Yapay biçimbilimsel özniteliklerin hiperspektral görüntü sınıflandırma başarımının üzerindeki etkisi On the effect of synthetic morphological feature vectors on hyperspectral image classification performance

Amir Abbas Davari*, Erchan Aptoula†, Berrin Yanikoglu*,
*Department of Computer Science and Engineering, Sabanci University, Istanbul, Turkey
adavari@sabanciuniv.edu, berrin@sabanciuniv.edu
†Department of Computer Engineering, Okan University, Istanbul, Turkey
erchan.aptoula@okan.edu.tr

***Özetçe* —Bu çalışmada yapay olarak üretilmiş özniteliklerin hiperspektral uzaktan algılanmış görüntülerin sınıflandırma başarımına olan etkisi incelenmiştir. Öznitelik olarak, bu alanda kendini ispatlamış olan biçimbilimsel öznitelik profillerinden yararlanılmıştır. Yapay öznitelik üretimi için, çalışmamızın bu erken aşamasında, görece basit olan Bootstrapping algoritması seçilmiştir. Farklı hiperspektral veri kümeleri ile yürütülen deneyler sonucunda, yapay öznitelik üretiminin, kısıtlı eğitim verileri durumunda sınıflandırma başarımına önemli oranda katkı sağladığı gözlemlenmiştir**

Anahtar Kelimeler—uzaktan algılama, hiperspektral, biçimbilimsel öznitelik profilleri, matematiksel biçimbilim, sınıflandırma.

***Abstract*—This paper studies the effect of synthetic feature vectors on the classification performance of hyperspectral remote sensing images. As feature vectors, it has been chosen to employ morphological attribute profiles, that have proven themselves in this field. At this early stage of our work, the relatively simple Bootstrapping algorithm has been used for synthetic feature vector generation. Based on experiments conducted on multiple hyperspectral datasets, it has been observed that synthetic feature vectors contribute considerably to classification performance in the case of limited training dataset sizes.**

Keywords—remote sensing, hyperspectral image, extended morphological attribute profile, bootstrap, resampling, classification.

I. Introduction

Remote sensing is nowadays of paramount importance for several application fields, including environmental monitoring, urban planning, ecosystem-oriented natural resources management, urban change detection and agricultural region monitoring [1]. The majority of the aforementioned monitoring and detection applications require at some stage a label map of the remotely sensed images, where individual pixels are marked as members of specific classes, e.g. water, asphalt, grass, etc. In other words, classification is a crucial step for several remote sensing applications.

It is widely acknowledged that exploiting both the spectral as well as spatial properties of pixels, improves classification performance with respect to using only spectral based features [2]. In this regard, morphological profiles (MP) are one of the popular and powerful image analysis techniques that enable us to compute such spectral-spatial pixel descriptions. They have been studied extensively in the last decade and their effectiveness has been validated repeatedly [3]–[5].

The characterization of spatial information obtained by the application of a MP is particularly suitable for representing the multiscale variations of image structures, but they are limited by the shape of the structuring elements. To avoid this limitation, morphological attribute profiles have been developed [6]. By operating directly on connected components instead of pixels, not only they enable us to employ arbitrary region descriptors (e.g. shape, color, texture, etc) but they pave the way for object based image analysis as well. In addition they can also be implemented efficiently.

In this paper, we employ morphological attribute profiles for content description and focus specifically on the case of imbalanced and/or limited training datasets. As remote sensing images possess most often heterogeneous content, the classes therein are almost always represented unequally, e.g. a lot of roofs, very few shadows in an urban area, or a lot of fields and very few trees, etc. Moreover, the small ratio between the number of available training samples and the number of features makes it impossible to obtain reasonable estimates of the class-conditional hyper-dimensional probability density functions used in standard statistical classifiers. Consequently, on increasing the number of features given as input to the classifier beyond a certain threshold (which depends on the number of training samples and the kind of classifier adopted), the classification accuracy decreases [7]; this behavior is known as the Hughes phenomenon [8]. We have chosen to explore synthetic feature vector generation along with mor-

This work was supported by the TUBITAK Grant 112E210.

phological attribute profiles in order to remedy these issues. In particular, we have investigated a popular synthetic data generation method from the machine learning state of the art, namely resampling [9]–[11].

This paper is organized as follows; the explored approach is detailed in Section II. Next, Section III presents the conducted experiments and their results. Finally in Section IV the paper concludes with a discussion of the obtained results.

II. Approach

In this work we used extended attribute profiles (EAP) [13] as feature vectors for every pixel. As our datasets are hyperspectral, i.e. consist of hundreds of spectral bands, in order to prevent the EAP dimension from being very high and impractical, the number of image bands have been reduced by means of principle component analysis to a number which contains more than 99% of the dataset energy. After computing the feature vectors, we have added synthetic data to our feature space and we took advantage of resampling by the bootstrapping algorithm for its generation. The entire workflow is depicted in Figure 1.

