FEATURE SUBSET SELECTION WITH APPLICATIONS TO HYPERSPECTRAL DATA

Hao Chen and Pramod K. Varshney

Electrical Engineering and Computer Science Department
Syracuse University, Syracuse, NY, 13244

ABSTRACT

Feature subset selection is very important in high dimensional datasets such as hyperspectral images. In this paper, we define a new feature redundancy measure. Two different feature selection algorithms are proposed based on this measure. Experimental results on a real hyperspectral dataset are presented to demonstrate the effectiveness of our methodology.

1. INTRODUCTION

There are many signal processing applications where high-dimensional datasets need to be processed. One such applications is hyperspectral data where hyperspectral sensors can provide hundreds of bands of data simultaneously. However, processing such high dimensional data is computationally very complex. Also due to a lack of sufficient training samples for this high dimensional dataset, the curse of dimensionality[1] becomes a serious issue. Therefore, reducing the dimensionality of the raw input data space is a very important step in hyperspectral data processing. Many different feature(band) selection algorithms have been proposed [2]-[5]. In these algorithms, significantly fewer bands are selected for further processing such as for classification or target detection.

Principal Components Analysis(PCA) has been widely used for feature reduction in the past. PCA transforms features $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$ into a set of new features $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_N]^T = \mathbf{W}^T\mathbf{X}$ which are statistically uncorrelated with each other. Here, $\mathbf{W} = (W_1, W_2, \ldots, W_N) = [W_{ij}]_{1\leq i,j\leq N}$ is the eigenmatrix of the covariance matrix of X. The variance of the first principal component(PC) is maximum, the variance of the second PC is the second largest and so on. After transformation, PCs with significant variances are selected. PCA does provide an approach to reduce data dimensionality. However, bands with highly discriminant information but with small variances may be concealed in PCA transformations. Furthermore, in some cases, it is of more interest to find the most useful original bands for specific applications. Therefore, some feature selection techniques based on PCA have been proposed.

In methods based on feature ranking in conjunction with PCA, a suitably defined score of each individual band is calculated and bands are ranked from high to low based on the value of the scores. Bands with largest scores are selected as a result of the feature selection algorithm. In [6], the author uses $F_x(i) = \sum_j |\lambda_j W_{ij}|$ as the score of the original band i. In [5], the discriminant power measure

$$\rho_x(i) = \sum_j \left|\lambda_j W_{ij}^2\right| \tag{1}$$

has been used. After a preliminary feature subset has been selected, a Divergence-Based Band Decorrelation procedure is applied in order to remove the highly correlated bands. Notice that the variance of feature X_i is

$$\sigma_{x_i}^2 = \sum_j \left|\lambda_j W_{ij}^2\right| = \rho_x(i). \tag{2}$$

From (2), we can see that features are selected based on their variance. However, the variance of a single band is not very useful because by simply multiplying the band by a constant a, for say image enhancement, the variance of that band will be changed from σ^2 to $a^2\sigma^2$. In this case, the classification performance will not change. As to the divergence measure, it only measures the difference between two different features by comparing their distributions. In fact histograms of the two images are used most of the time. Since a histogram only describes the intensity distribution of an image, and it does not provide any information in the spatial domain, this method may eliminate important features.

In [4], dominant eigenvectors with corresponding eigenvalues that are much larger than the others are selected. The angles between the original features and dominant eigenvectors are calculated as the feature scores. Obviously, since no comparison between the original features is used, this method still suffers from high redundancy between the selected features.

In this paper, we define a new feature redundancy measure. Two different feature selection algorithms are proposed based on this measure.These methods can be used as stand alone methods or can be combined with methods discussed in [4]-[6].The relation between PCA and our proposed method is also explored.

2. FEATURE SELECTION ALGORITHM

Let $X = [X_1, X_2, \ldots, X_N]^T$ represent a feature set. Without loss of generality, we assume that X_is are random vari-

ables with zero mean and unit variance. Therefore, the feature X_i is redundant if it can be expressed as follows:

$$X_i = f(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N). \quad (3)$$

Here, f is a suitable function. Obviously, removal of X_i from the original feature set does not cause any loss of information and the feature subset $\mathbf{X'} = [X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N]^T$ is sufficient. If f is a linear function, X_i is linearly redundant. To describe the redundancy of a feature X_i, a number of criteria can be used. In this paper, we use the MSE criterion $R(X_i) = \min(E(X_i - f(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N))^2)$ as a measure of redundancy of the feature X_i. Obviously, the lower $R(X_i)$ is, the more redundant the feature X_i is. When $R(X_i) = 0$, X_i is totally redundant given the other features. In this paper, we restrict the function f to be a linear function. This method can also be extended easily to non-linear functions.

