GENETIC ALGORITHMS AND LINEAR DISCRIMINANT ANALYSIS BASED DIMENSIONALITY REDUCTION FOR REMOTELY SENSED IMAGE ANALYSIS

Minshan Cui, Saurabh Prasad, Majid Mahrooghy, Lori M. Bruce, James Aanstoos

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

ABSTRACT

Remotely sensed data (such as hyperspectral imagery) is typically associated with a large number of features, which makes classification challenging. Feature subset selection is an effective approach to alleviate the curse of dimensionality when the number of features contained in datasets is huge. Considering the merits of genetic algorithms (GA) in solving combinatorial problems, GA is becoming an increasingly popular tool for feature subset selection. Most algorithms presented in the literature using GA for feature subset selection use the training classification accuracy of a specific algorithm as the fitness function to optimize over the space of possible feature subsets. Such algorithms require a large amount of time to search for an optimal feature subset. In this paper, we will present a new approach called Genetic Algorithm based Linear Discriminant Analysis (GA-LDA) to extract features in which feature selection and feature extraction are performed simultaneously to alleviate over-dimensionality and result in a useful and robust feature space. Experimental results with classification tasks involving both hyperspectral imagery and SAR data indicate that GA-LDA can result in very low-dimensional feature subspaces yielding high classification accuracies.

Index Terms— Remotely Sensed Data, Hyperspectral data, Genetic Algorithms, Feature Subset Selection, Feature Extraction.

1. INTRODUCTION

Hyperspectral Imagery (HSI) capture reflected radiation over a series of contiguous bands covering a wide range of the electromagnetic spectrum for every pixel in the image. Such imagery can provide features pertinent to the classification task at hand. However, analysis methods for such imagery first must reduce the dimensionality of this very high dimensional feature space to make any classification analysis tractable. Features derived from Synthetic Aperture Radar (SAR) imagery for analysis (e.g., raw backscatter, statistical features, texture features, discrete wavelet transform features etc.) can also result in very high dimensional feature spaces. Although this high dimensional data potentially provides relevant class-specific information for image analysis, it often also results in over-dimensionality and ill-conditioned statistical formulations.

Feature subset selection is hence a useful tool when the analysis involves high dimensional feature spaces. However, identifying and selecting relevant features from a large set of features is not a trivial task. Genetic algorithms (GA) have become popular tools for various feature subset selection problems [1]. However, the most popular GA based feature selection strategy uses the training classification accuracy to optimize and find the "best" feature subset. This greedy search is only suitable for classification tasks that do not operate in high dimensional feature spaces. As the feature-space dimensionality increases, the amount of time required for such a search increases significantly. Alternately, another class of GA-based feature subset selection algorithms employ a "filter" function that can be thought of as some metric that is optimized during the GA search. For classification problems, a good fitness function effectively measures the class-separation potential of feature subsets, thereby resulting in feature subsets that maximize class-separation.

In this work, we study two different metrics as potential filter functions for a GA based feature selection of high dimensional remotely sensed data. (1) Bhattacharyya distance (BD) [2], and (2) Fisher's ratio [3-5]. In particular, we propose an algorithm called Genetic Algorithm based Linear Discriminant Analysis (GA-LDA). Traditionally, for small-sample-size and high dimensional classification problems, a technique known as stepwise LDA (SLDA) [6] is commonly employed for dimensionality reduction. The key idea behind SLDA is that a preliminary forward selection and backward rejection is employed to discard redundant and less relevant features, and then a Linear Discriminant Analysis (LDA) projection is applied on this reduced subset of features to further reduce the dimensionality of the feature space. To test the efficacy of this method with different types of remotely sensed data, classification tasks involving both HSI and SAR data are studied. Classification experiments demonstrate that GA-LDA is much more successful at dimensionality reduction and feature selection compared to conventional approaches.

The rest of the paper is organized as follows. Section 2 presents the proposed GA-LDA based feature selection and

978-1-4577-1005-6/11/$26.00 ©2011 IEEE

dimensionality reduction algorithm. Section 3 describes the experimental data employed in this work. In section 4, classification results with experimental data are reported. Section 5 highlights the key contributions and successes of the proposed approach.

2. GENETIC ALGORITHM BASED LINEAR DISCRIMINANT ANALYSIS

Traditional feature selection techniques (such as forward selection and backward rejection) focus on evaluating the merits of each feature at a time and tend to ignore the importance of the relationship between features. The main advantage of a GA search compared with forward selection and backward rejection is that GA can take into account the relationships between features. In forward selection (FS) [7], one is unable to reevaluate the features that become irrelevant after adding some other features. Similarly, in backward rejection (BR), one is unable to reevaluate the features after they have been discarded. On the contrary, a GA search always attempts to evaluate the merits of combinations of features and their contribution to a fitness function. Genetic algorithms are a class of optimization techniques that search for the global minimum of a fitness function. This typically involves four steps – *evaluation*, *reproduction*, *recombination*, and *mutation* which are briefly explained below. The reader is referred to [8] for a detailed description.