Having the feature vector computed, we generated equal number of synthetic data, n, per each class using bootstrapping and added them to the original feature vector. The result was used as the new data set and training and test sets were extracted from it as is going to be described in section III-C. Feeding the aforementioned feature vector to the classifier, the classification performance was measured for different numbers of n values.

For classification purpose, random forest classifier was exploited. Random forest is basically an ensemble of decision trees. Unlike single decision trees which are likely to suffer from high Variance or high bias, random forests use ensemble methods to find a natural balance between the two extremes. In this work we used Bagging for ensembling which is based on averaging method. In averaging methods, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced [12].

III. Experiments

A. Data Sets

Pavia University, has been acquired by the ROSIS sensor in 115 spectral bands during a flight campaign over Pavia, northern Italy. 12 of these bands were removed due to noise and therefore 103 bands were used in this work. The scene has a size of 610 times 340 pixels with geometrical resolution of 1.3 m. Table I shows the classes and number of pixels per each class in Pavia University scene data set. We used the first four principle components of this dataset which had 99.16% of the total variance.

The Salinas Valley scene was collected by the 224-band AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution (3.7 m). In the corrected version 20 channels were removed and the experiments were conducted on a 204 channel HS image. The area covered comprises of 512 lines by 217 samples. It includes vegetables, bare soils, and vineyard fields. Table II shows its class based information. We used the first four principal components of this dataset which possessed 99.68% of the total variance.

Table I: Pavia University scene data set; class based information

#	Class	Samples
1	Asphalt	6631
2	Meadows	18649
3	Gravel	2099
4	Trees	3064
5	Painted metal sheets	1345
6	Bare Soil	5029
7	Bitumen	1330
8	Self-Blocking Bricks	3682
9	Shadows	947

Table II: Salinas scene data set; class based information

#	Class	Samples
1	Brocli_green_weeds_1	2009
2	Brocli_green_weeds_2	3726
3	Fallow	1976
4	Fallow_rough_plow	1394
5	Fallow_smooth	2678
6	Strubble	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_trellis	1807

B. Feature Extraction

The feature vector in this work was extended multi attribute profile (EMAP) with four attributes and four thresholds [13]. In detail:

- Area EAP: Area Extended Attribute Profile; the area threshold values for this attribute were tuned to be 100, 500, 1000, 5000.
- SD EAP: Standard Deviation Extended Attribute Profile; its threshold values were tuned to be 20, 30, 40 and 50.
- Hu First Moment EAP: The attribute threshold values were tuned to be 0.2, 0.3, 0.4, and 0.5.
- BBD EAP: Bounding Box Diagonal Extended Attribute Profile; its attribute threshold values were tuned to be 10, 25, 50 and 100.

For calculating the EMAPs, a hierarchical image representation named Max-tree [6] was employed and for filtering the constructed Max-tree based on each λ value, Max-filtering strategy [6] was used.

C. Classification

For classification a random forest classifier was used. Number of trees was chosen to be 100 and number of features

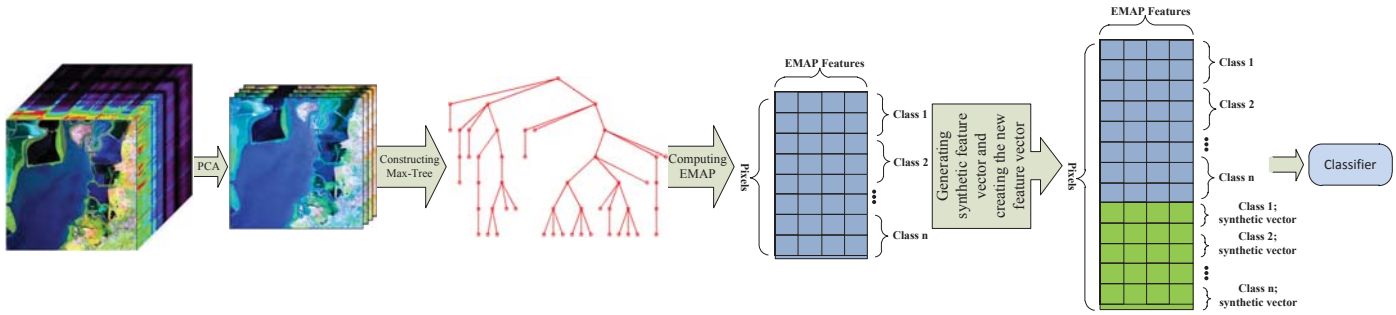

Figure 1: Our approach workflow.

to be used for training each tree was $\sqrt{\text{number of features}}$ by default for bagging. For training we randomly selected 20 pixels per class from the image as the training set and used them for the classification and all the ground truth pixels were used for testing. For each experiment, this procedure was repeated 25 times and Overall Accuracy (OA%), Average Accuracy (AA%) and Kappa averages and standard deviations of them over these 25 repetitions were calculated and reported as the classification performance. As a recall:

- OA: The overall accuracy is the number of correctly classified instances divided by the total number of data points (pixels).
- AA: The average accuracy is the average of class-based accuracies.
- Kappa: The kappa statistic is a measure of how closely the instances classified by the classifier matched the ground truth. By measuring the expected accuracy, it basically gives a statistic for the accuracy of a random classifier.