2.1. Linear Redundancy measure

Let $f(x) = \sum_{j=1, j\neq i}^{N} \alpha_i X_i$, then

$$\begin{aligned} R(X_i) &= E(\min_{\alpha}(X_i - (\sum_{j=1, j\neq i}^{N} \alpha_i X_i))^2) \\ &= \min_{\beta, \beta_i = 1} \beta^T \Sigma_X \beta \end{aligned} \quad (4)$$

where $\beta^T = [-\alpha_1, -\alpha_2, \ldots, -\alpha_{i-1}, 1, -\alpha_{i+1}, \ldots, -\alpha_N]$.

Let $\beta = \sum_{j=1}^{N} C_j W_j$. Since $\mathbf{W}$ is an orthogonal matrix and $\beta_i = 1$, (4) becomes $R(X_i) = \min_{C} \left(\sum_{j=1}^{N} C_j^2 \lambda_j \right)$ subject to $\sum_{j=1}^{N} C_j W_{ij} = 1$. Using the Lagrange multiplier method for this constrained minimization problem and with some calculations, we have, $C_j = \frac{W_{ij}}{\lambda_j \sum_{m=1}^{N} \frac{W_{im}^2}{\lambda_m}}$, and,

$$R(X_i) = \frac{1}{g_i} \quad (5)$$

where, $g_i = \sum_{j=1}^{N} \frac{W_{ij}^2}{\lambda_j}$. Obviously, the higher g_i is, the more redundant X_i is. When $g_i = \infty$, $R(X_i) = 0$, and X_i is totally redundant, i.e., elimination of X_i from the original feature set will not lose any information. Also, "normalizing" the column eigenmatrix $\mathbf{W} = [W_1, W_2, \ldots, W_i, \ldots, W_N]$ by dividing the $W_i s$ with $\sqrt{\lambda_i}$ to form a new matrix

$$\begin{aligned} \omega &= [\omega_1, \omega_2, \cdots, \omega_i, \cdots, \omega_N] \\ &= [\frac{W_1}{\sqrt{\lambda_1}}, \frac{W_2}{\sqrt{\lambda_2}}, \cdots, \frac{W_i}{\sqrt{\lambda_i}}, \cdots, \frac{W_N}{\sqrt{\lambda_N}}], \end{aligned} \quad (6)$$

we get,$R(X_i) = \frac{1}{\|\omega_i\|_2^2}$, where $\|.\|_2$ is the second norm of a vector.

2.2. Properties of the Redundancy Measure

Let S_N be an N-Dimensional vector space and X_is be the vectors that span it. Then, the expression $\sum_{j=1, j\neq i}^{N} \alpha_i X_i = \alpha^T X'$ in equation (4) is a new vector in the subspace S_{N-1} that is spanned by X'. From the definition of $R(X_i)$, it is easy to see that the vector that minimizes $R(X_i)$ is the orthogonal projection of vector X_i in the subspace S_{N-1}. Let $\Sigma_{X_i X'} = [\sigma_{i,1}, \sigma_{i,2}, \ldots, \sigma_{i,i-1}, \sigma_{i,i+1}, \ldots, \sigma_{i,N}]^T$ be the covariance between X_i and X' and $\Sigma_{X'}$ be the covariance matrix of X'. Therefore, from the definition of the orthogonal projection, $\alpha = (\Sigma_{X'})^{-\frac{1}{2}} \Sigma_{X_i X'}$, we have

$$R(X_i) = \beta^T \Sigma_X \beta = 1 - \Sigma_{X_i X'}^T \Sigma_{X'}^{-1} \Sigma_{X_i X'} = \frac{|\Sigma_X|}{|\Sigma_{X'}|}. \quad (7)$$

From equation (7), it is easy to see that the subset that gives the least redundancy actually corresponds to the largest value of $\Sigma_{X'}$ among all the possible feature subsets. By assuming that the selected features follow a multinormal distribution, the least redudant feature subset has the largest entropy $I = log((2\pi)^N |\Sigma_{X'}|)$. Both equations (7) and (6) will be used in our proposed algorithms based on different search algorithms.

2.3. Feature Selection Based on the Redundancy Measure

A feature subset with low redundancy can be found by sequentially eliminating the most redundant features. In practice, in order to define the size of the "best" feature subset, two criteria are often used. One is to predefine the size of the feature subset, the other is to predefine a threshold that will stop the selection procedure when this threshold is reached. Combining the redundancy measure with the sequential backwards search(**SBS**) and these stopping criteria, we propose a SBS feature selection algorithm as follows.

