FUSION OF DIVERSE FEATURES AND KERNELS USING LP-NORM BASED MULTIPLE KERNEL LEARNING IN HYPERSPECTRAL IMAGE PROCESSING

Muhammad Aminul Islam, Derek T. Anderson, John E. Ball, Nicholas H. Younan

Mississippi State University
Department of Electrical and Computer Engineering
Mississippi State, MS 39762

ABSTRACT

Multiple kernel learning (MKL) is an elegant tool for heterogeneous fusion. In *support vector machine* (SVM) based classification, MK is a homogenization transform and it provides flexibility in searching for high-quality linearly separable solutions in the *reproducing kernel Hilbert space* (RKHS). However, performance often depends on input and kernel diversity. Herein, we explore a new way to extract diverse features from hyperspectral imagery using different proximity measures and band grouping. The output is fed to ℓ_p-norm MKL for feature-level fusion, where larger p's are preferred for diverse vs sparse solutions. Preliminary results on benchmark data indicates that ℓ_p-norm MKSVM of diverse features and kernels leads to noticeable performance gain.

Index Terms— Multiple kernel learning, feature fusion, hyperspectral image analysis, bandgrouping

1. INTRODUCTION

Hyperspectral imaging has a wide range of applications from mineralogy to weather forecasting, agriculture, surveillance, etc. A hyperspectral image can be described as a high dimensional data cube. Each sub-image in this data cube (usually on the order of hundreds) informs us about the radiance (or reflectance) properties of a scene at different narrowly spaced bands in the *electromagnetic* (EM) spectrum. However, automated analysis of some geographical area might require several remote sensing systems with sensors in different regions of the EM, e.g., visible, near IR and SWIR hyperspectral imagery, lidar and *synthetic aperture radar* (SAR). The point is, numerous sensors are often involved in remote sensing and new theory is needed to fuse them. Herein, we focus on the production and fusion of disparate features in hyperspectral imagery for robust classification. However, without loss of generality, the underlying approach discussed in this article is equally applicable to multi-sensor fusion.

The last decade has seen a surge of interest in the development and use of *multiple kernel learning* (MKL) for tasks like heterogeneous data fusion, classification and input/feature selection in areas like machine learning, pattern recognition, signal processing, computer vision, etc. A good recent review of MKL mathematics and algorithms is [1]. In the context of *support vector machine* (SVM) based classification, it is often the case that a single kernel is not enough. In practice, a challenge is finding a quality kernel. This is where MK helps. Instead of being restricted to a limited set of known kernels, which may not solve a task, MK provides a solid foundation to combine (fuse) a set of known base kernels (those satisfying Mercer's conditions) to produce a new and more powerful tailored kernel. MK is both a homogenization transformation for different input spaces and it ultimately provides important flexibility for classification in terms of searching for quality linearly separable solutions in the *reproducing kernel Hilbert space* (RKHS). There are number of outstanding MKL problems, for example: how do we generate diverse inputs for MKL; what linear or nonlinear aggregation operators are needed to combine the base kernels; how are multiple kernels normalized; how do we mitigate overfitting in MKL; and what search algorithms are needed for fusion algorithm parameter estimation. To date, numerous algorithms have been put forth, e.g., MKLGL [2], ℓ_p-norm MKL [3], FIGA [4], GAMKL$_p$ [5], DeFIMKL [5], etc.

In terms of hyperspectral image processing, MKL has been used for tasks like classification [6], feature selection [7] and nonlinear unmixing [8]. In [9], Zhang et al. presents a multi-sensor fusion technique, where ℓ_1-norm MKL is used to fuse several multi-scale RBF kernels applied to each sensor data set. Majority voting is used to aggregate the MKL classification results. The main contribution was the use of *active learning* (AL) for the selection of training samples based on maximum disagreement. In [7], simpleMKL [10] was used to help learn image features. Multiple RBF kernels are applied to a single feature, a group of features and features from heterogeneous sources. In [11], Honeine and Richard proved that the angular kernel is a valid reproducing kernel, and it was explored for hyperspectral image processing because spectral angle is a popular tool used in the literature due to its invariance to the spectral energy.