D. Experiments and Results

As it was mentioned in section II, we used data augmentation via the bootstrapping algorithm in order to compensate the class-based insufficiency of training data instances by means of adding synthetic data to our feature space. Table III and Figure 2 show the classification performance for different number of synthetic feature vectors added to Pavia University scene data set. Similarly, Table IV and Figure 3 show the classification performance for different number of synthetic feature vectors added to Salinas scene data set.

IV. Conclusion

Having a look at the results for these two data sets, it can be understood that employing synthetic feature vector has a positive effect on HS image classification performance. Although we used a simple generation method without taking the statistics of data into account, for Pavia University data set we obtained around 1.5% and 2% performance improvement for OA% and Kappa respectively. For the Salinas data set, around 0.25% improvement was achieved for OA% and Kappa as well.

Furthermore, based on the results, adding synthetic feature vectors not only improves the accuracy and Kappa statistics,

Table III: Pavia University scene; classification performance for different number of synthetic data

# synth data	AA%	AA std	OA%	OA std	Kappa	Kappa std
0	94.24	0.56	92.46	2.98	0.9023	0.0372
20	94.57	1.10	92.47	3.13	0.9025	0.0392
40	94.25	1.30	92.78	1.65	0.9059	0.0211
60	94.69	0.91	93.43	2.18	0.9143	0.0273
80	94.82	1.21	93.35	2.92	0.9135	0.0365
100	94.85	1.08	93.81	2.18	0.9192	0.0275
125	**94.98**	0.84	**93.94**	1.90	**0.9208**	0.0242
250	94.49	1.10	93.75	1.70	0.9182	0.0218
500	94.04	1.31	92.41	2.53	0.9014	0.0321
1000	93.76	0.86	92.52	1.98	0.9024	0.0251
3000	92.12	1.63	91.63	2.51	0.8905	0.0320

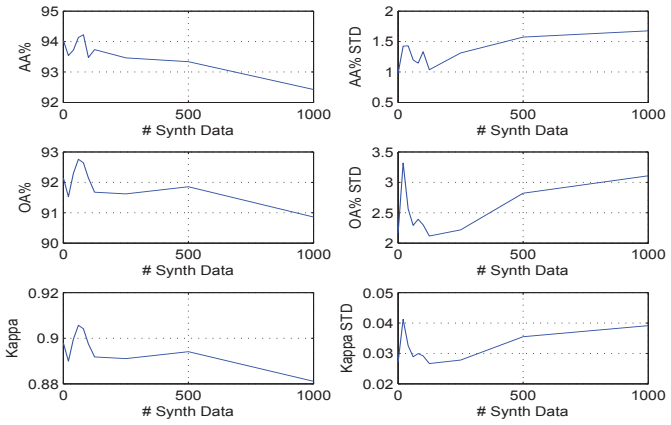

Figure 2: Pavia University scene; classification performance with different amount of added synthetic data.

Table IV: Salinas; classification performance for different number of synthetic data

# synth data	AA%	AA std	OA%	OA std	Kappa	Kappa std
0	95.61	0.47	91.71	1.05	0.9079	0.0117
20	95.39	0.73	91.03	1.47	0.9005	0.0162
40	95.50	0.65	91.37	1.42	0.9042	0.0157
60	95.53	0.60	91.41	1.49	0.9046	0.0165
80	95.40	0.81	91.37	1.61	0.9042	0.0178
100	**95.65**	0.60	**91.94**	1.18	**0.9105**	0.0131
125	95.41	0.51	91.24	1.40	0.9027	0.0154
250	95.25	0.67	91.04	1.43	0.9004	0.0159
500	95.19	0.57	91.15	1.66	0.9017	0.0183
1000	94.76	0.63	90.67	1.08	0.8963	0.0120
3000	93.60	0.76	89.21	1.13	0.8801	0.0125

