Multi-sensor dust detection using machine learning

Team 7:
Julie Bessac (Mathematics and Computer Science Division, Argonne National Laboratory),
Ling Xu (Department of Mathematics, North Carolina A&T State University),
Manzhu Yu (Department of Geography, Penn State University)

Faculty mentor:
Dr. Aryya Gangopadhyay (Department of Information Systems, University of Maryland Baltimore County)

External mentor:
Yingxi Shi (University of Maryland Baltimore County)

Research Assistant: Pei Guo (Department of Information Systems, University of Maryland Baltimore County)
Background

- Dust and sand storms originating from Earth’s major arid and semi-arid desert areas can significantly affect the climate system and health.
- Recently, researchers utilize machine learning techniques to detect dust in multispectral imagery from satellites based on Lidar-based dust profiles.
- Projects in previous Cybertraining classes have been studied the same problem, but focusing on classification of pixels along the tracks of CALIPSO.
- Instead of classification, this project focuses on using unsupervised machine learning to extract and segment dust regions from VIIRS granule imagery.
- Building on the foundations of existing outcomes, we will also use the collocated dataset from the two satellite observations: CALIOP and VIIRS.
Related works

- **Existing heuristic methods:**
  - utilize the brightness temperature difference (BTD) between Thermal Infrared (TIR) bands at around 11 μm and 12 μm wavelengths to detect dust clouds over land surfaces
  - **Pro:** simple criteria, easy to implement
  - **Con:** sensitive to the different dust events, study areas, or different season

- **Machine learning (including deep learning) methods:**
  - Lazri and Ameur (2018): classify cloud type and estimate rainfall intensity
  - Strandgren et al. (2017): using Artificial Neural Network (ANN) to study the characteristics of cloud and aerosol based on both SEVIRI and CALIOP
  - Kolios and Hatzianastassiou (2019): utilized an ANN model to detect dust outbreaks in the Mediterranean region
  - **Pro:** address the limitation of fixed thresholds
  - **Con:** detecting the extent of dust is still lacking
Data: CALIOP

- Lidar-based observation
- Provides aerosol vertical distribution
- ½ km, 1 km, or 5 km (to be used)

- Classes of aerosol include:
  - marine,
  - dust,
  - polluted continental/smoke,
  - clean continental,
  - polluted dust,
  - elevated smoke,
  - dusty marine

Figure credit: https://www-calipso.larc.nasa.gov/
Data: VIIRS

- The VIIRS sensor has 16 M bands with 750 meter native resolutions from 412 nm to 12 micron, and 5 I bands with 375 meter resolution
- Aerosols can be retrieved using split window methods
- Broader spatial coverage

Figure credit: worldview.earthdata.nasa.gov
Data: VIIRS data download and data preprocess

- Besides the collocated VIIRS and CALIOP data prepared by Team 5 of CyberTraining 2019, we also downloaded VIIRS granule (VNP02MOD and VNP03MOD products) using the API provided by https://sips.ssec.wisc.edu/#/products/api
- In this preliminary study, we selected three spatiotemporal ranges:

1) North Atlantic Ocean (74W-20W, 13N-43N) for the whole 2014
2) Asian dust (110.9E-135.85E, 28.26N-44.38N) in Spring season (March, April, and May) in 2014
3) Northern Africa, Europe, and the Mediterranean (30W-60E, 0N-60N) in the Summer season (June, July, and August) in 2014
Data: VIIRS data download and data preprocess

Illustration of data sets at a selected area in North Africa and Caribbean, (a) VIIRS dust composite, (b) VIIRS true color composite, (c) enlargement of the top left corner in (a), (d) enlargement of the top left corner in (b), (e) the dust category on CALIPSO track.
Methods: Workflow

1. Derive dust signature from all CALIOP-VIIRS collocated data within the study area:
   - On-track pixel categorization:
     - category 1: pure dust
     - category 2: pure polluted dust
     - category 3: dust with polluted dust
     - category 4: dust or polluted dust with other aerosols
     - category 5: other aerosol only
     - category 0: anything out of the first 5 categories

