BOOSTING WITH MULTIPLE CLUSTERING MEMBERSHIPS FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Giovanni Bellio
Department of Computer Science
Auburn University at Montgomery
Montgomery, AL 36117, USA
gbellior@aum.edu

Randy Russell
Department of Chemistry
Auburn University at Montgomery
Montgomery, AL 36117, USA
rrussell@aum.edu

Olcay Kursun
Department of Computer Science
Auburn University at Montgomery
Montgomery, AL 36117, USA
okursun@aum.edu

***Abstract*— A novel hyperspectral image classification algorithm is proposed and demonstrated on benchmark hyperspectral images. We also introduce a hyperspectral sky imaging dataset that we are collecting for detecting the amount and type of cloudiness. The algorithm is designed to be applied to the Such systems could improve the spatial and temporal resolution of cloud information vital to understanding Earth's climate. We discuss the nature of our HSI-Cloud dataset being collected and an algorithm we propose for processing the dataset using a categorical-boosting method. The proposed method utilizes multiple clusterings to augment the dataset and achieves higher pixel classification accuracy. Creating categorical features via clustering enriches the data representation and improves boosting ensembles. For the experimental datasets used in this paper, gradient boosting methods performed favorably to the benchmark algorithms.**

***Index Terms*— Hyperspectral Imaging, Cloud Segmentation, Categorical Boosting, Ensemble Learning, Clustering for Feature Extraction.**

I. INTRODUCTION

Compared to developing single models, ensemble learning algorithms that utilize decision trees (DTs) and boosting have received increasing interest due to many features including but not limited to their fast and accurate predictions, robustness to noise, ability to deal with diverse features such as both numerical and categorical features, having fewer parameters to optimize, and having a rule-based interpretability using if-then-like rules [1]. Gradient/Adaptive boosting methods based on decision trees, such as CatBoost and AdaBoost, can handle diverse data types and solve a wide range of machine learning problems involving categorical variables. The thrust of this work is to develop machine learning algorithms based on such boosting algorithms and test their applicability and prediction performances on hyperspectral image (HSI) datasets for pixel classification.

In Hyperspectral Imaging (HSI), a pixel is characterized by a high number of spectral channels/bands, thus allowing accurate and efficient classification of individual pixels [1][2][3]. HSI cameras vary in the number of wavelengths (bands) they have, but typically in an HSI dataset, every pixel is represented by several hundreds of bands. Typically, for every HSI image a ground truth image is provided that contains the class labels of individual pixels. The spectral signature in those bands (reflectivity/irradiances in different wavelengths) for a pixel can be used as a powerful predictor of the class-label (i.e., for classification of that individual pixel). Since different classes have different certain hyperspectral signatures, HSI can serve as an important pattern recognition goal, for example, for scanning a large field by taking aerial pictures. An exemplary application could be classification of individual pixels into various types of vegetation/soils/fields/trees [2], and another application could be classification of individual pixels into dark, regular, or thin clouds versus clear sky [3]. In such HSI applications, single pixel classification can help, for example, in monitoring the state of crops (wet/dry/rotten) or in finding irregularities/outliers such as a metal object camouflaged in the field.

To be able to take full advantage of aforementioned categorical boosting/ensemble methods in HSI images, we propose to use clustering for feature extraction. Clustering algorithms such as K-means can be used as a form of preprocessing (data summarization/quantization) tools [4][5][6]. In general, HSI datasets do not contain any categorical features (except some catalog information such as when and where the image is captured and other auxiliary information such as weather conditions). The proposed boosting algorithm applies multiple clustering runs and use the cluster memberships of HSI image pixels as super-features. These additional categorical features improve classification accuracy of the subsequent boosting classifier [7]. We propose and demonstrate that creation of categorical features using clustering algorithms proves very useful in enriching the data representation for categorical-boosting ensembles. This work presents a method that can be further developed for achieving dimensionality reduction, accurate classification, and parallelism needed for easy implementation in high-performance computing frameworks for hyperspectral image classification [1][8].

II. BACKGROUND AND DATASETS

The dataset we are currently collecting as a publicly available HSI dataset has 462 bands with the goal of cloud detection, classification, and segmentation. The images

collected are recorded with the Resonon Pica XC2 camera, which imaging system acts as a push-broom scanning spectrometer with 462 narrow wavelength bands ranging from 400 nm to 1000 nm. Semi-supervised pixel classification can be used to identify clear-sky pixels and different types of clouds for segmenting these images. As can be seen in Figures 1 and 2, even a gray image obtained using a single wavelength contains a wide variety of types of cloudy pixels. Therefore, we have decided to develop a boosting-based classifier that can be apply weak classifiers to various subsets of the bands. Once the single pixel classification is done effectively, then some form of postprocessing such as median filtering can be performed for segmenting a hyperspectral image into regions according to cloud type or clear sky.

