Kernel Rotation Forest for Hyperspectral Image Classification

Jeremy Kim

jihyungkim@knights.ucf.edu

***Abstract*—Classification of hyperspectral imagery has utility in land cover mapping. Kernel Rotation Forest (KRF) using Gaussian kernel Principal Component Analysis (GkPCA) can resolve non-separable cases and be a suitable small-sized ensemble method when there is high within-label spectral signature heterogeneity.**

Introduction

Hyperspectral imagery amassed from airborne and space-based platforms are three-dimensional spectral stacks of the same scene. Its remotely-sensed spectral and spatial resolution is appropriate for land cover mapping [1]. The supervised task of identifying numerous ground categories is compounded however, by their typically limited training records, relative to the dimension of the input space (i.e. hundreds of spectral bands) [2]. The ensuing Hughes phenomenon—the curse of dimensionality—confers tremendous mastering of *sparse* data that cannot generalize. And in spite of comprising massive quantities of information (e.g. 145 x 145 pixels and 220 spectral bands work out to roughly 4.6 million data points), the hyperspectral cubes' overlapping layers can be highly correlated and noncontributory. To address hyperspectral data sets being quite multitudinous and not valuable in their entirety, having many computationally-efficient learners probing for relevance can be advantageous [1].

The machine learning community has largely abandoned pursuit of a universally-capable model and have instead pledged to hetero- and homogeneous aggregation paradigms. Ensembles, multiple classifier systems (MCS), are many predictors funneling their predictions into a consensus mechanism that decides final outputs. Ensemble size can be a gauge of operating complexity and regarded as an ensemble hyperparameter. When the number of predictors within an ensemble approaches the order of thousands, distinctions between ensembling methods are muted. Thus the community strives for a consistently good, *small* ensemble that requires low (fixed complexity) training effort and delivers fast, near-optimal performance [3]. Two indicators of successful ensembles are accurate members with obvious between-member diversity (e.g. if all researchers specialized in the same narrow field, they would be unlikely to offer one another expansive insight) [2]. The trade-off between accuracy and diversity, known as the accuracy-diversity dilemma, was addressed by the hopeful inventors of Rotation Forest (RF) that documented its achievement in smaller ensembles [4].

First we will describe the accuracy and diversity mechanisms in well-regarded ensemble methods. In bagging, each classifier independently trains on a bootstrap sample, a smaller sample that is "bootstrapped" from a larger sample. The distribution encountered during learning echoes the original distribution and thus base classifiers can inference unseen data. But, the slim provenance

of inter-classifier dissimilarity is the relative object frequencies between samples. Bagging's resampling produces insufficiently diverse ensembles and expects large membership in order to contend [3].

Random Forest extends bagging and generates diversity through random subspaces, randomizing the subset of features to consider during node-splitting. Its base classifier, decision tree (DT), decides the most information-gaining feature to branch on at each stage of growth [4]. Limiting the candidate pool of variables decorrelates resultant trees and mitigates computational complexity [2]. Random subspaces is an effective diversifying heuristic that does not compromise accuracy [3].

AdaBoost's performance has been attributed to its diversity-generating pipeline. AdaBoost orchestrates an inaccurate but heterogeneous network, directing the sequential focus of each learner on previously misclassified observations. Its individually weak members pick up one another's slack and collectively outperform bagging ensembles. And this concludes our summary of previous ensemble methods' handling of the accuracy-diversity dilemma [3].

RF ambitiously generates accuracy and diversity using principal component analysis (PCA), also known as Karhunen-Loeve transformation, to explore distinct feature spaces. The forest cultivates DT that are receptive to its rotation heuristic that applies feature extraction to subsets of features and reconstructs a full feature set. The randomized feature subsets activate different realms, as even when the same feature subsets are proposed, constrained sample size avoids coefficient repetitions. Finally, accuracy is sought by learning on the entire data [2], [3]. Although feature interpretability is sacrificed, RF attains accuracy and diversity with minor information loss [4].

Previous literature reasoned PCA as being inadequate for feature extraction as it compresses dimension (i.e. the most discriminatory components might correspond to small variance and be discarded). However, all principal components are kept during RF ensembling to function as transformation instruction and thus discriminatory information is preserved. However, having all components and being able to observe the scatter of the data in the extracted feature space does not necessarily expedite classification [3].

KRF augments RF's feature-extraction-based promotion of diversity and accuracy by introducing kernel PCA to extract nonlinear features. While linear PCA seeks directions of maximum variance in the dimension of the input space, kernel PCA searches in the feature space. Nonlinear feature projection renders classes much more separable, thereby facilitating drawing of decision boundaries. The authors of KRF favored Gaussian kernel, and as such we proceed with GkPCA [4].

