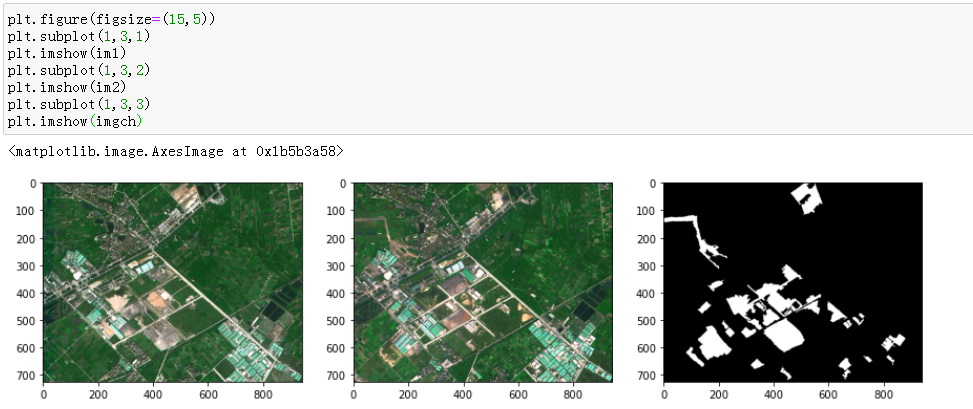
1. 导入相关库，其中包括对图像进行读取的PIL库中的Image对象。

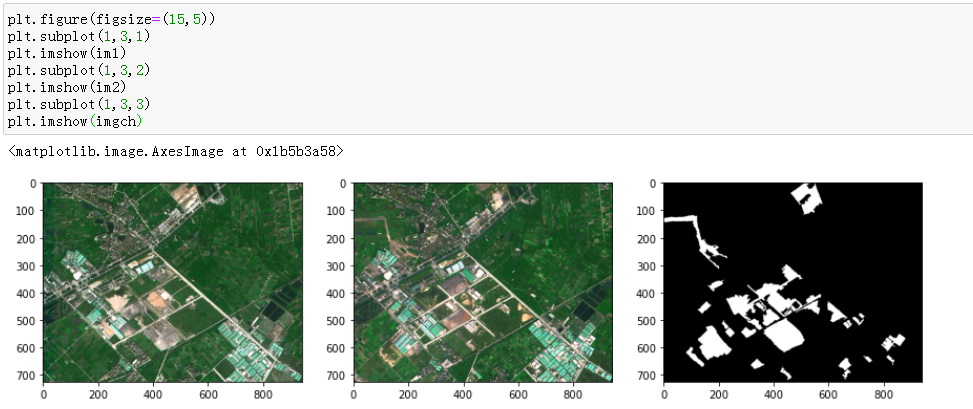


1. 利用Image对象打开图像文件。其中im1，im2为前后两期遥感影像。imgch为变化检测标注结果。

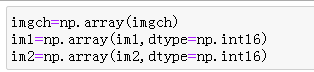


1. 查看图像。Imgch中白色（像素值为255）的区域为变化区域，黑色（像素值为0）区域位未变化区域。

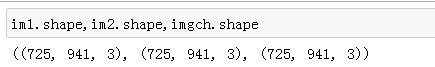




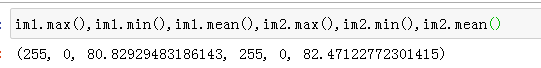
1. 将图像转换为数组对象，数据类型选择两个字节整数，这个非常重要否则，Image对象的每个像素的数据类型默认为np.uint8，也就是无符号1个字节整数，用0-255表示光谱范围的值。如果不转换，两个图像相减时负值将表示为无符号型整数，出现混乱。



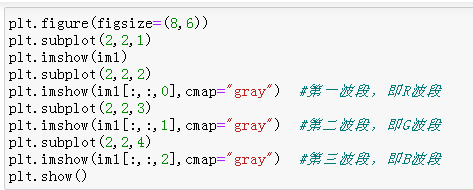
1. 查看图像数组的特征。

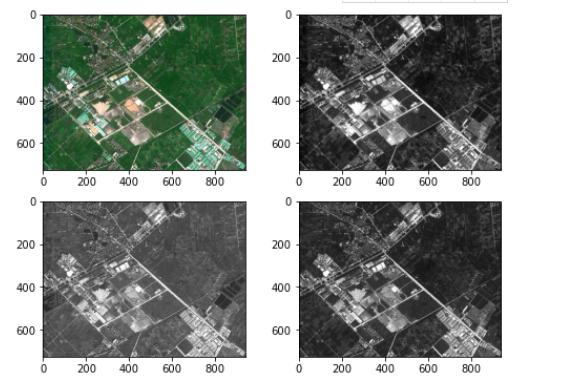


从中可以看出，彩色图像中每个像素的彩色值是由三个整数组成的，分别为R、G、B三个波段，是一个三维数组。图像为725行，941列，也就是图像的长宽分别为941和725。