Gu et al. published a series of papers on MKL for classification in hyperspectral imagery [6, 12, 13]. These articles

are efficient algorithms to learn the optimum weights in ℓ_1-norm MKL. The weights of the kernels are obtained by using maximum variance kernel with minimum F-norm error [6], applying low rank *non-negative matrix factorization* (NMF) in [13], and regularizing the weights using cardinality based constraints in [14], which is the extension of [6] for sparse MKL. In [13], *kernel based NMF* (KNMF) uses the non-linear mapping of the base kernels. Specifically it uses the polynomial kernel to map the base RBF kernels to a higher dimensional RKHS. In all the papers, multi-scale RBF kernels based on Euclidean distance is used as the base kernels.

Kloft *et al.* provided an extensive analysis on the different variants of MKL [3]. They showed theoretically and analytically that ℓ_1-norm MKL has higher performance in noisy situations where the noisy kernels are eliminated via the sparse weights. On the other hand, higher norm MKL tends to give equal weights to all the kernels, and therefore outperforms the sparse MKL when the kernels are good and diverse. In many cases in hyperpsectral image processing, we can have diverse feature sets that can be used to generate quality kernels. This in turn signifies that dense MKL has a huge potential to improve the classification results over the sparse MKL, but it has not been explored in the hyperspectral community to date. Herein, we employed the ℓ_p-norm MKL and found similar results that support the analysis in [3].

In summary, while there has been interest in using MKL for hyperspectral image processing, work to date has primarily focused on the fusion of homogeneous kernels, e.g., multiple RBFs with Euclidean distance. However, it is likely that different kernels are required. In addition, to the best of our knowledge little-to-no work has focused on how to generate a diverse set of features for MKL via bandgrouping in hyperspectral image processing. We show that diversity with respect to both features and kernels is important for MKL as well as ℓ_p-norm MKL outperforms sparse MKL in aggregating them. Figure 1 is a high-level illustration of our approach.

2. METHODS

In this section, we describe the three major parts of our approach—(i) proximity measure calculation, (ii) band grouping for feature extraction, and (iii) feature space fusion using ℓ_p-norm MKL. For notational purposes, the 3D hyperspectral data cube is remapped into a 2D space such that each row represents a pixel and each column is a band. Let the re-shaped 2D data set be $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]^T \in \mathbb{R}^{n\times b}$, where n is the number of pixels in the image and b, the number of bands.

2.1. Proximity Measure Calculation

Hyperspectral sensors are wonderful because each pixel has a wealth of information and tells a story, versus traditional single channel or "RGB" imagery. However, hyperspectral imagery also suffers from the curse of dimensionality, spatially, spectral and sometimes temporal. Instead of using all bands, or features, it is often the case that selecting individual bands, feature projection (e.g., *Principle Component Analysis* (PCA), random projection, etc.), or grouping bands leads to a better solution (higher accuracy and more robust). While numerous approaches have been proposed, it has been shown that band group partitioning is of utility and it can be derived, in a supervised or unsupervised fashion, from a proximity measure [15]. For example, one can compute the correlation matrix between the different bands (unsupervised approach). Structure in this matrix can be used to derive a band group partitioning. However, there are a number of unsupervised proximity measures, such as Euclidean, correlation, Jeffrey K. Matusita, Bhattacharyya, *spectral angle mapper* (SAM), etc. In general, it has been demonstrated that selection of proximity measure depends in part on the data set and task, meaning there does not appear to be a global best. Herein, we explore the square of Euclidean, which measures the distance between a pair of vectors, and correlation, which measures angular similarity. We selected these two proximity measures for demonstration as they capture different aspect of the data via its features. However, in future work this is likely a parameter that needs to be included in our algorithm.

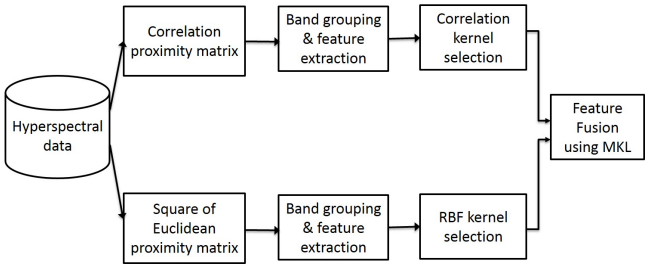

Fig. 1. High-level illustration of the proposed MKL approach.

Proximity Measure 1: Square of Euclidean The square of Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{L} (x_{ik} - x_{jk})^2.$$

Note, the square of Euclidean distance is always positive and depends in large on L, e.g., the length of the vectors or the number of pixels in the training data set.