Before moving into semisupervised setting, in this paper, we first demonstrate the proposed machine learning algorithm on two benchmark datasets for supervised HSI pixel classification, Indian Pines and Salinas datasets. These two datasets are well-known HSI datasets captured by the AVIRIS (Airborne Visible Infrared Imaging Spectrometer) sensor. The first dataset, called Indian Pines, is composed of images of 145×145 pixels in size, with each pixel of the image represented with 204 spectral channels (bands) in the 400-2500 nm range of wavelengths [2]. The dataset includes 17 classes (class-0 is unlabeled and the other 16 classes are various crops, grass, and woods). Table I lists the class names and the number of pixels per class in the dataset.

The second dataset, called Salinas, consists of images of 512×217 pixels in size with 204 spectral bands [2]. It includes 17 classes (class-0 is unlabeled and the other 16 classes are different types of vegetation). Table II lists the class names and the number of pixels per class in the dataset.

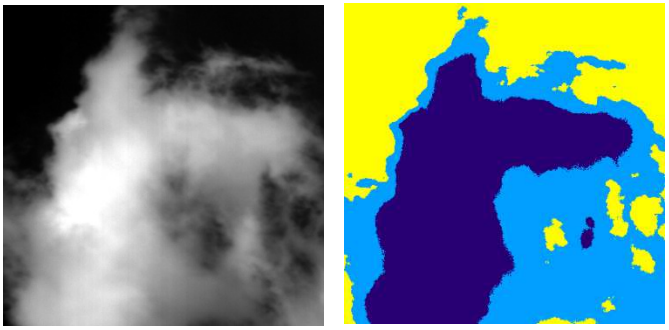

Figure 1. (Left Panel) One of HSI images collected rendered as gray-scale. This wavelength of 585 nm is selected to maximize the contrast between cloudy and clear sky based on the analysis summarized in Figure 2. (Right Panel) Ground-truth obtained by K-means clustering the image given in the left panel and classify each pixel into cloudy, thin-clouds, and clear sky classes.