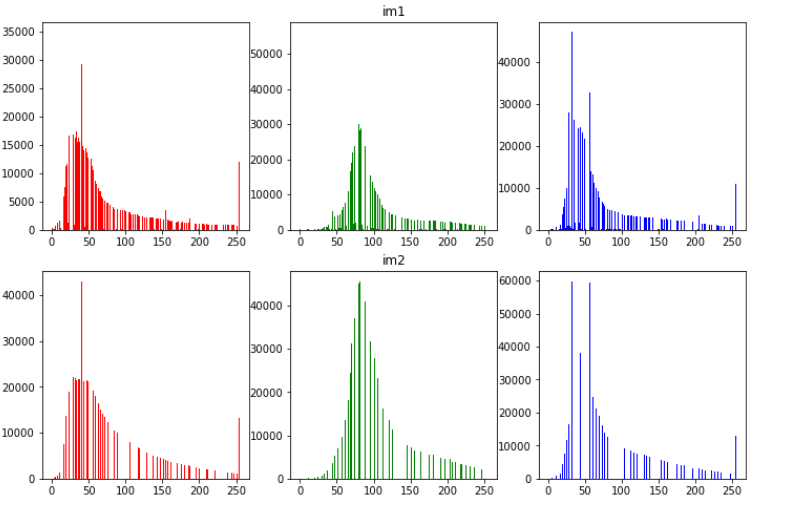
从中可以看出，图像光谱值的最大值和最小值分别为0和255。光谱均值分别为80和82差别不大。可以查看R、G、B每个波段上的灰度分布情况。

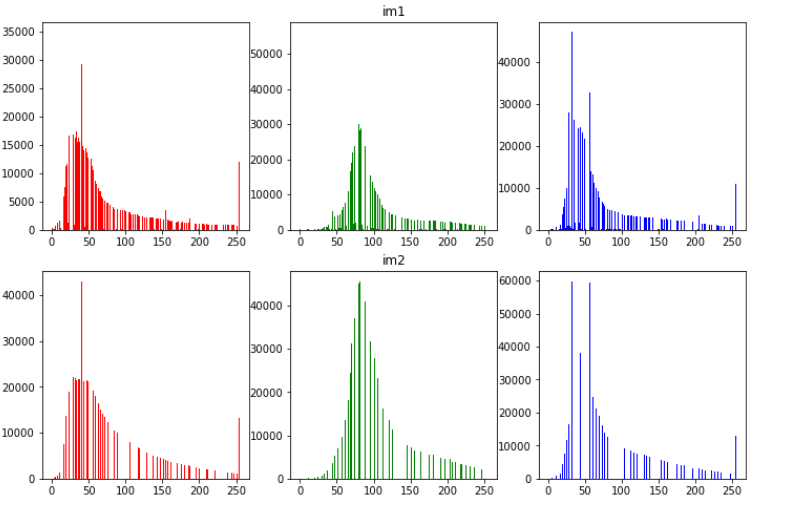




1. 查看图像的灰度分布。

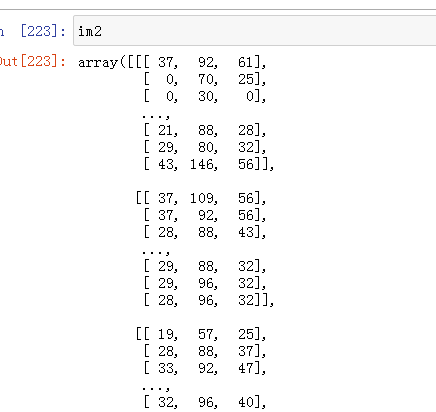
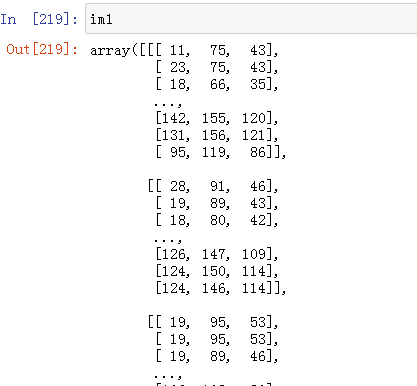