- *Evaluation*: In this step, a random initial set of individuals will be selected, and each individual will be evaluated by a fitness function and will be assigned a fitness value. Then, all individuals will be ranked on the basis of the fitness values.
- *Reproduction*: During this step, a number of individuals with best fitness values in the current generation will be copied to the next generation. These individuals are called elite children.
- *Recombination*: In this step, some individuals with high fitness values other than elite children will be combined to produce new individuals. This step attempts to extract best genes from different individuals and recombine them into potentially superior children.
- *Mutation*: In this step, small portions of individuals undergo mutation according to some mutations rules. This step not only prevents the algorithm from getting trapped in a local minimum but increases the likelihood that the algorithm will generate individuals with better fitness values.

The Proposed Fitness Functions

In this paper, we will study two different metrics as filter functions for a GA-LDA based feature selection and dimensionality reduction: Bhattacharyya distance and Fisher's ratio.

Bhattacharyya distance (BD) [4] uses the first and second order statistics to measure the separation between two probability distribution functions. For two normally distributed classes, BD is defined as

$$\mathrm{BD} = \frac{1}{8}(\mu_2 - \mu_1)^{\mathrm{T}}\left(\frac{\Sigma 1 + \Sigma 2}{2}\right)^{-1}(\mu_2 - \mu_1) + \frac{1}{2}\ln\frac{\left|\frac{\Sigma 1 + \Sigma 2}{2}\right|}{\sqrt{|\Sigma 1|\,|\Sigma 2|}} \tag{1}$$

where μ_i and Σi are the mean vector and covariance matrix of class i, respectively. When BD is used as a fitness function in GA, GA will search for a feature subset that maximizes the BD value. Feature subsets producing higher BD value will hence likely be useful for the classification task at hand.

Another metric we studied in this work is Fisher's ratio [4]. LDA seeks to find a linear transformation w to a reduced dimensional subspace such that the ratio of within-class scatter to between class scatter, $J(w)$ in this projected subspace (provided by Fisher's ratio) is maximized:

$$J(w) = \frac{W^T S_B\, W}{W^T S_w\, W}, \tag{2}$$

where S_B and S_w are the between-class and within-class scatter matrices [4]. When Fisher's ratio is used as a fitness function in GA, it will search for features that maximize the Fisher's ratio, selecting a subset of features that yield the highest Fisher's ratio (and hence class-separation) after the LDA projection w is applied on them.

After GA based feature selection (using BD or Fisher's ratio), we apply an LDA projection to further project the subset of features on a reduced dimensional subspace optimized for classification. This GA-LDA approach is very similar to conventional SLDA, where a forward selection/backward rejection prunes redundant and less-useful features following which an LDA projection is carried out. Such algorithms are particularly useful when LDA cannot be directly applied on the input feature space owing to its very high dimensionality. We will demonstrate with our experimental results that GA is much more efficient at such a pruning compared to traditional stepwise selection approaches.

3. EXPERIMENTAL DATA

Two different types of remotely sensed datasets are used for experimentation. These datasets include both HSI and SAR imagery.

The experimental HSI data employed in this study was acquired using NASA's AVIRIS sensor [9] and was collected over northwest Indiana's Indian Pine test site in June 1992. The image represents a vegetation-classification scenario with 145x145 pixels and 220 bands in the 400 to 2450 nm region of the visible and infrared spectrum. Figure-1 depicts the spectral signatures for the eight classes extracted from this imagery.

The SAR data used in this experiment is from NASA Jet Propulsion Laboratory's Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVSAR) instrument, a polarimetric L-band synthetic aperture radar flown on a

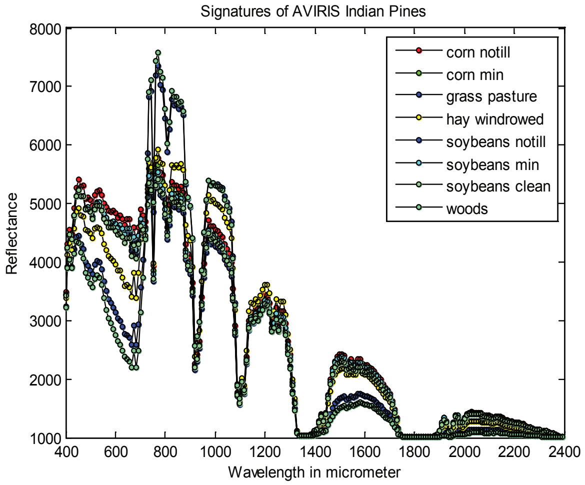

Figure 1: A plot of reflectance versus wavelength for eight classes of spectral signatures from AVIRIS Indian Pines data.

