BIODIVERSITY ASSESSMENT USING HIERARCHICAL CLUSTERING OVER HYPERSPECTRAL IMAGES

Ollantay Medina †, Vidya Manian ‡ and J. Danilo Chinea §

† Computing and Information Sciences and Engineering, University of Puerto Rico
at Mayaguez, Call box 9000, Mayaguez, Puerto Rico 00681
‡ Department of Electrical & Computer Engineering, University of Puerto Rico
at Mayaguez, Call box 9000, Mayaguez, Puerto Rico 00681
§ Department of Biology, University of Puerto Rico
at Mayaguez, Call box 9000, Mayaguez, Puerto Rico 00681

ABSTRACT

Hyperspectral images represent an important source of information to assess ecosystem biodiversity. In particular, plant species richness is a primary indicator of biodiversity. This paper aims to use spectral variance to predict vegetation richness, known as Spectral Variation Hypothesis. A hierarchical clustering method based on minimum spanning tree computations retrieve clusters whose Shannon entropy reflects the species richness on a given zone. These entropies correlate well with the ones calculated directly from field data.

Index Terms— Hyperspectral Images, Biodiversity, Minimum Spanning Tree, Clustering

1. INTRODUCTION

Hyperspectral images represent an important tool to assess ecosystem biodiversity. Improvements in spectral and spatial sensors resolution allow more precise analysis of biodiversity indicators that should agree with indicators obtained using field data. The development of accurate analysis tools would be advantageous to extend the analysis to larger zones by hyperspectral image processing; furthermore, for future scenarios, given that the actual data gathering is expensive in time and resources, a pre-analysis would give extra knowledge to plan a more effective campaign to collect field data.

One of the most important indicators of biodiversity in ecosystems is plant species richness. This indicator can be measured in hyperspectral images considering the Spectral Variation Hypothesis (SVH) proposed in [1], SVH states that spectral heterogeneity is related to spatial heterogeneity and thus to species richness. Subsequent efforts to validate SVH can be found in [2] and [3]. This paper aims to capture spectral heterogeneity by means of hierarchical clustering and then use the result for prediction of plant species richness. The information of heterogeneity can be used to compute entropies like Shannon, Gini-Simpson or Renyi among others [4]; in our case, Shannon entropy is chosen, given that is commonly used in biodiversity, and is the one used in the available field data.

There have been other studies trying to capture plant diversity starting from spectral heterogeneity like [5] and [6], one noteworthy difference with this study is the focus on the clustering method. The mentioned studies represent the spectral variability using the radius of the cluster, with no further explanation, this means the implicit assumption that the clusters are spherical, which is not always the case, in non spherical clusters the radius lose meaning. The method proposed in this paper doesn't rely on this assumption.

The clustering method to be used has to deal with clusters, not necessarily spherical, of irregular boundaries due to the nature of high dimensional data and the phenomenon of spectral mixture. A density based method like DBSCAN [7] could be a first option but it is computationally expensive; minimum spanning tree based methods [8] are also a good option for our problem, and have an appealing computational cost. The later method is used in an iterative hierarchical clustering process.

The study is conducted over data from the Guanica forest, a subtropical dry forest located in southwest Puerto Rico, considered the best preserved subtropical forest in the Caribbean.

This paper is organized as follow: Section 2 gives a brief theoretical background about minimum spanning tree, clustering and Shannon entropy. Section 3 explains the main algorithm used to process hyperspectral imagery. Section 4 explains the details of experiments and shows the results. Section 5 presents the conclusions.

2. THEORETICAL BACKGROUND

2.1. Minimum spanning trees

Let G = (V, E) the complete graph where V is the set of vertices and E is the adjacency matrix. V considers the pixels in the hyperspectral image and E is the dissimilarity matrix that considers the distances under some metric d(.) calculated for every pair of pixels. The minimum spanning tree T is the acyclic subset of edges that connects all V, whose weight d(T) is minimum.

$$d(T) = \sum_{(u,v)\in T} d(u,v) \tag{1}$$

2.2. Minimum spanning tree clustering

This algorithm is capable of detecting clusters with irregular boundaries. The first step of the process constructs a minimum spanning tree T; Kruskal's algorithm [9] is used to compute T. Then a partition of the set of vertices is produced by removing edges in T that satisfy a predefined criterion under the rationale: shorter edges should go between members of the same cluster and larger edges should go between members in different clusters; a well chosen criterion should transform T into a set of disjoint trees that represent clusters.

2.3. Shannon entropy

The Shannon entropy is a popular biodiversity index that represents the weighted geometric mean of the proportional abundances of the species. It is calculated as follows:

$$H = -\sum_{i=1}^{n} p_i \ln p_i \tag{2}$$

Where n is the number of species and p_i is the proportion of individuals belonging to the ith species.

