STOCHASTIC FEATURE SELECTION WITH DISTRIBUTED FEATURE SPACING FOR HYPERSPECTRAL DATA

Jeffrey D. Clark, Michael J. Mendenhall, Member, IEEE, and Gilbert L. Peterson

Department of Electrical and Computer Engineering
Air Force Institute of Technology
Wright-Patterson AFB, OH.

ABSTRACT

Feature subset selection is a well studied problem in machine learning. One short-coming of many methods is the selection of highly correlated features; a characteristic of hyperspectral data. A novel stochastic feature selection method with three major components is presented. First, we present an optimized feature selection method that maximizes a heuristic using a simulated annealing search which increases the chance of avoiding locally optimum solutions. Second, we exploit local cross correlation pair-wise amongst classes of interest to select suitable features for class discrimination. Third, we adopt the concept of distributed spacing from the multi-objective optimization community to distribute features across the spectrum in order to select less correlated features. The classification performance of our semi-embedded feature selection and classification method is demonstrated on a 12-class textile hyperspectral classification problem under several noise realizations. These results are compared with a variety of feature selection methods that cover a broad range of approaches.

Index Terms— Hyperspectral, feature selection, detection, dimensionality reduction

1. INTRODUCTION

Hyperspectral data provides an abundance of spectral information per spatial location; however, processing this type of data can be computationally time intensive. This computational burden may by reduced if a subset of meaningful features can be extracted. Many methods exist to extract meaningful feature subsets from hyperspectral data, their accuracy depends on the ability to retain important classification information.

Feature selection methodologies are classified into three types: filter, wrapper, and embedded [1]. The filter method incorporates properties of the data to select a feature subset; some examples include Relief F and Bhattacharyya methods. The wrapper method uses a feedback of classification accuracy to select the appropriate feature subset; for example genetic and Best First Search methods. Embedded methods determine a feature subsets' "goodness", and continually updates the feature subset to produce a better feature subset until the stopping criterion is met [1]. Examples of these methods are C4.5 [1] and generalized relevance learning vector quantization improved (GRLVQI) [2]. All of the previously mentioned methods train on the data obtained from the a specific imaging system and provide feature subsets based on the resolution of that imaging system.

The resolution of typical lab collected hyperspectral data is on the order of a magnitude better than most imaging systems in the field. The discriminating ability of the feature subsets produced from higher resolution lab data are typically better than what might be obtained with lower spectral resolution fielded systems. The feature selection method presented in this paper allows for the unique ability to select features realizable by a low resolution imaging system based on the high resolution laboratory data, allowing for a more informed selection of features resulting in excellent classification accuracy. This is accomplished by adjusting an analysis window over the laboratory data to accommodate the resolution capability of the target imaging system used to collect the hyperspectral data. This novel methodology incorporates a stochastic search approach (simulated annealing search algorithm [3]) with a heuristic that guides the search. The result is a non-greedy local search of the spectral domain that provides a locally optimal solution. The feature subset is less-redundant than other methods, due to a distributed spacing algorithm incorporated into the search.

One of the goals of feature selection is to produce a feature subset that is not correlated, where correlation indicates redundancy. Redundant features are often noted as adding nothing new to the discriminating capability of the feature subset and are typically considered unnecessary [4]. The feature selection method presented in this work produces a less-redundant/less-correlated feature subset that demonstrates good discriminating capability in the presence of

The views expressed in this paper are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This work is sponsored by the Air Force Research Laboratory, Sensors Directorate.

noise. This is accomplished by incorporating the distributed spacing concept, typically used to solve Multi-Objective optimization problems, as outlined by Coello Coello *et.al.* [5], and Deb and Srinivas [6].

A correlation-based detector is specified that compliments the proposed feature selection methodology. The capability of the detector is compared to that of the Minimum Eucledian Distance (MED) classifier with the features selected using GRLVQI, Bhattacharyya, Relief F and the feature selection method presented in this work. The result of the detector/classifier is shown over a range of additive white Gaussian noise realizations.

2. PROPOSED FEATURE SELECTION/DETECTION

The Non-correlated Aided Simulated Annealing Feature Selection (NASAFS) method selects features in a pairwise manner, where the results of each pair is combined into a database of ‘distinguishable features’ for discriminating the reference class from all other classes. The database is then used by the detection algorithm to categorize unseen data samples. Although Simulated Annealing (SA) can be processing intensive, it is not a limiting factor in the processing of the feature set.