1. Initialize $\mathbf{Y} = X$, calculate its covariance matrix Σ_Y
2. Find the most redundant feature to be eliminated from $\mathbf{Y}$.
   (a) Find the eigendecomposition of $\Sigma_Y = W_Y \cdot diag(\lambda_1, \lambda_2, \cdots, \lambda_{|Y|}) \cdot W_Y^T$ and calculate ω by (6)
   (b) for $i = 1, \ldots, |\mathbf{Y}|$, calculate $g_i = \|\omega_i\|_2^2$
   (c) Find the feature Y_K that corresponds to $K = argmax_i(g_i)$
3. Update the feature subset $\mathbf{Y}$ by either one or both of following criteria,$A\backslash a$ means remove element a from the set A.
   (a) If $g_K > \tau$, then $\mathbf{Y} = \mathbf{Y}\backslash Y_I$, here, τ is a predefined threshold to indicate the degree of redundancy we can omit.

(b) If $|\mathbf{Y}| > S$, then $\mathbf{Y} = \mathbf{Y} \backslash Y_I$, here, S is a predefined subset size.

4. If $|\mathbf{Y}|$ is decreased in step 3, go to step 2.

5. Output the desired feature subset $\mathbf{Y}$.

The SBS based algorithm is quite suitable to be used in combination with other feature selection algorithms,e.g, we may use the algorithm of [4] to select a "preliminary" feature subset with possibly highly redundant bands and use our algorithm to further reduce the dimensionality of the feature subset. Using g_i with the Sequential Forward Search(SFS) algorithm, we may also obtain the algorithm to find the best feature subset as follows:

1. Initialize the output feature subset $\mathbf{Y} = \Phi$, the null set.

2. Among all possible feature pairs X_i, X_j, find features X_I, X_J that are most uncorrelated, i.e., select $1 \leq I < J \leq N = \underset{i,j}{argmin} |\Sigma_{X_i,X_j}|$. Here, Σ_{X_i,X_j} is the covariance matrix for features X_i, X_j. Let $\mathbf{Y} = \mathbf{Y} + \{X_I, X_J\}$ and $X = X \backslash \{X_I, X_J\}$.

3. Find the next best uncorrelated feature. For $i = 1, 2, \ldots, |X|$,do

   (a) Find the correlation vector Σ_{YX_i} (these coefficients are calculated in step 2). By using (7), calculate $R(X_i)$.

4. Select $I = \underset{i}{argmin}\, R(X_i)$.

5. Update the feature subset $\mathbf{Y}$ by either one or both of the following methods.

   (a) If $|\mathbf{Y}| < S$, then $\mathbf{Y} = \mathbf{Y} + X_I$ and $X = X \backslash X_I$. Here, S is a predefined subset size.

   (b) If $R(X_I) > \tau$, i.e, the new feature is not redundant to the acceptable degree with the existing features in $\mathbf{Y}$, then $\mathbf{Y} = \mathbf{Y} + X_I$ and $X = X \backslash X_I$.

6. If $|\mathbf{Y}|$ is increased in step 5, goto step 2.

7. Output the desired feature subset $\mathbf{Y}$.

3. EXPERIMENTS AND RESULTS

In this section, some experimental results are presented to demonstrate the effectiveness of our algorithm. The data used here is a segment of an AVIRIS data scene taken of NW Indiana's Indian Pine test Site. From the 220 spectral bands of original data, 185 bands are used, discarding twenty water absorption bands as well as fifteen noisy bands by visual inspection. To test the classification accuracy, four different classes are chosen, they are "Corn-notill","Soybean-notill","Soybean-min" and "Corn". The ML classifier is used in both experiments, 50% of the pixels are randomly chosen as training data set and the rest are chosen as testing data set. The classification accuracy results presented are the averages of 50 runs.

In the first experiment, four different feature selection algorithms are tested. They are Kermani et al. ([4]), Chang et al.([5]), Campbell's ([6]), and our proposed SFS based selection algorithm. Figure 1 displays the overall classification accuracy achieved by these different algorithms versus different feature subset sizes($|\mathbf{Y}| = 1, 2, 3, \ldots, 50$).