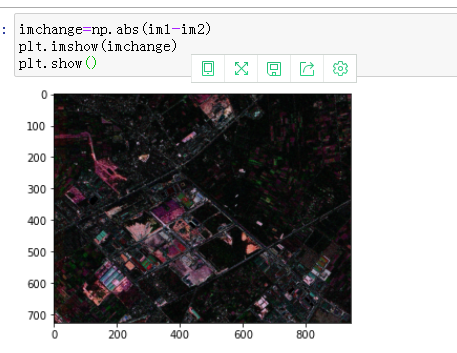


从图中可以看出，两个时相图像中灰度分布都不是标准的正态分布，红色波段数据集中于20-40部分。

1. 查看两时相图像对应的数据。

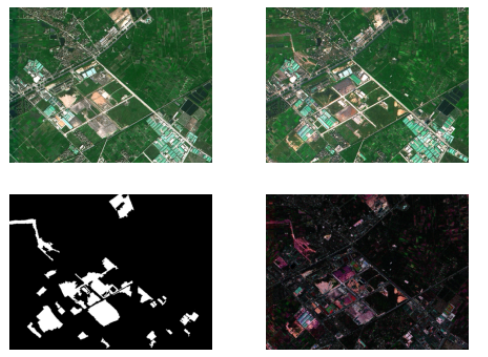


1. 利用im1和im2生成差异图像。

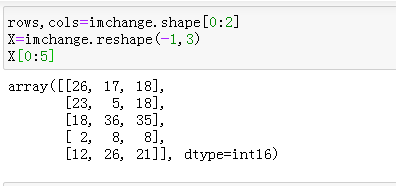


从中可以看出，发生变化的部分变化幅度较大，其明暗程度能够指示区域的变化情况。

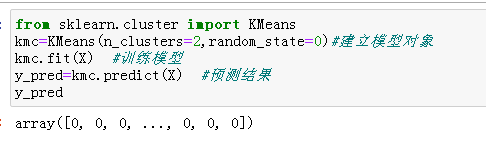
1. 对比查看原始两时相数据、变化标注、差异图像。



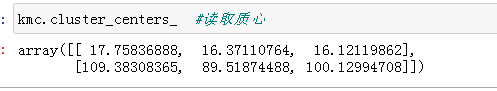
1. 利用差异图像生成聚类样本集，每个像素一个样本，每个样本三个光谱分量。



1. 利用KMeans算法对差异样本，进行聚类，将所有像素分类两类，变化类和未变化类。

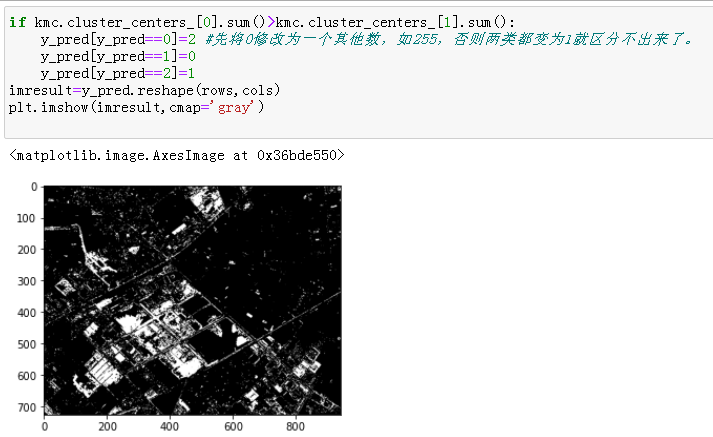


1. 显示两个类别的质心。



其中第一个是0标注类的质心，第二个为1标注类的质心。从其中可以看出0标注类的质心为变化幅度，小于1标注类质心的变化幅度。但是，并不是每次都这样，KMeans为每类标注的结果0和1是随机的，也有可能0标注类的质心变化幅度大于1标注类，这样时0表示变化类，1表示未变化类。而进行显示时，如果想用白色显示变化类，那么就需要将变化类的值修改为255，为变化类的值修改为0。