2. Cluster each single VIIRS granule subset:
   - VIIRS granule subset, collocated with CALIOP
   - 256 pixels
   - Unsupervised machine learning / clustering
     - Determining optimal K value
     - Clustering

3. Segmentation result:

4. Dust cluster determination:
   - Calculating similarity between each cluster's M1-M16 bands and on-track dust pixels' 16 bands

5. Dust extent:

6. Validation:
   - Validate clustering result with on-track pixels
   - Validate dust extent with the VIIRS Aerosol Environmental Data Record (EDR)
Methods: Step 1-4

In **Step 1**, based on the VIIRS CALIPSO collocated data, pixels on CALIPSO tracks are categorized into groups related to dust. Category 1 (pure dust) will be considered as the dust pixels, and the other categories are considered as dust-free in our first trials of experiments.

In **Step 2**, each prepared VIIRS granule subset is clustered using K-means, where the number of clusters (K) is determined using the L-curve method for optimization.

In **Step 3**, the segmentation result is generated, where each cluster occupies a proportion of the VIIRS granule subset.

In **Step 4**, dust signature of the study area is generated based on all dust pixels on CALIPSO tracks, and the dust signature is essentially a matrix with each dust pixel as a row, and their corresponding VIIRS spectral band values as each column.
Methods: Step 5-7

In **Step 5**, to determine which resulting cluster is more likely to be dust, similarities of the VIIRS spectral band values between each cluster in the segmentation result and the dust signature. Cluster(s) with high similarity values with the dust signature will be considered as dust cluster(s).

In **Step 6**, the resulting dust extent is generated.

In **Step 7**, pixels on track of CALIPSO are used to validate the resulting dust extent. The validation with existing aerosol products, such as the VIIRS Aerosol Environmental Data Record (EDR), is still ongoing.
**Methods: Unsupervised machine learning**

Unsupervised techniques are used when no extra information is known about the quantity of interest to learn or predict

- K-means clustering is a method that partitions a dataset into K sub-group (cluster)
- Each cluster \( C_k \) is identified by its mean \( m_k \) value and generally an arbitrary label \( k \)
- Observations from the dataset are assigned to the cluster with the nearest mean (through most of time the Euclidean distance)

- The optimal number of clusters \( K \) is determined empirically through the L-curve or elbow method
- Many variations of the K-means have been proposed in the literature: based on different initialization methods, distances, and different cluster representants such as the K-medoids where the median of each cluster is used instead of the mean
Methods: Dust cluster determination

- After obtaining the clusters based on the unsupervised machine learning algorithm, it is essential to determine which cluster (or potentially multiple clusters) represents dust.
- The cluster determination process relies on the collective dust signature within the 16 spectral bands reflected by the CALIOP-VIIRS collocated data.
- **Similarities (based on Euclidean distance)** between each cluster and the dust signature matrix are calculated:
  - The cluster that has the highest similarity to the dust signature matrix is considered a dust cluster.
  - If the similarity values of other clusters to the dust signature matrix are within a valid range, i.e., the similarity values are also high enough, then these clusters are considered as potential dust clusters.
  - Potential dust clusters can complement the small dust region effect when the number of clusters (K) is set large.
Methods: Dust cluster determination

- Colored pixels represent the centroids of each cluster when a K-means clustering is performed with 4 clusters on the example dataset.
- In this example, cluster C0 is visually the closest from the bands corresponding to pure dust (category 1 in central column).
- Euclidean distances computed confirms the closeness of cluster C0 to the bands categorized as pure dust.

