Land Cover Classification and Identification

Xintian (Stella) Li, David (Yuanrong) Pan, Jacey Chang
CONTENTS

01 Introduction
   Problem and Dataset

02 Data Preprocessing
   Relabeling and PCA

03 Machine Learning
   K Means, SVM and Random Forest

04 Deep Learning
   CNN (Binary & Multiclass), Transfer Learning (MobileNet & VGG)
Introduction

Problem and Dataset
<table>
<thead>
<tr>
<th>Problem</th>
<th>Motivations</th>
<th>Technical Possibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to human activities, land cover is changing rapidly over the recent decades.</td>
<td>Monitoring land cover changes help us understand and evaluate the urban development process and its associated environmental impacts</td>
<td>The proliferation of satellite image data and the emergence of advanced machine learning technologies make it possible to identify land cover and quantify the changes automatically.</td>
</tr>
</tbody>
</table>
EuroSAT Dataset

a dataset based on Sentinel-2 satellite images and consisting out of 10 classes with in total 27,000 labeled images, each 64 pixel * 64 pixel * 3 channels.

The 10 labels: Industrial Buildings, Residential Buildings, Annual Crop, Permanent Crop, River, Sea & Lake, Herbaceous Vegetation, Highway, Pasture, Forest
Visualize the Images
Examples from each label

AnnualCrop
AnnualCrop
Forest
Forest
HerbaceousVegetation

HerbaceousVegetation
Highway
Highway
Industrial
Industrial

Pasture
Pasture
PermanentCrop
PermanentCrop
Residential

Residential
River
River
SeaLake
SeaLake
02
Data Preprocessing
Relabel and PCA
We relabel the data to make the data **balanced** and faster to train.

<table>
<thead>
<tr>
<th>Original Label</th>
<th>New Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnnualCrop</td>
<td>Agriculture</td>
</tr>
<tr>
<td>PermanentCrop</td>
<td></td>
</tr>
<tr>
<td>Pasture</td>
<td></td>
</tr>
<tr>
<td>River</td>
<td>Water</td>
</tr>
<tr>
<td>SeaLake</td>
<td></td>
</tr>
<tr>
<td>Residential</td>
<td>Building</td>
</tr>
<tr>
<td>Industrial</td>
<td></td>
</tr>
<tr>
<td>Forest</td>
<td>Forest</td>
</tr>
<tr>
<td>HerbaceousVegetation</td>
<td>HerbaceousVegetation</td>
</tr>
<tr>
<td>Highway</td>
<td>Highway</td>
</tr>
</tbody>
</table>
Dimensionality Reduction: PCA

Dimensionality reduction is needed because huge data size and thus a long training time with the full dataset.

The first PC explains over 60% of the variance
The first 30 PCs explain over 82% of the variance
The first 100 PCs explain about 88% of the variance

We choose to use first 30 PCs in our models.
Visualize PCA Components

First 30 PC standardized, balanced dataset

It is shown in the PCA that the average hue of images matters, especially how much “yellowness” in the image.
03

Machine Learning

K-Means, SVM and Random Forest
KMeans Clusters

We first split the 30-pc dimensionality-reduced data into a 70% training set and a 30% testing test. Then we tried 6-cluster KMeans as an unsupervised model to see whether the data were easy to separate.

<table>
<thead>
<tr>
<th>actual_label</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agriculture</td>
<td>129</td>
<td>219</td>
<td>146</td>
<td>53</td>
<td>166</td>
<td>7</td>
</tr>
<tr>
<td>Building</td>
<td>202</td>
<td>24</td>
<td>170</td>
<td>67</td>
<td>257</td>
<td>0</td>
</tr>
<tr>
<td>Forest</td>
<td>8</td>
<td>136</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>576</td>
</tr>
<tr>
<td>HerbaceousVegetation</td>
<td>242</td>
<td>161</td>
<td>124</td>
<td>18</td>
<td>151</td>
<td>24</td>
</tr>
<tr>
<td>Highway</td>
<td>235</td>
<td>232</td>
<td>38</td>
<td>41</td>
<td>135</td>
<td>39</td>
</tr>
<tr>
<td>Water</td>
<td>160</td>
<td>235</td>
<td>4</td>
<td>0</td>
<td>102</td>
<td>219</td>
</tr>
</tbody>
</table>

Except for recognizing "Forest" as a distinct cluster (cluster 5) with an rough accuracy around 67%, K-Means did a poor job to distinguish different labels.
The best accuracy is about 67%, with an rbf kernel and parameter C=15.

The confusion matrix for the prediction of the best SVM model:
Random Forest Classifier

Random forest classifier had a similar performance with SVM classifier: The best model has a 68% testing accuracy.

Confusion matrix

Rank of model accuracies using different parameters
Deep Learning

CNN, Transfer Learning Models
First, we tried to use CNN to solve a simple 2-label classification question. We picked two classes which are very different from each other, buildings and water. Then we created a subset of data with 2,400 images in each of the two categories and perform CNN on the dataset.

**Building vs. Water**
Even a simple CNN network with only 1 convolutional layer and 1 dense layer can predict the 2 labels with a test accuracy of **96%**
Convolutional Neural Network (Binary Classification)

The model predicts all the buildings correctly. The ROC curve is very close to the upper left corner, also indicating a good diagnosis ability of the 2-label classifier.
The multi-class classification (6 labels) has a test accuracy of **83%**. The model is best at predicting Label 1 (Forest), not very good at predicting Label 0 and 5 (Agriculture and Water).

The accuracy is much higher relative to the ML classifiers.
Transfer Learning Models

Then we imported two pre-trained models (VGG16 and MobileNetV2) to fit and test on our own datasets.

We set input shape to (64, 64, 3), froze the convolutional layers of the pre-trained model, transferred to our datasets and updated the dense layers to get the output labels.

VGG16 accuracy: 85.5%
MobileNetV2: accuracy: 84.8%
Transfer Learning Models

MobileNetV2: accuracy: 84.8%

VGG16 accuracy: 85.5%
Thank you!

If you have any questions, please contact us at:
- xintianl@upenn.edu
- yrpan@upenn.edu
- jzchang@upenn.edu