2.1. Feature Selection Methodology

The proposed NASAFS works as follows:

1. Train on the reference class samples to determine the covariance threshold (k) for the heursitic function.
2. Randomly select a subset of feature ‘bins’ (*Fig. 1*), ensuring optimal distribution across the signal domain, and then evaluate them with the heuristic.
3. The heuristic evaluates the feature set and returns a value (*Eqn. 3*) to guide the SA search.
4. A feature ‘bin’ in the feature subset is replaced with a random pick of the remaining bins in a manner that maintains the distributed spacing requirement, then the new feature subset is evaluated by the heuristic.
5. The value returned from the heuristic is used to determine if the new subset is either kept or discarded.
6. A different feature ‘bin’ in the feature subset is selected and steps four through six repeated until convergence (currently a fixed number of training steps).

NASAFS trains on the reference class samples by finding the worst cross covariance of every combination of the samples for each corresponding bin of the reference class (*i.e.,* $R_1(A)$ with $R_2(A)$, $R_1(A)$ with $R_3(A)$ *etc.*, where in the form $R_s(b)$, R is the reference class, s is the sample, and b is a specific bin, see Fig. 1). This covariance value becomes the threshold (k) used in the heuristic in Eqn. 3. A less-correlated feature subset is ensured by distributing newly selected features across the spectrum. Feature distribution across the spectrum is accomplished by the method presented by Coello Coello *et.al.* [5] and proposed by Deb and Srinivas [6], and is accomplished by computing the measure of distributed spacing (ι) [6] which is the value placed on the distribution of the features in the feature subset according to the division/sub-regions of the spectral bands:

$$\iota = \sqrt{\sum_{i=1}^{q} \left(\frac{n_i - \overline{n}_i}{\sigma_i} \right)^2} \tag{1}$$

where q is the number of sub-regions (*Fig. 1*) that the signal is divided into, n_i is the actual number of selected feature points in the i^{th} sub-region of the signal(s), $\overline{n}_i$ is the expected number feature points in the i^{th} sub-region of the signal (if sub-regions are unequal, a weighting must be applied), and σ_i^2 is the variance of the selected feature points of the i^{th} sub-region of the signal. The variance is calculated similar to that in [6]:

$$\sigma_i^2 = \overline{n}_i \left(1 - \frac{\overline{n}_i}{P} \right) \text{ for } i = 1, 2, ..., q \tag{2}$$

where P is the total number of spectral bands. The number of sub-regions (q) of the signal is user defined, but could be determined in some other manner such as the correlation structure of the data. The best (optimal) less-correlated case, based on the number of allowed features in the feature subset and the number of sub-regions over the signal, is calculated. This best case value (*e.g.,* if the feature subset is to contain six features, then for the case in Fig. 1, there would be two features per sub-region) is used as a baseline when determining the actual correlation of the selected feature subset. A feature subset that meets the desired percentage of optimality (user defined) is allowed to proceed to the heuristic function of the feature selection algorithm. Otherwise, the previously picked feature must be returned to the open list and a replacement feature is randomly picked and the process repeated.

The heuristic returns values based on a sequential set of calculations. The cross covariance $C_{R,T}^{f}$ is calculated using a feature subset of the reference class corresponding to the feature subset of the target class (*e.g.,* $R_1(A, C, E, I)$ to $T_1(A, C, E, I)$, where T is the target class, Fig. 1). This value is compared to the covariance threshold k. If $C_{R,T}^{f} \geq k$, then $(1 - C_{R,T}^{f})$ is returned per Eqn. 3, otherwise the autocorrelation ($\Re_x$) of the bins of each class (reference and target) is accomplished and the absolute distance of these values, N_d, is computed where $N_d = |\Re_x(T_1(A, C, E, I) - \Re_x(R_1(A, C, E, I))|$. The result is compared to a distance threshold D_t (which is updated every time this threshold is exceeded) to determine the appropriate value to return. The heuristic h is expressed as:

$$h = \begin{cases} 1 - C_{R,T}^{f} & \text{if } C_{R,T}^{f} \geq k, \\ 1 & \text{if } C_{R,T}^{f} < k \text{ and } N_d - D_t \geq 0, \\ 1 - C_{R,T}^{f} & \text{if } C_{R,T}^{f} < k \text{ and } N_d - D_t < 0. \end{cases} \tag{3}$$

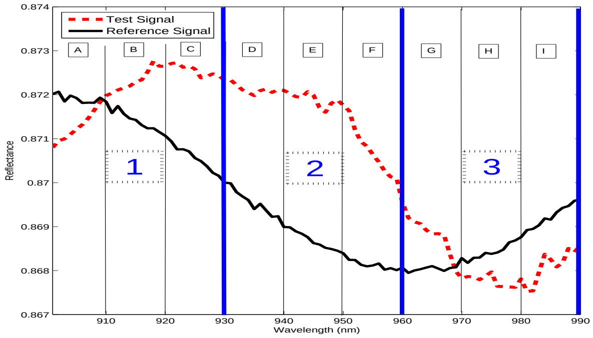


Fig. 1. Example of a signal segment divided into bins and sub regions. Here, black (solid line) is the reference signal and red (dashed line) is the test signal, 1 . . . 3 are the subregions and A . . . I are the bins.