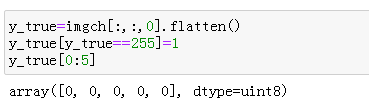
1. 处理两个类的像素点值。当0标注类的质心变化幅度大于1标注类的质心时，将0的值修改为1,1的值修改为0。



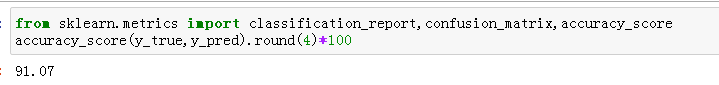
1. 将检测结果转换为Image对象后，保存图像。

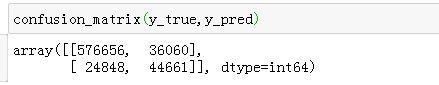


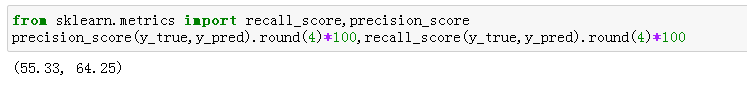
1. 利用标注图像，生成变化类/非变化类标注结果y\_true。由于标注图像是黑白图像，三个波段值都相同。采用0波段的值制作标注结果集，同时标注图像中用255来表示变化类，需要修改为1。



1. 利用变化类预测结果和真是标注进行模型准确性评估。

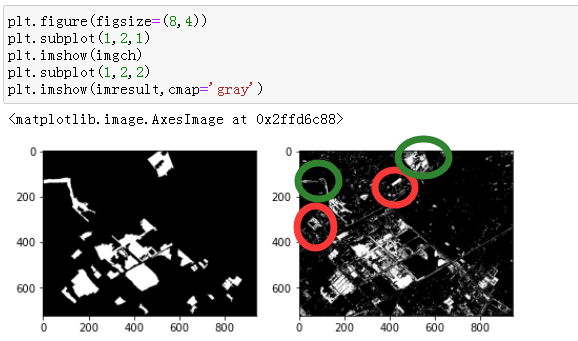






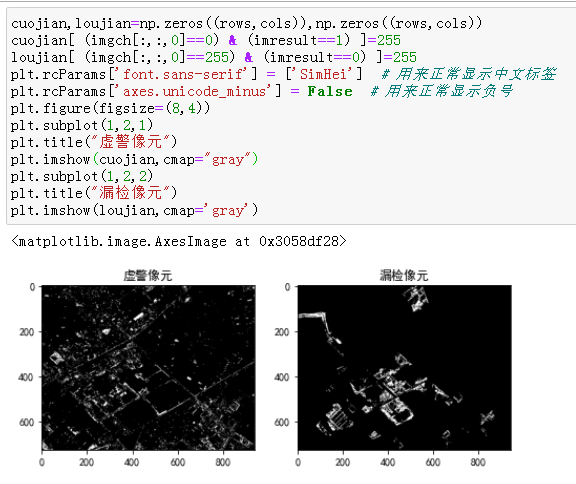
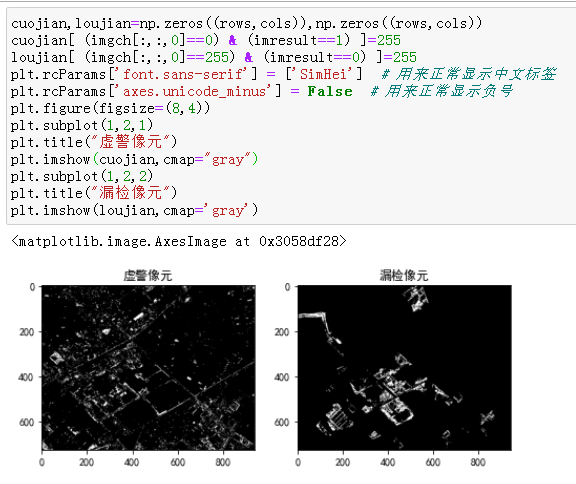
从中可以看出，准确率为91.07，混淆矩阵如上图所示。有36060个未变化像素被错误的分为了变化像素，同时有24848个变化像素被错误的分为了未变化像素，精确率为55.33%，召回率为64.25%，也就是说，虽然准确度为91.07%，比较高。但是，检测变化类的像素中只有55.33%是正确的，只有64.25%的真实变化像素被检测出来，精确率和召回率都相对较低。

