NC STATE UNIVERSITY

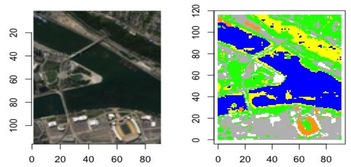
Department of Computer Science

GMM+Rule-Based Classifiers for Settlement Mapping

Amit Sanjay Watve, Felix Kim, Jonathan Raynor, Omkar Pradeep Acharya, Sridutt Bhalachandra, Sean Mahaffey
North Carolina State University

1. Introduction

- Map settlement and classify it based on satellite images
- Use GMM to determine relative pixel frequency of each component.
- The optimal number of components (k) is determined as one with the lowest Sum of Squared Errors (SSE) that can classify all major features in an image (k = 6 in current case)

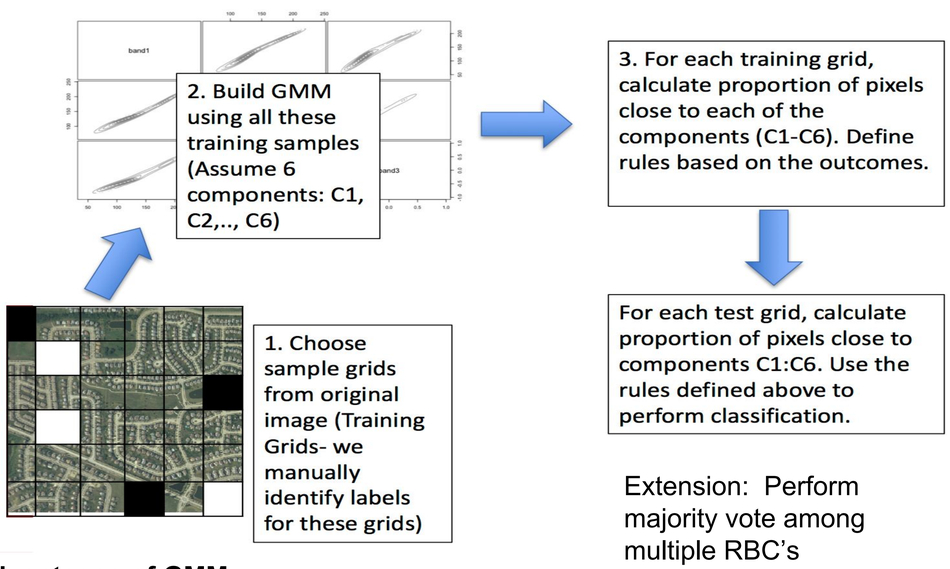


- Use Rule-Based Classifier to determine settlement class based on relative pixel frequencies
- Classify entire image as urban or rural based on settlement class frequencies

2. Data description

- Base data is a 3 band (Red/Green/Blue) raster GeoTiff file
- Spatial resolution of 5898x7696 with pixel size of 1m
- Test and training sets derived by extracting new rasters using the patch sizes designated and the embedded extent coordinates
- 10 training patches and 15 test patches (5 of each patch size)
- Random sampling and representative sampling were used

3. Method



Extension: Perform majority vote among multiple RBC's

Advantages of GMM

- Can handle clusters of varying sizes and shapes, and that overlap
- Component responsibilities more informative than hard assignments

Advantages of Rule-Based Classifiers

- Perform well when feature space can be divided into rectilinear regions
- Can handle imbalanced classes
- Relatively easy to interpret

4. Results

Residential Type 1 | Residential Type 2 | Water | Vegetation | Commercial

- RBC Accuracy for random sampling at different patch sizes:

Patch Size	50x50	75x75	100x100
Correct	4	1	0
Total	5	5	5
Accuracy	80.00%	20.00%	0.00%

- Found that as the patch size increased, accuracy decreased, as each patch was more likely to capture more than one class type for larger patch sizes due to the limitations of 1Rule RBC's.
- GMM did very well in soft-clustering the pixels into the 6 components. It did sometimes struggle to distinguish portions of the image having similar RGB values for certain components (eg. interpreting building shadows as water)
- Our RBC performed better on the 50x50 samples as compared to RIPPER from WEKA which has around 30% accuracy.

5. Parameter choices

- Number of components in Gaussian Mixture Model = 6
- Patch size = 50x50, 75x75, and 100x100 pixels
- 2 training patches for each class
- K-means used to determine the initial component responsibilities for the EM algorithm
- Number of patch class labels = 5
- Higher level class labels for entire image based on patch labels = 2 (Urban, Rural)

6. Conclusions

- We were able to draw a better understanding of the flexibility and usefulness of GMMs as opposed to solutions like k-means.
- We were able to use majority vote instead of a single RBC outcome to improve certainty of our results.
- Accuracy can be further improved by using weighted similarity rather than single rules.

7. References

[1] Gaussian mixture models. http://scikit-learn.org/stable/modules/mixture.html.

[2] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining, 2013.

[3]Reynolds, Douglas. "Gaussian mixture models." Encyclopedia of biometrics (2015): 827-832.

[4] Method image courtesy of Krishna Karthik Gadiraju (https://moodle-courses1617.wolfware.ncsu.edu/pluginfile.php/1012139/mod_folder/content/0/w2-c2-project-details.pdf?forcedownload=1)