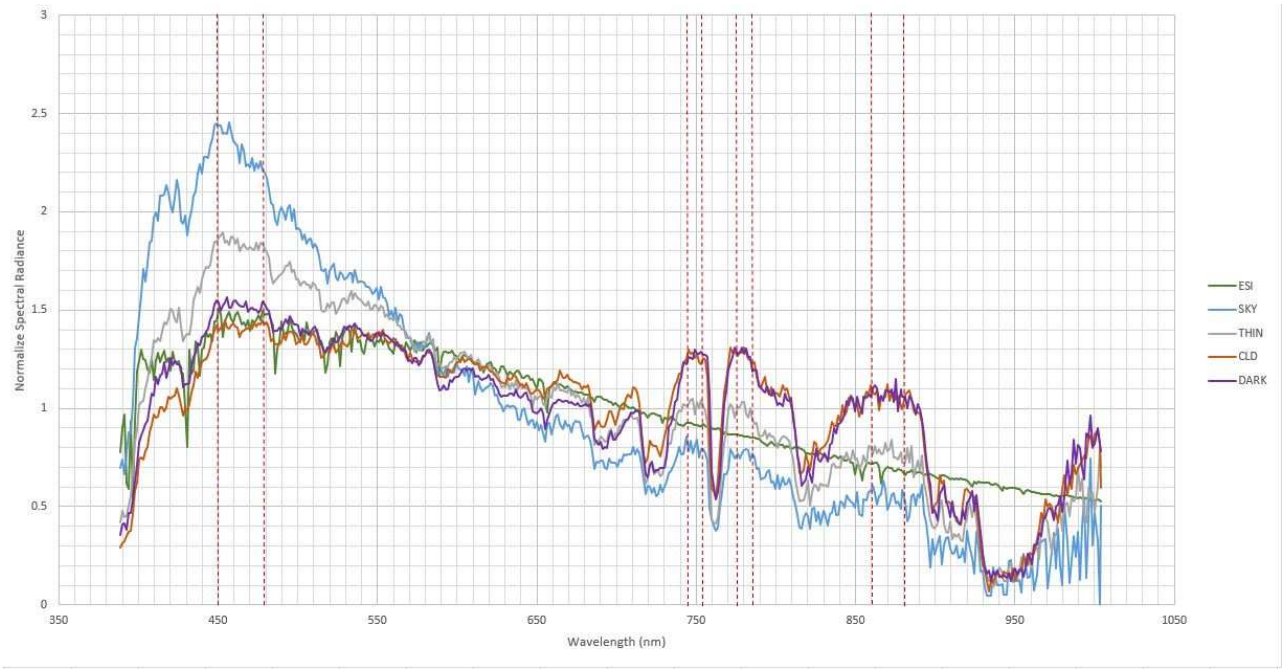

Figure 2. Signatures of clear sky (SKY), thin clouds (THIN), clouds (CLD), and dark clouds (DARK) in 462 bands of the HSI images being collected.

III. PROPOSED MACHINE LEARNING MODEL

While the high-resolution representation of an individual pixel in HSI (having many narrow wavelengths covering a large portion of the spectrum from near-ultraviolet to near-infrared range) makes discrimination of many more classes from each other using just a single pixel, reflectivity/irradiances in nearby wavelength intervals are generally very redundant and dimensionality reduction methods [4] are needed for band selection for HSI systems [1].

Applying some feature selection algorithms with high time complexity such as sequential backward selection (with time complexity of $O(n^2)$) is costly for such high dimensional datasets. Moreover, feature selection algorithms need to be further adapted to HSI domain, because the band selection for the classification task should also help determine important ranges of the spectrum. That is, feature selection process should not necessarily treat each one of the hundreds of wavelengths of the spectrum as separate or unrelated variables, because selection of individual wavelengths of the spectrum may not be justified and the task could be simplified by finding a few wavelength ranges of greatest importance. We identified four intervals to be the most useful for cloud classification (Figure 2 shows these intervals highlighted with red dashed vertical lines) Representing each interval as one categorical variable (i.e., creating a categorical variable for that interval by representing it with the index of clustering applied to the dataset using the variables of that interval) is the approach taken to preprocess the HSI dataset before applying categorical boosting.

TABLE I. INDIAN PINES DATA DISTRIBUTION

Class ID	Class Name	Number of Pixels
1	Alfalfa	54
2	Corn-notill	1434
3	Corn-mintill	834
4	Corn	234
5	Grass-pasture	497
6	Grass-trees	747
7	Grass-pasture-mowed	26
8	Hay-windrowed	489
9	Oats	20
10	Soybean-notill	968
11	Soybean-mintill	2468
12	Soybean-clean	614
13	Wheat	212
14	Woods	1294
15	Build.-Grass-Trees-Drv.	380
16	Stone-Steel-Towers	95

TABLE II. SALINAS DATA DISTRIBUTION

Class ID	Class Name	Number of Pixels
1	Brocoli (green 1)	2009
2	Brocoli (green 2)	3726
3	Fallow	1976
4	Fallow (rough plow)	1394
5	Fallow (smooth)	2678
6	Stubble	3959
7	Celery	3579
8	Grapes (untrained)	11271
9	Soil (vinyard develop)	6203
10	Corn (senesced green weeds)	3278
11	Lettuce (romaine 4wk)	1068
12	Lettuce (romaine 5wk)	1927
13	Lettuce (romaine 6wk)	916
14	Lettuce (romaine 7wk)	1070
15	Vinyard (untrained)	7268
16	Vinyard (vertical)	1807

IV. EXPERIMENTAL RESULTS

Instead of working with data-specific intervals that may require higher degrees of domain expertise and to demonstrate the general applicability of the algorithm, the procedures outlined in Figure 3 applies a shifting window with a stride amount of *s*-bands and uses the bands in each interval as input and clusters the dataset. Thus, each stride produces a new categorical feature, which are then all stacked to augment the original dataset.

As shown in Table III, for each window, a single categorical variable with K categories is learned by K-means and these are stacked horizontally to create a dataset with categorical features to train and test the boosting model. The window length, w, and the stride amount, s, determines number of categorical features (the number of window positions on the spectrum) created. Also, the results report the average accuracy for 10 different trials with the respective standard deviation using the Salinas dataset (we obtained comparable results with the Indian Pines dataset as well).