1. 利用图像对参考图像和预测图像进行对比。



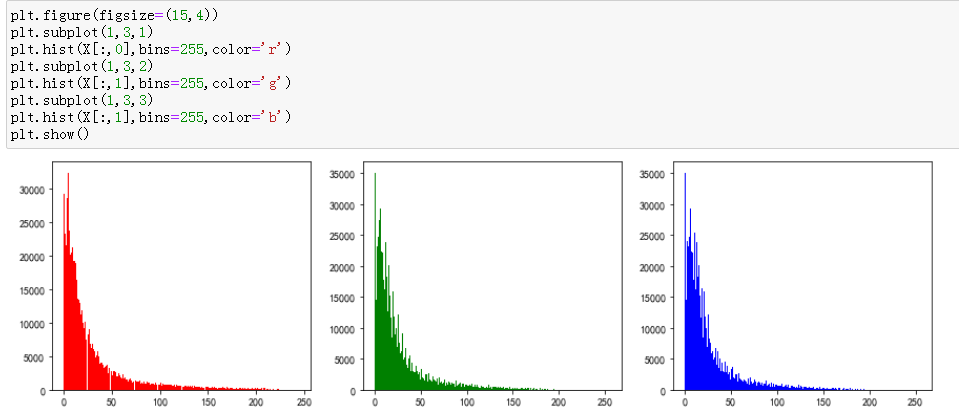
从中可以看出，预测结果中包含大量的虚警情况（红圈所示区域），图像中散布着很多比较零散的变化区域。同时变化区域内部不纯粹，其中很多变化像素没有检测出来。但是变化区域基本都能检测到，不存在漏检区域。但是，变化区域的边界较小，不纯粹。

1. 虚警（漏检）像元和漏检像元的分布情况。



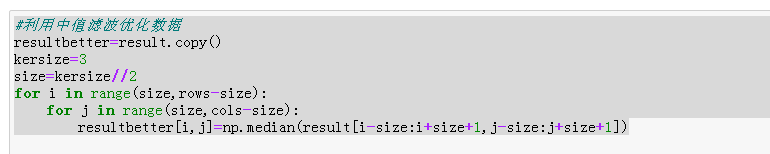
从图中可以看出，变化检测的结果中虚警像元的分布比较零散，这是由于高分辨率遥感影像在对地物进行精细化表达的同时，也会导致大量噪声的产生，带来变化检测的误诊。漏检像元都是由于变化区域周边变化幅度较小的区域，在聚类时错误的分为了未变化类。

1. 查看差值图像中各波段数据的分布情况。

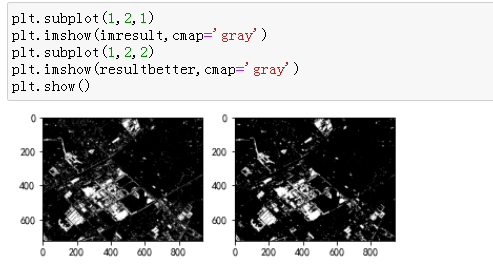


从中可以看出，变化类像素和未变化类像素之间的数据连续性较大，区分度并不好，所以导致很多错分像元的产生。

1. 可以利用中值滤波对检测结果进行优化。中值滤波的原理是，设置一个滤波范围size（比如为3），对于每一个像素，利用以其为中心周围3行3列的9个像素的像素值进行排序，取最中间的值作为该像素的值。