2.4. Pearson's correlation coefficient

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

2.5. Metrics

Metrics used to construct the dissimilarity matrix.
Spectral angle distance:

$$SAD(X,Y) = \cos^{-1}\left(\frac{X.Y}{\|X\|\|Y\|}\right) \tag{4}$$

Euclidean distance:

$$d(X,Y) = \|Y - X\| \tag{5}$$

3. ALGORITHMS

3.1. Hierarchical Clustering algorithm

The hierarchical clustering algorithm has as core routine the minimum spanning tree clustering algorithm; some needed concepts are explained first.

3.1.1. Cut criterion
Let G=(V,E) a complete graph.
Let T be the minimum spanning tree.
Let $\bar{d}$ be the average distance of all edges of T.
Let σ be the standard deviation of all edges of T
The edge in T with distance d is removed if $d > \bar{d} + \sigma$

3.1.2. Gap criterion
A boolean predicate that will be true only if T remains the same after applying cut criterion with condition $d > \bar{d} + \sigma$ and $d > \bar{d} + 2\sigma$. The purpose of this gap is to make sure the algorithm iterates until more well defined clusters appears using representatives and hierarchical clustering.

3.1.3. Cluster representative
Given a cluster C, the representative point r of C is the closest point to the centroid of C. Cluster representatives will be used in the next level of the hierarchical clustering.

3.1.4. Algorithm
This algorithm is an iterative process that will produce a hierarchical clustering. Every step adds a new level to a hierarchy of clusters, the process stops when the gap criterion is met. For higher levels in the hierarchy, representative points of the clusters in the previous level are used. Consider input G as established in Section 2.1.

Hierarchical Clustering (G)
1. Let H be a hierarchy of clusters
2. Add a new level to H using data in G
3. Repeat
4. Compute minimum spanning tree T using the data in the highest level of H
5. Apply cut criterion to T
6. If gap criterion is not met
7. Compute representatives for T
8. Add another level to hierarchy H
9. Until gap criterion is met
10. Return clusters from hierarchy H

The complexity of this algorithm is given by the number of iterations or levels in the hierarchy which depends on the condition used in the gap criterion; let k be this number; inside the loop. The most important operation is the computation of the minimum spanning tree T, since the algorithm of Kruskal is used for this computation, every iteration has a cost of O(E log E) [9], giving a total running time of $O(k\ E \log E) = O(k\ V^2 \log V^2)$, as the graph G is complete.

3.1. Shannon entropy computation

Once the clusters are identified, the computation of the Shannon entropy index is almost straightforward. Considering Section 2.3, assign the number of clusters to n and for every cluster compute p_i as the ratio between the number of points in the cluster and the total number of points in the image.

4. EXPERIMENTAL RESULTS

The purpose of these experiments is to show that the algorithm proposed computes a biodiversity index that reflects to a certain degree the species richness in the hyperspectral image. The validation of the index is done by calculating the Pearson correlation coefficient between the Shannon entropy of actual field data and the Shannon entropy computed in the corresponding image using clustering results.

The data comes from the Guanica forest in Puerto Rico, the field data comes from 20 circular zones, each circular zone has a radius of 10 m., field data considered vegetation, specifically plants with stem radius larger than 5 cm.. The corresponding hyperspectral images comes from AISA airborne imagery, the images have 128 bands, wavelengths from 400 – 1000 nm, and spatial resolution of 1 m. The data is divided in 3 groups because there are 3 separated hyperspectral images taken at different times producing differences in the reflectance.

Under these specifications, every hyperspectral zone possesses 324 pixels, where every pixel is a vector of 128 components. A preprocessing stage using NDVI index is applied in order to filter out non vegetation pixels. For the dissimilarity matrix, two well known metrics were used, the spectral angle distance and the Euclidean distance. The hierarchical clustering algorithm processed the whole data batch in 25 s. approximately. Tables 1-3 show the results.

Figure 1 shows computed Shannon entropies for a hyperspectral image of Guanica using the proposed method.

Field Data	SAD	Euclidean
Shannon	Shannon	Shannon
0.440	0.336	0.593
1.134	0.561	1.011
1.950	0.371	0.919
1.352	0.454	0.748
0.376	0.121	0.608
1.787	0.718	1.450
Pearson	**0.66**	**0.73**

Table 1. Group 1. Field data entropy and computed entropies. Corresponding Pearson's correlation coefficients.

Field Data	SAD	Euclidean
Shannon	Shannon	Shannon
1.354	0.371	0.667
1.987	0.375	0.692
1.685	0.146	0.391
2.237	0.541	1.085
2.481	0.613	0.806
1.923	0.271	0.796
Pearson	**0.71**	**0.59**

Table 2. Group 2. Field data entropy and computed entropies. Corresponding Pearson's correlation coefficients.

Field Data	SAD	Euclidean
Shannon	Shannon	Shannon
1.133	0.163	0.396
1.984	0.499	1.630
2.230	0.582	1.402
2.219	0.386	0.691
1.315	0.167	0.363
2.290	0.454	1.132
1.852	0.437	1.072
2.339	0.230	0.563
Pearson	**0.65**	**0.52**

Table 3. Group 3. Field data entropy and computed entropies. Corresponding Pearson's correlation coefficients.