Table 1: Illustrating some salient characteristics of UAVSAR [10]

Parameter	Value
Frequency	L-band
Bandwidth	80 MHz
Range Resolution	1.8 m
Polarization	Full quad polarization
Quantization	12 bits
Antenna size	0.5 m range/1.5 azimuth
Power	> 2.0 kW

Gulfstream-3 research aircraft. Some important characteristics of the sensor are listed in Table 1. Normally, the UAVSAR flies at an altitude of 12.5 km and the ground swath is 20km. This dataset represents two classes – healthy levees and levees with landslides on them to represent a classification system that can identify and map problems (such as landslides in levee systems). In order to capture most of image on both sides of the river levees, the data was supposed to be collected in two straight-line flights. Flying in a “racetrack” pattern looking toward the river from opposite directions, a range of local incidence angles along the levees was achieved.

4. EXPERIMENTAL RESULTS

In experiments reported in this paper, we study the overall classification accuracy as a function of varying training sample sizes, for different number of features selected by the (GA or forward selection/backward rejection) algorithm. For both datasets, training and testing samples were selected randomly from the pool of available labeled samples. The random draw of training/test samples was repeated 10 times and the average classification accuracies are reported in these figures. System parameters for the GA search are as follows. The population size (defined as the number of individuals in each generation) was set to 100. We compare the performance of four algorithms in this work. GA-LDA-BD implies that BD was used as the fitness function in the GA-LDA algorithm, while GA-LDA-Fisher represents GA-LDA algorithm using fisher’s ratio as the fitness function.

Figure-2 shows the overall classification accuracies obtained with the HSI dataset. A Gaussian maximum-likelihood classifier was used in this work. From this figure, we can infer that the GA-LDA approach for dimensionality reduction is much better than traditional SLDA and LDA, and the improvement is specifically more pronounced when a few features are chosen, and when the amount of sample-size is small. In other words, the GA-LDA approach is very good at identifying the most relevant features – for example, when only a few features are selected, the performance of GA-LDA is much better than SLDA and LDA, indicating that the subset of features identified by the GA search is much more relevant and useful to the classification problem

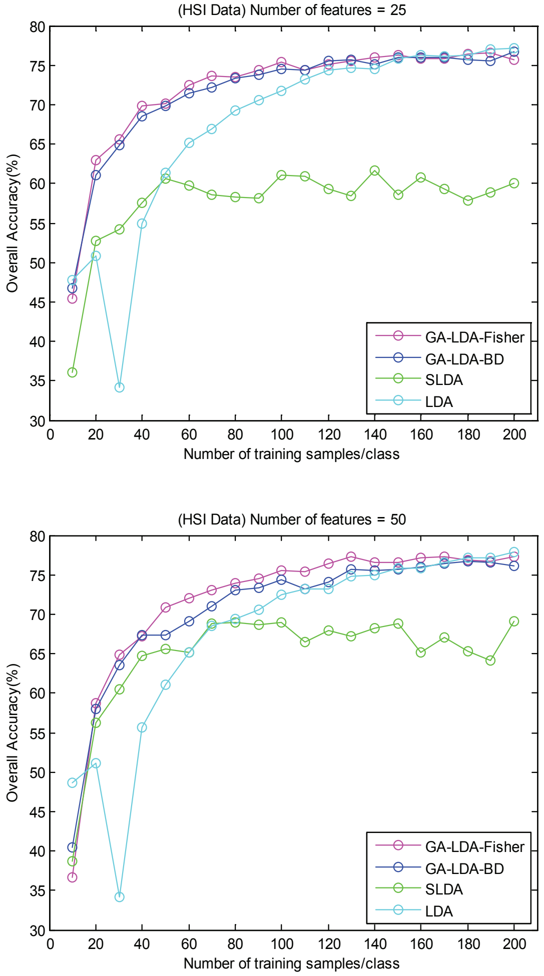

Figure 2: Illustrating the benefits of the proposed GA-LDA feature reduction approach for HSI land-cover classification using feature size 25 (*Top*) and 50 (*Bottom*).