Since the heuristic uses features that are bins of the spectral domain, the user must define a bin size, typically set to the bandwidth of the targeted collection system, used for the detection task. This bin size is used to parse up the signal, an example of bin size and sub-regions of the spectral bands is shown in Fig. 1. The feature selection algorithm selects a random (specified by the user) number of features as the starting feature subset. The feature subset is evaluated by the heuristic (Eqn. 3), which ranks the feature subset as a group. Each feature, within the original feature subset, is replaced by a randomly selected feature, one at a time and sequentially from the open list of features. Each feature replacement creates a new feature subset that is evaluated by the heuristic. If the new feature subset is kept, according to the SA algorithm, the feature that it replaced is put back on the open list. If the new feature subset is not kept, the replaced feature is restored to the feature subset and the new feature is placed on the open list.

2.2. Detector Methodology

The correlation detection method (CoDeM) is based on the principals of the feature selection method presented earlier. Due to the pairwise process of NASAFS, CoDeM also performs a pairwise detection that labels a sample as either in-class or out-of-class. If the sample is out-of-class, no other information is known, just that it is not the same as the reference class. CoDeM uses the average value of each bin for its calculations. Thus, for a feature subset consisting of six bins from NASAFS, the feature subset of CoDeM will have six scalar values.

CoDeM determines a covariance threshold (k_d) using noisy reference data, where the noise is a fraction of the noise power of the test samples. It determines the worst cross covariance of every combination of the samples for each corresponding bin of the feature subset of the reference class (*i.e.*, $R_1(A)$ with $R_2(A)$, $R_1(C)$ with $R_2(C)$, *etc.*). A distance threshold (d_d) is determined by finding the absolute distance of the autocorrelation of the non-noisy bins of the reference class to the noisy bins of the reference class. The cross covariance $C_{R,T}$ is calculated for the mean of the non-noisy reference class bins to the mean of the noisy test class bins. T_d is the absolute distance of the autocorrelation of the non-noisy reference class bins to the noisy test class bins. If $C_{R,T} \leq k_d$ then the test sample is determined to be out-of-class. If $C_{R,T} > k_d$ and $T_d \geq d_d$, then the test sample is determined to be out-of-class. If $C_{R,T} > k_d$ and $T_d < d_d$ then the test sample is determined to be in-class. This process is expected to produce the most realistic results if an appropriate target sensor noise model is incorporated.

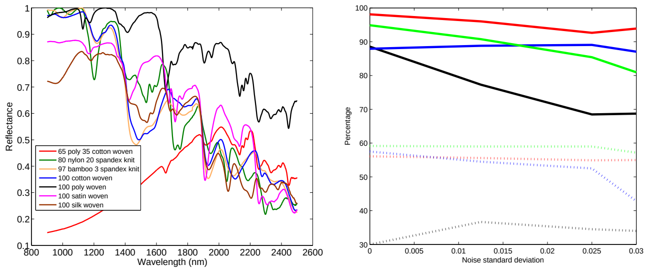
3. EXPERIMENTAL SETUP

A 12 class hyperspectral data set was used with NASAFS in comparison to GRLVQI, Relief F, and the Bhattacharyya methods. NASAFS is implemented using a bin size of $10nm$, which is the average bandwidth of an imaging collection system in the field; for example the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), which has a $10nm$ nominal channel bandwidth. The data was collected by a hand held reflectometer with a sampling interval of $1nm$ (as such, spectral mixing is not considered at this time). In order to present a fair comparison, the data used in the other feature selection methods was re-sampled to $10nm$ bins *via* an averaging technique. Relief F, GRLVQI and the Bhattacharyya methods were computed using a three-fold cross validation.