Proximity Measure 2: Correlation Correlation is a similarity measure between two signatures. The Pearson's correlation coefficient is the co-variance of two vectors normalized by the product of the standard deviation of two distributions,

$$s(\mathbf{x}_i, \mathbf{x}_j) = corr(\mathbf{x}_i, \mathbf{x}_j) = \frac{cov(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j}}.$$

The correlation coefficient is in $[-1, 1]$. Distance, or the dissimilarity measure, $d(\mathbf{x}_i, \mathbf{x}_j)$ is obtained herein by simply subtracting the $s(\mathbf{x}_i, \mathbf{x}_j)$ from 1.

2.2. Feature Extraction

In this step, we first partition similar bands from a given proximity measure into groups and we then apply a feature extraction or reduction technique to each band group to extract a single feature from that group. Herein, we use the algorithm proposed by Ball *et al.* that performs unsupervised grouping of contiguous similar bands with respect to a provided proximity measure (see [15] for full algorithm details). After band grouping, we can apply a number of feature extraction techniques such as *stepwise linear discriminant analysis* (SLDA) [16], mean or weight, to get features equal to the number of band groups. In this paper, we calculate the mean of all bands in a groups as the feature. While mean might not be the most sophisticated technique, its advantage is that it is simple to realize in hardware and gives rise to a simpler multispectral versus hyperspectral sensor.

2.3. Feature Space Fusion Using ℓ_p-Norm MKL

In the kernel approach, inputs are ideally projected into a high, possibly infinite, dimensional RHKS space, where the patterns for different classes are now linearly separable. The trick is that we can do this all via a "kernel function" in the original low(er) dimensional space and we never have to do the actual lifting. However, in reality we do not know what kernel to select and in general the choice of kernel is task specific. There is currently no straightforward way to select a kernel for a given set of data. As already mentioned, MKL provides one such path to help search for the idea kernel by the simple concept of combining simple known (base) kernels to form custom (tailored) kernels. However, we must search for this kernel and the space is both extremely large and if we are not careful, MKL tends to succumb to overfitting (can learn the training data perfectly but not generalize well to new test data). Herein, we restrict our analysis to a *linear convex sum* (LCS) of kernels. This is by far the predominant MKL approach. While a few nonlinear approaches have been put forth, e.g., FIGA, for various reasons (such as proving that a given aggregation operator always yields a valid Mercer kernel) nonlinear MKL is still an unsolved problem.

For a function to be a kernel, it must satisfy the Mercer's kernel properties such as continuity, symmetry, and positive semi-definiteness. There are numerous kernel functions, e.g., *radial basis function* (RBF), polynomial, etc. In this paper, we use RBF and correlation kernel. The RBF function is

$$k_r(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{||\mathbf{x}_i^2 - \mathbf{x}_j^2||}{2\sigma^2}\right),$$

where σ is the so-called width parameter of the RBF kernel. The correlation kernel is

$$k_c(\mathbf{x}_i, \mathbf{x}_j) = exp\left(-\frac{1 - corr(\mathbf{x}_i, \mathbf{x}_j)}{2\sigma^2}\right)$$

where $corr(\mathbf{x}_i, \mathbf{x}_j)$ is the Pearson's correlation coefficient for $\mathbf{x}_i$ and $\mathbf{x}_j$. In [17], the authors have shown that the correlation kernel satisfies the Mercer's kernel properties. Note, our two kernels are already more-or-less to scale by design. However, if one is using heterogeneous kernels that produce very different scales, then the zero mean and unit variance RHKS approach in [3] can be used.

The convex sum of M kernels is also a Mercer's kernel. This is because both the sum and multiplication by positive constant are *positive semidefinite* (PSD) preserving operators (on M different Gram matrices). The combined kernel with ℓ_p-regularized weight w_m is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} w_m k_m(\mathbf{x}_i, \mathbf{x}_j)$$

subject to $||\mathbf{w}||_p \leq 1$ and $w_m \in \mathbb{R}^+$, where $||\mathbf{w}||_p$ is the p-norm of $\mathbf{w}$. Though the above expression is notationally for M kernels on the same set of features, it is trivially generalized to multiple features, e.g., different kernels on different subsets of features [5]. Optimization-based MKL solutions, versus fixed rule or heuristic approaches, optimize (using alternating optimization typically) the weights of the kernels and the SVM criteria function. Again, we use ℓ_p-norm MKL [6, 7] to derive the LCS weights. However, we could use a number of other search algorithms for feature level fusion, such as MKLGL$_p$ or GAMKL$_p$, or decision-level MKL, e.g., DeFIMKL$_p$. The ℓ_p-norm condition is more-or-less the same across a number of solvers. In general, the different approaches represent variations in search, e.g., Group Lasso (MKLGL), genetic algorithm based (GAMKLp), and nonlinear decision-level fusion via DeFIMKL.