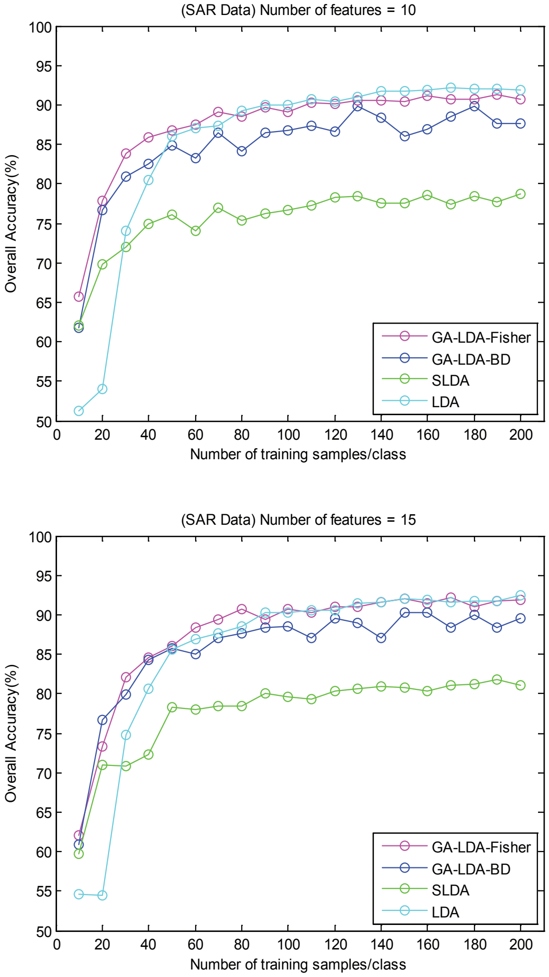

Figure 3: Illustrating the benefits of the proposed GA-LDA feature reduction approach for a SAR land-cover classification task using feature size 10 (*Top*) and 15 (*Bottom*).

at hand.

Figure-3 describes similar results for the 2-class SAR classification problem. Since this is a relatively simple classification problem, the overall accuracies using all the four methods is very good – however, the GA-LDA based approach is still better than traditional SLDA and LDA, especially when the number of training samples employed is very small, using a few features.

5. CONCLUSIONS

Experimental results presented in this paper indicate that a GA search is very effective at selecting the most pertinent features, while pruning out the most redundant features for classification tasks when an appropriate fitness function is employed. Akin to the conventional stepwise-LDA approach, we proposed a GA-LDA approach where GA first identifies a smaller dimensional subset of features upon which LDA is applied for final dimensionality reduction. Given a moderate feature space dimensionality and sufficient training samples, LDA is a good projection based dimensionality reduction strategy. However, as the number of features increases and the training-sample-size decreases, methods such as GA-LDA can assist by providing a robust intermediate step of pruning away redundant and less useful features. Consistent improvements in classification performance when using GA-LDA can be noted in our results. Finally, although the Fisher's ratio and BD provide similar information (by quantifying the class-separation ability of features), since LDA optimizes the Fisher's ratio, we note that GA-LDA-Fisher slightly outperforms GA-LDA-BD. This is expected because when the GA uses Fisher's ratio to perform its search, the final subset of features it identifies is already optimizing Fisher's ratio, resulting in a slightly better performance when LDA is applied on these features.

REFERENCES

[1] Ho-Duck Kim, Chang-Hyun Park, Hyun-Chang Yang, Kwee-Bo Sim, "Genetic Algorithm Based Feature Selection Method Development for Pattern Recognition," in *SICE-ICASE*, 2006.

[2] Chulhee Lee and Daesik Hong, "Feature Extraction Using the Bhattacharyya Distance," in *IEEE International Conference on Systems, Man, and Cybernetics*, 1997.

[3] Tran Huy Dat, Cuntai Guan, "Feature selection based on fisher ratio and mutual information analysis for robust brain computer interface," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[4] R.O. Duda, P.E. Stark, D.G. Stork, *Pattern Classification*, Wiley Inter-science, October 2000.

[5] S. Prasad and L. M. Bruce, "Limitations of Principal Components Analysis for Hyperspectral Target Recognition," in *IEEE Geoscience and Remote Sensing Letters*, vol. 5, pp. 625-629, 2008.

[6] S. Kumar, J. Ghosh, M.M. Crawford, "Best-bases feature extraction algorithms for classification of hyperspectral data," in *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 39, No. 7, pp 1368-1379, July 2001.

[7] Nakariyakul, S. ,Casasent, D.P., "Improved forward floating selection algorithm for feature subset selection," in *Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition*, Hong Kong, 30-31 Aug. 2008.

[8] K.S. Tang, K.F. Man, S. Kwong, Q. He, "Genetic algorithms and their applications," in *IEEE Signal Processing Magazine*, Vol. 13, Nov 1996.

[9] NASA Jet Propulsion Laboratory Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) web page http://aviris.jpl.nasa.gov/

[10] Kevin Wheeler, Scott Hensley, Yunling Lou, Tim Miller, Jim Hoffman, "An L-band SAR for repeat pass deformation measurements on a UAV platform", 2004 IEEE Radar Conference, Philadelphia, PA, April 2004.