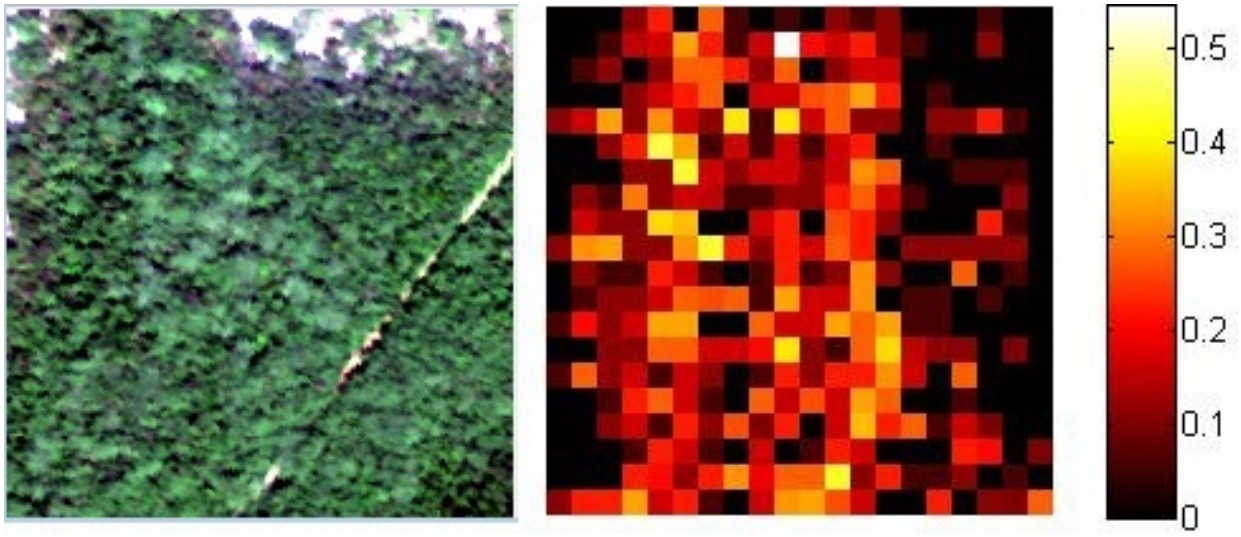

Figure 1. Left: A portion of the Guanica forest, size 200x200 pixels. Right: Computed Shannon entropies using the proposed method, zone size 10x10 pixels. Higher entropy means more proportional spectral heterogeneity.

5. CONCLUSIONS

A hierarchical clustering method is presented to capture the spectral variability of vegetation in hyperspectral images; the information obtained is used to assess plants species richness. The clustering method works with spherical and non spherical clusters.

Experimental results show that there is a significant correlation between the Shannon entropy of the field data and the one computed using the proposed method. Spectral angle distance exhibits stronger correlation than Euclidean distance, the explanation might be in the property of invariance to multiplicative scaling in spectra that SAD possess; these changes in spectra can be originated by differences in observation angle and illumination, both of them reasonable to expect for vegetation. The algorithm has a short running time.

6. ACKNOWLEDGMENTS

This work was supported by NASA under grant NNX09AV03A and by Laboratory for Applied Remote Sensing and Image Processing, University of Puerto Rico at Mayaguez.

7. REFERENCES

[1] M., Wohlgemuth, T., Earls, P., Arevalo, J.R., Thompson, S. D. Palmer, "Opportunities for long-term ecological research at the Tallgrass Prairie Preserve, Oklahoma," in *Proceedings of ILTER Regional Workshop, Budapest, Hungary*, 2000, pp. 123-128.

[2] M., Earls, W., Hoagland, B.W., White, P.S., Wohlgemuth, T. Palmer, "Quantitative tools for perfecting species lists," *Envirometrics*, vol. 13, pp. 121-137, 2002.

[3] D., Chiarucci, A., Loiselle, S.A. Rocchini, "Testing the spectral variation hypothesis by using satellite multispectral images," *Acta Oecologica*, vol. 26, pp. 117-120, 2004.

[4] L. Jost, "Entropy and diversity," *OIKOS Synthesising Ecology*, vol. 113, no. 2, pp. 363-375, 2006.

[5] D. Rocchini, "Effects of spatial and spectral resolution in estimating ecosystem ecosystem α-diversity by satellite imagery," *Remote Sensing of Environment*, vol. 111, 2007.

[6] D., Blakenhol, N., Carter, G., Foody, G., Gillespie, T. Rocchini, "Remotely sensed spectral heterogeneity as a proxy of species diversity: Recent advances and open challenges," *Ecological Informatics*, 2010.

[7] M., Kriegel, H. P., Sander, J., Xu, X. Ester, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *2nd International Conference of Knowledge Discovery and Data Mining*, 1996.

[8] O., Zhou, Y., Jorgensen, Z. Grygorash, "Minimum Spanning Tree Based Clustering Algorithms," in *18th IEEE International Conference on Tools with Artificial Intelligence*, 2006.

[9] T., Leiserson, C., Rivest, R., Stein, C. Cormen, *Introduction to Algorithms*., 2003.