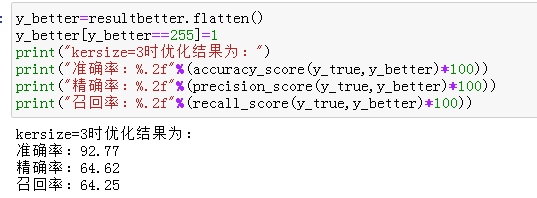


1. 对优化结果对比显示。

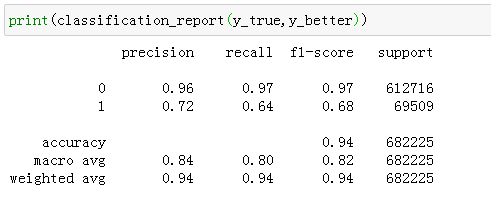


从中可以看出，图像中的部分虚警像元被处理掉了。

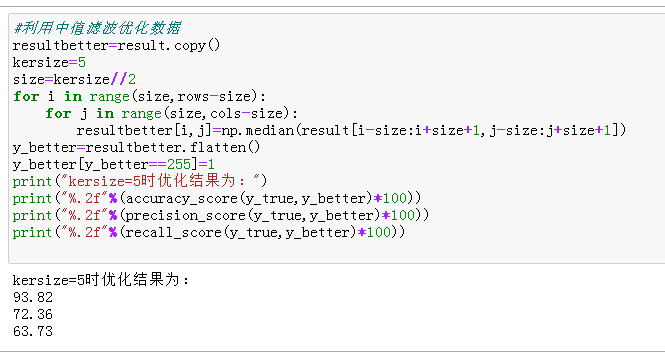
1. 对优化结果的准确性进行评估。



从中可以看出，利用卷积核为3的中值滤波进行优化后，准确率从原来的91.07上升到92.77，精确率从55.33%上升到64.62%。详细分类结果为：



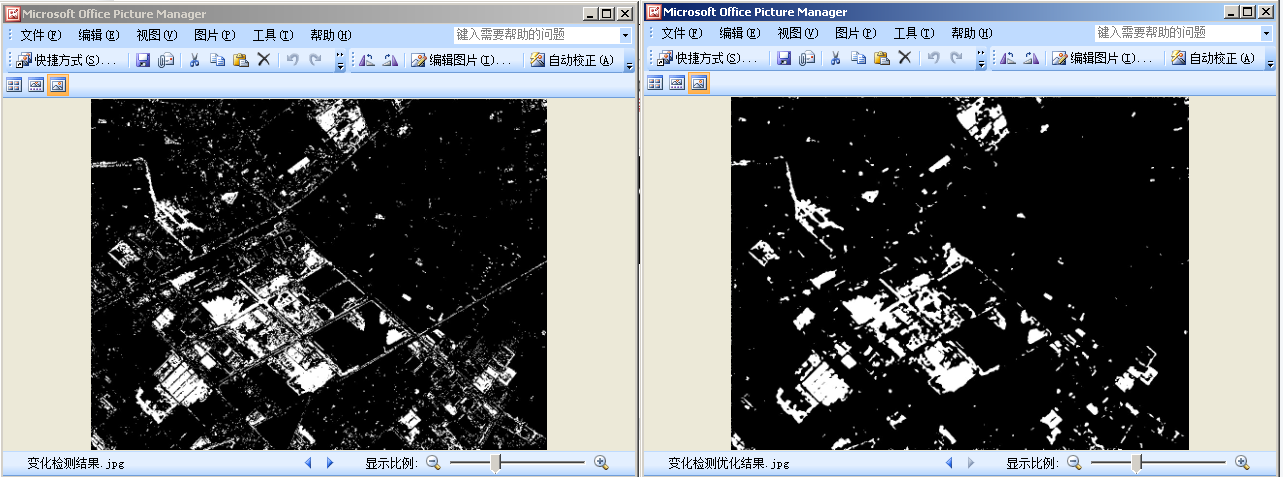
1. 增大卷积核，将其设置为5，对优化结果进行评估。从中可以看出，增大卷积核后，准确率上升到93.82%，精确率从55.53%上升到72.36%，但是召回率从64.25%下降到63.73。也就是说，利用中值滤波进行优化时，可以起到提高精确率的效果，但是会导致更高的漏检率。综合评估，利用卷积核为5的中值滤波结果作为最终的检测结果。



1. 将优化后的结果转化为图像进行保存。



打开优化前后的图像进行对比分析后发现，优化后零散的虚警像元处理的效果良好。



实验结论：

（1）利用KMeans聚类方法可以在没有任何先验知识的情况下进行变化区域的检测，且准确率良好。

（2）利用中值滤波可以对变化检测结果中的零散噪声问题进行优化。

（3）中值滤波对漏检问题的改良性能较差。

**四、实验结果与分析：**

（1）分析案例中聚类算法解决分类问题的准确性，并从实际角度对漏检和错检问题进行解释。

（2）针对分析过程中出现的问题和错误，写出问题分析和处理过程。