TABLE III. SENSITIVITY OF THE HYPERPARAMETERS FOR SALINAS DATASET

	Stride = 5 Window length = 10	Stride = 10 Window length = 10	Stride = 5 Window length = 20	Stride = 10 Window length = 20
K = 10	86.93 ± 0.32	85.39 ± 0.21	86.74 ± 0.48	86.06 ± 0.58
K = 20	88.06 ± 0.24	86.70 ± 0.16	88.73 ± 0.17	87.89 ± 0.28
K = 30	88.36 ± 0.13	86.95 ± 0.15	89.13 ± 0.13	88.5 ± 0.19

As the categorical dataset has fewer features, it is more suitable for feature selection and classification using boosting, which is the goal this work is trying to achieve. In the pixel classification task of 16 target classes (various vegetations etc.) on the test set, accuracies of the proposed boosting method yields favorable results. For comparison with the use of the original raw variables (i.e., 204 bands), various popular benchmark classifiers [5][6] are also used. The results on the Indian Pines and the Salinas datasets are reported in Table IV.

TABLE IV. COMPARISONS ON HSI CLASSIFICATION RESULTS

	Indian Pines	Salinas
K-NN (K=1)	63.49 ± 0.28	86.61 ± 0.11
K-NN (K=3)	65.87 ± 0.40	87.30 ± 0.14
K-NN (K=5)	66.81 ± 0.32	87.40 ± 0.14
Boosting on the original features	71.88 ± 0.55	88.38 ± 0.12
Proposed K-means + Boosting	71.89 ± 0.60	89.13 ± 0.13

```
Algorithm Clustered-Shifting-Window Boosting

Input
   X[N,D]: Train-set of N samples and D features
   y[N]: Class-labels of the N samples
   X_test[M,D]: Test-set of M samples and D features
   y_test[M]: Class-labels of the M test samples
   w: Window length
   s: Stride of windows
   K: Number of clusters in each window
Output
   Model: Trained Boosting Classifier (and centroids used)
   Test_Accuracy: Test accuracy of the trained classifier Model

Num_Windows = 0
for win_start from 1 to D in steps of s
   win_end = win_start + w
   windowed_data = X[:, win_start:win_end]
   centers = Kmeans(windowed_data, K)
   train_centers = Find_Nearest_Center(X, centers)
   test_centers = Find_Nearest_Center(X_test, centers)
   Categorical_Trainset[:,Num_Windows] = train_centers
   Categorical_Testset[:, Num_Windows] = test_centers
   Num_Windows = Num_Windows + 1
end for

Model = Train_Boost_Classifier(Categorical_Trainset, y)
Test_Accuracy = Test_Classifier(Model, Categorical_Testset, y_test)
```

Figure 4. The proposed Clustered-Shifting-Window Boosting Algorithm for a hyperspectral image dataset

V. CONCLUSION

We proposed a method for hyperspectral image pixel classification using cluster-ensemble-based categorical feature extractor and a categorical boosting classifier using those features. Each run of clustering splits the dataset of pixels into unsupervised categories of pixels. The resulting cluster indices of each run are used as categorical features by a categorical-boosting classifier. As the hyperspectral datasets come with several hundreds of features corresponding to a sequence of narrow wavelengths, we applied a sliding window to create diverse set of clustering runs. The experimental results showed that the proposed method improves the classification accuracy.

ACKNOWLEDGMENT

This work performed as part of an MS thesis [9] was supported by NSF under Grant No. 2003740.

REFERENCES

[1] Samat, A., Li, E., Du, P., Liu, S. & Xia, J. (2021) GPU-Accelerated CatBoost-Forest for Hyperspectral Image Classification Via Parallelized mRMR Ensemble Subspace Feature Selection. *IEEE J. of S. Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3200-3214.

[2] Grana, M., Veganzons, M., & Ayerdi, B. Hyperspectral remote sensing. https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes (Accessed on 9/24/2022).

[3] A Modular MultiLayer Framework for Real-Time Hyperspectral Image Segmentation. https://deepp0925.github.io/project_website/ (Accessed on 02/24/2023).

[4] Alpaydin, E. (2014). *Introduction to machine learning, third edition*. The MIT Press, Cambridge.

[5] Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181.

[6] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python, *JMLR,* 12, pp. 2825-2830.

[7] Ostroumova, L., Gusev, G., Vorobev, A., Dorogush, A.V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. NeurIPS.

[8] Sellami, A., Farah, M., Riadh Farah, I., & Solaiman, B. (2019). Hyperspectral imagery classification based on semisupervised 3-d deep neural network and adaptive band selection. *Expert Systems with Applications*, 129:246-259.

[9] Bellio, G. (2022) Boosting with Original and Clustered Categorical Features for Machine Learning on Large Datasets. Auburn University at Montgomery, AL, USA.