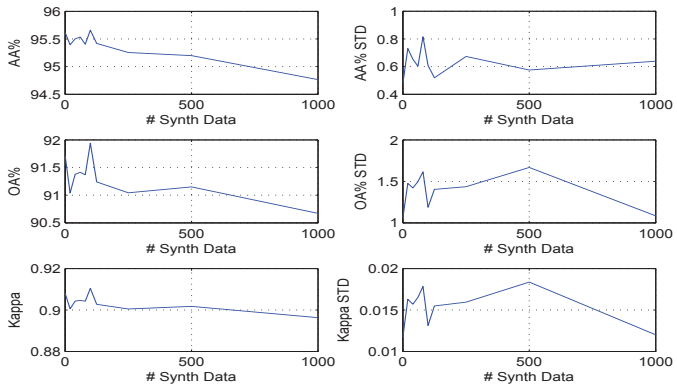

Figure 3: Salinas scene; classification performance with different amount of added synthetic data.

but also it whether decreases the standard deviation of the results or keep it low while the accuracy and Kappa is increased. Having a look at Figure 2 it is obvious that the minimum standard deviation is obtained when the optimum number of synthetic feature vectors was added, i.e. maximum performance was achieved. On the other hand for Salinas valley dataset, Figure 3 shows that the standard deviation is kept almost in the same level as the case of not using synthetic feature vector while the accuracy and Kappa is improved. Figure 4 and Figure 5 visualizes the aforementioned remarks.

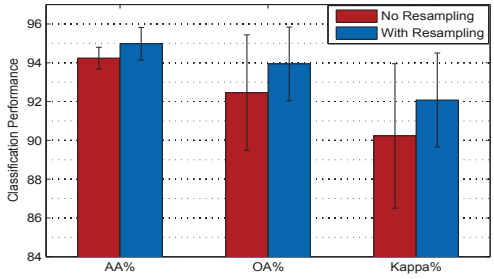

Figure 4: Effectiveness of using synthetic data using resampling for Pavia University scene data set. In this chart the best performance of resampling in III was exploited.

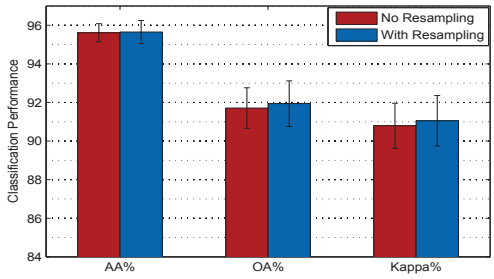

Figure 5: Effectiveness of using synthetic data using resampling for Salinas scene data set. In this chart the best performance of resampling in IV was exploited.

V. Future Works

In the context of synthetic data generation, there are three main different aspects that can affect the classification performance positively. First, is the algorithm for generating synthetic data. The wiser choosing this algorithm, the more the synthetic data will be informative and will enrich the main data. Second, is balancing the classes via adding different number of synthesized data and the last one is determining the optimal number of synthetic data is an important parameter which plays an important role to get the most out of synthetic data generation approach.

As it was noted, in this work we used a simple generation method to study the usability of synthetic data in this field. We are planning to study the aforementioned three aspects and compare their effectiveness in HS RSI supervised classification in our future works.

References

[1] S. Valero, P. Salembier, and J. Chanussot, "Hyperspectral image representation and processing with binary partition trees," *Image Processing, IEEE Transactions on*, vol. 22, no. 4, pp. 1430–1443, 2013.

[2] D. A. Landgrebe, *Signal theory methods in multispectral remote sensing*, vol. 29. John Wiley & Sons, 2005.

[3] M. Fauvel, J. A. Benediktsson, J. Chanussot, and J. R. Sveinsson, "Spectral and spatial classification of hyperspectral data using svms and morphological profiles," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 46, no. 11, pp. 3804–3814, 2008.

[4] J. A. Benediktsson, M. Pesaresi, and K. Amason, "Classification and feature extraction for remote sensing images from urban areas based on morphological transformations," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 41, no. 9, pp. 1940–1949, 2003.

[5] K. Tan, E. Li, Q. Du, and P. Du, "Hyperspectral image classification using band selection and morphological profiles," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 1, pp. 40–48, 2014.

[6] M. Dalla Mura, J. A. Benediktsson, B. Waske, and L. Bruzzone, "Morphological attribute profiles for the analysis of very high resolution images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 10, pp. 3747–3762, 2010.

[7] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 8, pp. 1778–1790, 2004.

[8] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *Information Theory, IEEE Transactions on*, vol. 14, no. 1, pp. 55–63, 1968.

[9] T. M. Padmaja, N. Dhulipalla, R. S. Bapi, and P. Radha Krishna, "Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection," in *Advanced Computing and Communications, 2007. ADCOM 2007. International Conference on*, pp. 511–516, IEEE, 2007.

[10] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18–36, 2004.

[11] S. Kotsiantis, D. Kanellopoulos, P. Pintelas, *et al.*, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[13] M. Dalla Mura, J. Atli Benediktsson, B. Waske, and L. Bruzzone, "Extended profiles with morphological attribute filters for the analysis of hyperspectral data," *International Journal of Remote Sensing*, vol. 31, no. 22, pp. 5975–5991, 2010.