NASAFS used ten reference class samples to obtain the covariance threshold used in the heuristic, and generated the feature sets based on one sample from each of the other classes. For NASAFS, we divided the spectral domain into seven equal sub-regions for the feature spreading aspect. For all the methods, no more than six features were allowed to be in the feature set (this selection is chosen arbitrarily for computational considerations). To determine the accuracy of the chosen feature sets, the MED and the CoDeM presented in this paper were used. Both of these methods were compiled using a range of noise power added to the test signals. The noise powers used were related to the average noise levels of a field imaging system. Since the Bhattacharyya, Relief F, and the GRLVQI produce global feature sets, the NASAFS method (which produces pairwise feature sets) was implemented in a pairwise manner with the MED and the results for each class were averaged to obtain a single global accuracy measure. Since the CoDeM is set up in a pairwise manner, the results of the feature sets from each of the different methods are averaged to get single global accuracy results.

4. RESULTS AND CONCLUSION

Figure 2 shows representative samples for the 12 classes used in this experiment. NASAFS produced discriminative feature sets with a low correlation value. Figure 3 shows the hyperspectral signal for 80% Polyester 20% Rayon blend with the respective feature subset coefficients for the different feature selection methods. Visually, it is seen that the features of the feature subset denoted by NASAFS are spread over the reference signal, whereas the feature subsets of the other methods tend to be close together. The correlation matrix of the data is used to determine the correlation coefficient for each feature subset of each feature selection method as shown in Table 1. The single correlation coefficient value for each method was obtained by averaging the correlation coefficients for all combinations of each feature within each feature subset. NASAFS produces less correlated feature subsets than the other methods.



(a) Data

(b) Accuracy

Fig. 2. (a) Representative samples from the 12-class textile data set used in the evaluation of NASAFS. (b) Results of the detection/classification methods for each feature selection methodology using CoDeM (solid line) and MED (dashed line). Each feature selection method is represented with a different color: Relief F (blue), NASAFS (red), GRLVQI (green), and Bhattacharyya (black).

Figure 2 shows the average accuracy of the feature selection methods employed in this work as compared with the MED and the CoDeM. NASAFS classification results are better than the other methods when using CoDeM, and comparable to GRLVQI with the MED process. Currently, a global feature set for NASAFS is being sought after that shows promise. Presently, the feature spreading function of NASAFS produces much better feature subset from a correlation standpoint. Future work aims at extending the feature spreading methodology to existing feature selection methods (*e.g.*, GRLVQI). It also aims to better define less-correlated sub-regions, for example using the correlation matrix of the data to define the sub-regions of the signal. Further, adding more realistic noise, atmospheric effects and accurate resampling of the data for a specific collection imaging system is currently underway.

Method	Mean Corr Coef
Relief F	0.9955
GRLVQI	0.9025
Bhattacharyya	0.9998
NASAFS	0.6402

Table 1. Average correlation coefficients.

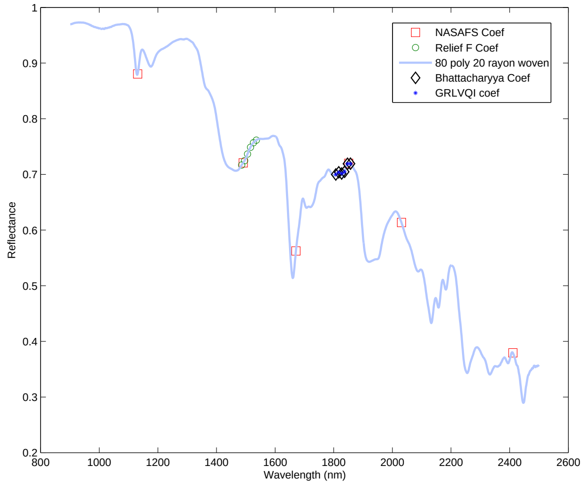


Fig. 3. Hyperspectral signal for 80% Polyester 20% Rayon blend with the respective feature set coefficients for the different feature selection methods.

5. REFERENCES

[1] M. Dash and H. Liu, “Feature selection for classification,” Tech. Rep., Department of Information Systems and Computer Science, National University Of Singapore, Singapore 119260, March 1997.

[2] M.J. Mendenhall and E. Merényi, “Relevance-based feature extraction for hyperspectral images,” *IEEE Transactions of Neural Networks*, vol. 19, no. 4, pp. 658–672, April 2008.

[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, New Jersey 07458, 2003.

[4] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 1157–1182, March 2003.

[5] C.A. Coello Coello, G.B. Lamont, and D.A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, Springer Science + Business Media, New York, NY, 2nd edition, 2007.

[6] N. Srinivas and K. Deb, “Multiobjective optimization using nondominated sorting in genetic algorithms,” *Evolutionary Computation*, vol. 2, no. 3, pp. 221–248, fall 1994.