Boxplot of the 16 bands extracted on CALIPSO-track.
Methods: Dust cluster determination

- Statistics exploring each cluster:
  - Repartition of clusters on the whole area and along the CALIOP track
  - Cluster C0 minimizing the Euclidean distance between the centroids and the bands means in each dust-aerosol category is the most prevalent cluster
Methods: Dust cluster determination

Observations:
- Several clusters distributions differ significantly from the pure dust bands distribution on the left column.
- We use this set of statistics and metrics to determine the candidate cluster containing the most dust information.
Experiments and results

1. K-means clustering on single images using 16 VIIRS radiative bands
2. Compares accuracy results with two other methods, K-Medoids and Fuzzy C-means on several images and given several land-types
3. K-means on single images using 3 selected VIIRS radiative bands
4. Clustering on larger images in order to explore greater spatial extent of dust
Experiment 1: Dust extent extraction using K-means

- As a first set of experiments, the K-means clustering is performed on 256*256 pixels images

(North Atlantic region) Composite images of VIIRS granule subset at 2014234t1724, dust categories on CALIPSO track, and resulting dust extents segmented from our methods

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dust-free</td>
<td>1.00</td>
<td>0.41</td>
<td>0.58</td>
<td>137</td>
</tr>
<tr>
<td>Dust</td>
<td>0.59</td>
<td>1.00</td>
<td>0.74</td>
<td>118</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.68</td>
<td>255</td>
</tr>
<tr>
<td>Macro avg</td>
<td>0.80</td>
<td>0.70</td>
<td>0.66</td>
<td>255</td>
</tr>
<tr>
<td>Weighted avg</td>
<td>0.81</td>
<td>0.68</td>
<td>0.66</td>
<td>255</td>
</tr>
</tbody>
</table>
Experiment 1: Dust extent extraction using K-means

(Asian spring dust) Composite images of VIIRS granule subset at 2014147t0606, dust categories on CALIPSO track, and resulting dust extents segmented from our methods.
Experiment 1: Dust extent extraction using K-means

(Northern Africa summer) Composite images of VIIRS granule subset at 2014152t1112, dust categories on CALIPSO track, and resulting dust extents segmented from our methods.
Experiment 1: Dust extent extraction using K-means

Average silhouette: 0.749
Silhouette analysis for KMeans clustering on sample data with K = 4

(a) North Atlantic region example

Average silhouette: 0.377
Silhouette analysis for KMeans clustering on sample data with K = 4

(b) Asian spring dust example

Average silhouette: 0.707
Silhouette analysis for KMeans clustering on sample data with K = 4

(c) Northern Africa summer example
Experiment 2: Average accuracy using K-means within different study areas

- All three study regions have a median accuracy value around 0.6
- Northern Africa summer study area shows a higher median precision (~0.8) over the other two study areas (~0.6)
- However, the Northern Africa summer study area generally has a wider range of accuracy values than the other two study areas

Box plots of the accuracy, precision, recall, and F1-score using the proposed method over the datasets of three different study areas
Experiment 2: Average accuracy using K-means over different surface types

- The proposed method performs better over barren with a precision of \(~0.7\), whereas the accuracy over water bodies and other surface types result in \(~0.2\).

Box plots of the accuracy, precision, recall, and F1-score for all the images over different surface types.
Experiment 2: Average accuracy using K-means, K-medoids, and Fuzzy C-means

- Accuracy using different clustering methods, including K-means, K-medoids, and Fuzzy C-means did not show significant differences, therefore we continue our experiments using K-means.

Box plots of the accuracy, precision, recall, and F1-score for all the images using different clustering methods.
Experiment 3: K-means clustering on one single image using 3 VIIRS bands

- No significant difference between using 3 and 16 bands

(a) Predicted dust composite (using VIIRS 3 bands)  
(b) Predicted dust composite (using VIIRS 16 bands)  
(c) Dust composite
Experiment 4: Experiment using larger VIIRS granule subset

- Generally, with larger scale, the on-track accuracy improves.
- This accuracy improvement is expected because the sample size increases, and dust is easier to detect as a mid-scale meteorological phenomena.

Mean silhouette score: 0.3618
Future directions

- Investigate on semi-supervised techniques
- Additional variants of the proposed experiments setup can be tested to improve the interpretation of the clusters and the accuracy of the classification
- Further validate the resulting dust extents by comparing with other existing aerosol products