3. PRELIMINARY RESULTS AND DISCUSSION

The Indian Pines data set consists of 145×145 pixels with a spatial resolution of 20 meters and 220 spectral channels (bands). During data pre-processing of the data, 20 water absorption bands, $104 - 108$, $150 - 163$ and 220 were removed. We considered the following 9 classes for classification — Corn-notill, Corn-min, Grass-pasture, Grass-trees, Hay-windowed, Soybean-notill, Soybean-mintill, Soybean-clean and woods. Random jack-knife partitioning is used, where 20% is training and the remainder is testing. Hereafter, the squared Euclidean is denoted as 'SqE' and correlation is 'Corr'. Proximity matrices are computed on the training data based on SqE and Corr. The number of band groups and thus features extracted was 11 for SqE and 16 for Corr (forming two feature vectors). The training data is standardized for each feature to have zero mean and unit variance and

Table 1. Producer's accuracies for ℓ_p-norm MKL based fusion.

ℓ_p-norm	Method (SqE = Square of Euclidean, Corr = Correlation)	# of kernels	corn (notill)	corn (min)	grass (pasture)	grass (trees)	hay (windowed)	soybeans (notill)	soybeans (min)	soybeans (clean)	woods
NA	SqE	1	64.52	46.18	75.57	92.80	97.19	63.82	78.57	28.31	97.78
	Corr	1	62.42	34.93	81.11	85.09	96.16	62.02	66.46	30.55	97.49
$p = 1.1$	Fusion of SqE & Corr	$1+1$	68.35	46.48	79.60	90.79	96.16	64.99	79.08	45.42	**97.87**
	SqE	2	65.91	52.62	87.41	92.29	97.44	64.34	80.09	37.27	97.78
	Corr	2	62.77	36.13	82.12	87.27	96.16	62.02	66.41	30.55	97.49
	Fusion of SqE & Corr	$2+2$	68.70	52.02	88.41	93.13	96.42	64.99	80.45	46.64	97.87
$p = 2$	Fusion of SqE & Corr	$1+1$	69.92	53.37	82.87	91.96	96.16	69.77	80.14	52.55	97.20
	SqE	2	69.14	57.12	88.66	92.80	97.70	69.51	81.81	46.64	97.78
	Corr	2	65.82	39.28	87.15	88.11	97.19	63.70	66.67	37.07	97.49
	Fusion of SqE & Corr	$2+2$	73.50	62.82	91.18	93.97	97.70	72.22	83.23	60.08	97.10
$p = 100$	Fusion of SqE & Corr	$1+1$	71.49	60.57	83.63	93.63	96.93	72.87	81.16	60.29	96.14
	SqE	2	72.97	64.32	89.67	94.47	97.70	73.00	83.13	55.80	97.29
	Corr	2	68.88	46.03	89.92	89.61	97.44	66.93	67.68	45.42	97.39
	Fusion of SqE & Corr	$2+2$	**77.24**	**69.42**	**92.44**	**94.97**	**97.95**	**76.49**	**85.11**	**66.40**	95.75

the testing data is standardized using the mean and standard deviation of the training data. For SqE and Corr, we used 10 kernels each with $\sigma = \{2^{-3}, 2^{-2}, \cdots, 2^{6}\}$. Top performing kernels for each feature were selected using SVM accuracy and fused using ℓ_p-norm MKL. 'SVMLight' and 'MKL' implementations in the Shogun toolbox [18] were used. For the ℓ_p-norm, we tried $p = 1.1$ (approximately city block distance), $p = 2$ (Euclidean) and $p = 100$.

Table 1 shows that inter-method fusion, i.e., fusion of SqE and Corr, is the top performer for all classes. Also, a larger p produces the best results for all classes except 'woods'. Inter-method fusion has an improvement of approximately 2% to 10% relative to intra-method for corn-notill, corn-min, grass-pasture, soybeans-notill, soybeans-min and soybeans-clean. It has almost the same performance for grass-tress and hay-windowed. The behavior of wood is different from all other classes. It has the best result at $p = 1.1$, and it continues to degrade with increasing p. Note, 'wood' is easily classifiable with a single kernel. $\ell_{1.1}$ MKL, which promotes sparse solutions, is more suitable for this task. In [3], Kloft et al. showed that when kernels are diverse, higher norm MKL is more appropriate and yields better results. In our case, results improve for 8 out of 9 classes as ℓ_p-norm increases, which supports our claim that diversity in features and kernels is useful for MKL-based hyperspectral classification.