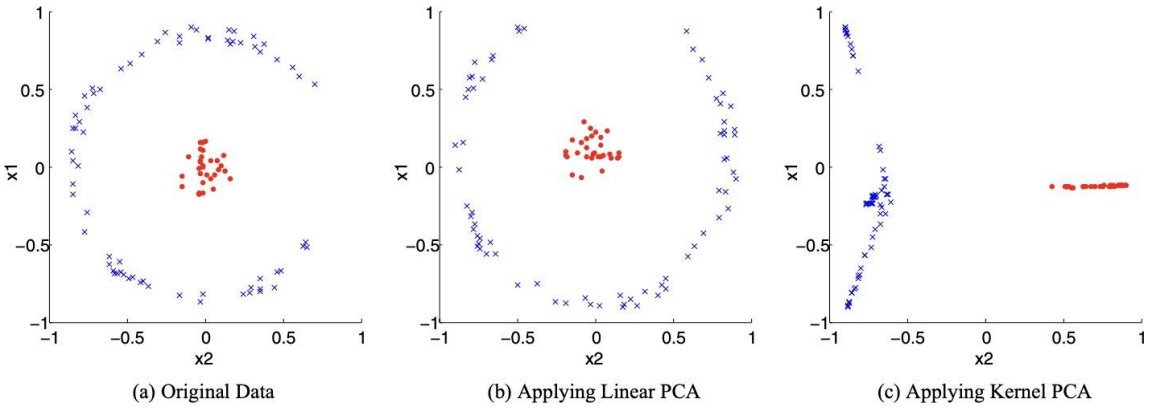
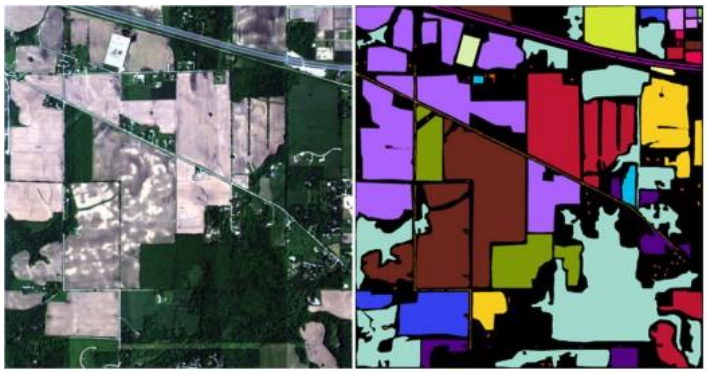


Figure 1. Kernel PCA can yield outstanding class separability over linear PCA.

We selected the Indian Pine dataset to investigate kernel PCA transformation because linear PCA's suboptimal performance on that particular scene was recorded [2]. Indian Pine, the Purdue University Agronomy farm northwest of West Lafayette and the surrounding area, was imaged by the NASA Airborne Visible / Infrared Imaging Spectrometer (AVIRIS) sensor on June 12, 1992 to support soil research. The Indian and Pine Creek watersheds are pictured, thus it is commonly referred to as the Indian Pine dataset [5]. As 20 water absorption bands were removed, there remain 200 spectral bands for analysis. The 16 land use labels are reported at field magnitude, but heterogeneous intra-label signatures (e.g. corn and soybean crops were planted shortly prior to scene acquisition and soil intersects with vegetation references) provide an opportunity to evaluate intra- and inter-label variability at medium spatial resolution.



ID	Class	#Labeled Samples
	C1: Corn-high residue	47,168
	C2: Corn-mid residue	183,530
	C3: Corn-low residue	356
	C4: Soybean-high residue	226,130
	C5: Soybean-mid residue	120,400
	C6: Soybean-low residue	29,210
	C7: Other residues	5,795
	C8: Wheat	3,387
	C9: Hay	63,135
	C10: Grass/Pasture	6,512
	C11: Grass	26,853
	C12: Wood-uniform	64,947
	C13: Wood-rugged	288,500
	C14: Highway	10,637
	C15: Local road	6,570
	C16: Power station	6,929
	C17: Power towers	411
	C18: Houses/Buildings	2,128
	C19: Urban areas	1,532

Figure 2. Indian Pine scene (left) and ground truth map and class legend (right).

Method

Given dataset $\{x_i\}_{i=1}^{N}$ and feature map ϕ, we assume $\Sigma_{i=1}^{N} x_i = 0$. The sample covariance matrix in the feature space is $C = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i)\phi(x_i)^T$. Kernel PCA solves the eigenvalue problem $Cu_j = \lambda_j u_j,\ \ j = 1, \ldots, N$ where λ_j is an eigenvalue of C and u_j is the corresponding eigenvector. The eigenvalue problem can be written as $K\alpha_j = \lambda_j N\alpha_j,\ j = 1, \ldots, N$ where $K_{mn} = k(x_m, x_n)$, k is a kernel function and α_j satisfies $u_j = \Sigma_{i=1}^{N} \alpha_{ji}\phi(x_i)$. Once the eigenvalue problem is solved, each eigenvector α_j is normalized to satisfy the condition $u_j^T u_j = \alpha_j^T K\alpha_j = \lambda_j N\alpha_j^T\alpha_j = 1,\ j = 1, \ldots, N$. Each eigenvector α_j corresponds to the j-th principal component. The j-th principal component score of a test instance x_t, $s_j(x_t)$ is resolved by projections onto the eigenvectors,