4. FUTURE WORK

As stated above, our results are preliminary but promising—ran on a single well-known benchmark data set that was not trivial to solve using a single kernel. In future work, we will apply our technique to additional data sets. We will also investigate a search procedure for MKL parameter selection, including kernel type and associated parameters (a critical aspect of MKL that is typically overlooked due to complexity). Here, we explored, for demonstration, one kernel based on Euclidean distance and another based on angle for diversity. We believe it is also of interest to explore different, or a combination of different band group selection algorithms and what particular proximity measures are fed to these techniques to ultimately generate diverse features for MKL. Last, we are currently using all features produced by band grouping. However, sometimes some bands (or band groups) are not useful for a task at hand and performance can be raised if we do a feature selection step before fusion.

5. REFERENCES

[1] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[2] Z. Xu, R. Jin, H. Yang, I. King, and M. R. Lyu, "Simple and efficient multiple kernel learning by group lasso," in *Proc. Int. Conf. Machine Learning*, 2010, pp. 1175–1182.

[3] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "Lp-norm multiple kernel learning," *The Journal of Machine Learning Research*, vol. 12, pp. 953–997, 2011.

[4] L. Hu, D. T. Anderson, T. C. Havens, and J. M. Keller, "Efficient and scalable nonlinear multiple kernel aggregation using the choquet integral," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Anne Laurent, Olivier Strauss, Bernadette Bouchon-Meunier, and RonaldR. Yager, Eds. 2014, vol. 442 of *Communications in Computer and Information Science*, pp. 206–215, Springer.

[5] A. Pinar, T. C. Havens, D. T. Anderson, and L. Hu, "Feature and decision level fusion using multiple kernel learning and fuzzy integrals," in *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Aug 2015, pp. 1–7.

[6] Y. Gu, C. Wang, D. You, Y. Zhang, S. Wang, and Y. Zhang, "Representative multiple kernel learning for classification in hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 7, pp. 2852–2865, 2012.

[7] D. Tuia, G. Camps-Valls, G. Matasci, and M. Kanevski, "Learning relevant image features with multiple-kernel classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 10, pp. 3780–3791, 2010.

[8] J. Chen, C. Richard, and P. Honeine, "Nonlinear unmixing of hyperspectral images based on multi-kernel learning," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2012 4th Workshop on*, June 2012, pp. 1–4.

[9] Y. Zhang, H. L. Yang, S. Prasad, E. Pasolli, J. Jung, and M. Crawford, "Ensemble multiple kernel active learning for classification of multisource remote sensing data," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 8, no. 2, pp. 845–858, 2015.

[10] R. Alain, R. B. Francis, C. Phane, and S. G. Yves, "Simple mkl," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[11] P. Honeine and C. Richard, "The angular kernel in machine learning for hyperspectral data classification," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, June 2010, pp. 1–4.

[12] Y. Gu, G. Gao, D. Zuo, and D. You, "Model selection and classification with multiple kernel learning for hyperspectral images via sparsity," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 6, pp. 2119–2130, 2014.

[13] Y. Gu, Q. Wang, H. Wang, D. You, and Y. Zhang, "Multiple kernel learning via low-rank nonnegative matrix factorization for classification of hyperspectral imagery," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 8, no. 6, pp. 2739–2751, 2015.

[14] P. Gurram and H. Kwon, "Optimal sparse kernel learning in the empirical kernel feature space for hyperspectral classification," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 4, pp. 1217–1226, 2014.

[15] J. E. Ball, D. T. Anderson, and S. Samiappan, "Hyperspectral band selection based on the aggregation of proximity measures for automated target detection," in *SPIE Defense+ Security*. International Society for Optics and Photonics, 2014, pp. 908814–908814.

[16] J. E. Ball and L. M. Bruce, "Level set hyperspectral segmentation: Near-optimal speed functions using best band analysis and scaled spectral angle mapper," in *Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006. IEEE International Conference on*. IEEE, 2006, pp. 2596–2600.

[17] H. Jiang and W. Ching, "Correlation kernels for support vector machines classification with applications in cancer data," *Computational and mathematical methods in medicine*, vol. 2012, 2012.

[18] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. Bona, A. Binder, C. Gehl, and V. Franc, "The shogun machine learning toolbox," *The Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, 2010.