$$s_j(x_t) = u_j^T\phi(x_t) = \sum_{i=1}^{N} \alpha_{ji} k(x_i, x_t).$$

Given a training dataset $D = \{x_i, y_i\}_{i=1}^{N}$, where x_i is an input vector and $y_i \in \{0, \ldots, c\}$ is the corresponding label, we train L classifiers independently. The i-th classifier C_i is trained by randomly splitting the set of features F into K feature subsets from $F_{i1}, \ldots, F_{iK}$. For each feature subset F_{ij}, a submatrix X_{ij}^{S} is formed from a bootstrap sample X^S of X. Kernel PCA is applied to form the rotated submatrix X_{ij}^{ROT}. Classifier C_i is trained on $R_i = [X_{i1}^{ROT}, \ldots, X_{iK}^{ROT}]$ and its label vector Y, until all classifiers $C_1, \ldots, C_L$ are built. Pseudocode is outlined in Algorithm 1 [4].

Algorithm 1 Kernel Rotation Forest

Input: X ($N \times p$ hyperspectral instances), Y ($N \times 1$ corresponding labels), F (set of p features), L (ensemble size), K (number of feature subsets)
Output: L classifiers $C_1, \ldots, C_L$

for $i = 1$ to L:
  $\{F_{ij}\}_{j=1}^{K} \leftarrow K$ disjoint random subsets from F
  for $j = 1$ to K:
    $X^S \leftarrow$ bootstrap sample drawn from X
    $X_{ij}^{S} \leftarrow$ submatrix of X^S consisting of features in F_{ij}
    $X_{ij}^{ROT} \leftarrow$ rotation matrix of X_{ij}^{S} from kernel PCA extracted components
  $R_i \leftarrow [X_{i1}^{ROT}, \ldots, X_{iK}^{ROT}]$
  $C_i \leftarrow i$-th classifier trained on (R_i, Y)

DT were unpruned for maximum interpolation. Entropy, $-\sum_{i=1}^{n} p(c_i) log(p(c_i))$, was the impurity criterion for sensitive information gain measurements. The number of features within a subset was fixed to 4 to evenly divide the 16 labels and create disparate subsets. Gaussian, or radial basis function, was the kernel. Because kernel PCA provides a far greater number of nonlinear features than the input dimensionality, we retained only the first M_{ij} features, where M_{ij} is the dimensionality of X_{ij}. Classification was carried out through majority voting which returns the prediction mode for each test instance. Overall accuracy was the evaluative metric.

Results

The 5-fold cross-validation accuracies for GkPCA KRF and PCA RF for ensemble sizes of 10, 50, and 100 are shown. As a reminder, a lucky guess for 1 in 16 would be 6.25.

	$L = 10$	$L = 50$	$L = 100$
PCA	8.67 ±0.24	11.27 ±0.28	12.37 ±0.09
GkPCA	23.66 ±0.33	25.59 ±0.39	25.89 ±0.27

Discussion

GkPCA outperformed PCA by a factor of 2.73x for an ensemble size of 10. However, this leading performance was diminished to 2.09x for an ensemble size of 100. The assertion that differences in methodologies tapers off for larger ensemble sizes is exhibited in our trial. We thus recommend GkPCA for efficient learning outcomes in smaller ensembles.

One of GkPCA's shortcomings is its high computational burden. Our local machine, an M1 Max with 10-core CPU and 64 GB memory, experienced significant degradation when attempting to apply kernel PCA on larger samples. Our implementation worked around this hurdle by feeding small windows of resolutions not to exceed 100 x 100. GkPCA might then not be a viable option for bigger ensembles.

A final mention is that KRF is similar to RF in that it is ultimately constructing an uninterpretable black box model with no recuperable knowledge (e.g. Random Forest has feature-ranking capability) [4]. In future experiments, devising synergizing concepts with KRF to help it regain human-interpretable intelligence would be invaluable.

References

[1] E. Pasolli, S. Prasad, M. M. Crawford, and J. C. Tilton, "Advances in Hyperspectral Image Classification Methods for Vegetation and Agricultural Cropland Studies," in *Hyperspectral Image Classification Methods and Approaches*, 2018, CRC Press.

[2] J. Xia, P. Du, X. He, and J. Chanussot, "Hyperspectral Remote Sensing Image Classification Based on Rotation Forest," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 239-243, January 2014. doi: 10.1109/LGRS.2013.2254108.

[3] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619-1630, October 2006, doi: 10.1109/TPAMI.2006.211.

[4] J. Shim, S. Kang, and S. Cho, "Kernel Rotation Forests for Classification," In Proc. IEEE International Conference on Big Data and Smart Computing (BigComp) '20, 2020, pp. 406-409, doi: 10.1109/BigComp48618.2020.00-40.

[5] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3," 2015, Purdue University Research Repository. doi:10.4231/R7RX991C