We can see it clearly that when the feature subset size is not very large $|\mathbf{Y}| < 15$, the performance of our SFS feature selection algorithm is significantly better than all the other algorithms. The reason is that our algorithm selects some of least redundant features while the desired feature subset size is relatively small. This ensures a higher classification accuracy. This property can be seen much more clearly from Table 1[1], where, in order to achieve 60% classification accuracy, the size of the feature subset selected by our algorithm is 7, which is much less than the size of feature subsets selected by other algorithms(15,17 and 20 respectively). As shown in Figure 2, due to the large dimension of the original feature set $|X|$, the score of each individual feature is pretty small. Therefore, the size of the selected feature subset will have to be relatively large to reach a reasonable threshold. However, our redundancy measure drops dramatically after a few steps, therefore, it is easier to select a near optimal feature subset by using our SFS selection algorithm. In this experiment, we also notice that when $10 \leq |\mathbf{Y}| \leq 28$, the classification performance of our SFS selected feature subset increased very slowly and there is a significant increase in classification accuracy when $|\mathbf{Y}| = 29$. This is because with the increase of the feature subset size, the redundancy of the newly selected feature will be very high, when $i \geq 10$, $R(Y_i) << 0.01$. Therefore, the additional information provided by each additional band Y_i is limited for $i \geq 10$. However, when the feature subset size $\mathbf{Y}$ is large enough, due to the well known "XOR" phenomenon[2], classification performance will be increased. In this experiment, when $|Y| = 29$, this phenomenon occurred.

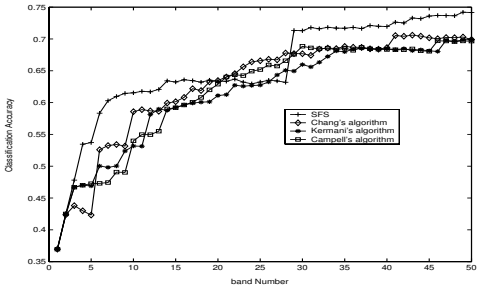

Fig. 1. Performance of different algorithms

[1]Size: the size of feature subset. CA: classification Accuracy.

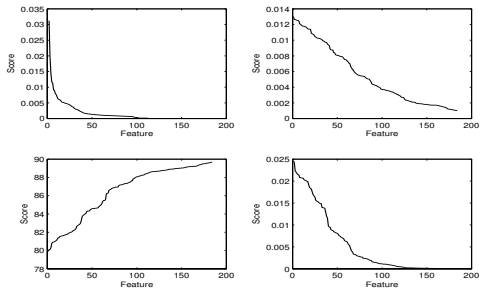

Fig. 2. Normalized Scores of original features in descendant order. upper left: $R(X_i)$ in SFS based method, upper right:[6], bottom left:[4], bottom right:[5]

In the second experiment, we first choose three different feature subsets selected by using the three algorithms using the corresponding thresholds. In Kermani's method, we choose $\tau_1 = 84$(angle between the original feature and the major PCs) to select a feature subset $Y1, |Y1| = 47$. In Chang's method, we select $\tau_2 = 0.7$(the score percentage of the feature subset versus the score of the whole feature set) to select a feature subset $Y2, |Y2| = 39$ and in Campbell's method, we select $\tau_3 = 0.38$(the percentage of the score of the feature subset versus the score of the whole feature set) to select a feature subset $Y3, |Y3| = 33$. SBS based feature elimination is then used to prune the redundant feature subsets. The classification performance for the three cases is shown in Figure 3. Again, when the size of the pruned feature subset is relatively small,our SBS feature elimination method prunes many highly redundant features in each feature subset without losing much classification accuracy.

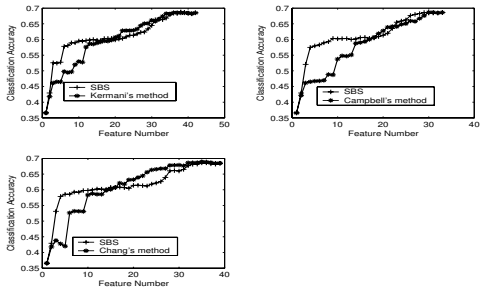

Fig. 3. The performance of SBS feature elimination in conjunction with methods in [4],[6],[5]

4. CONCLUDING REMARKS

In this paper, a new redundancy measure is proposed for feature subset selection in high-dimensional datasets. Based on this measure, the SFS feature selection algorithm and the SBS feature elimination algorithm are proposed. A couple of experiments using hyperspectral data show the effectiveness of our algorithm. Several extensions to this work are planned.

Table 1. Comparison of the size of the selected feature subsets to achieve the same classification accuracy.

	Size	CA
Kermani's ([4])	20	0.6071
Campbell's ([6])	17	0.6000
Chang's ([5])	15	0.6011
SFS Based	7	0.6029

5. ACKNOWLEDGMENTS

This research was supported by NASA under grant NAG5-11227. We would like thank the Laboratory for application of Remote Sensing at Purdue University for providing the AVIRIS data used here.

6. REFERENCES

[1] D. W. Scott, "The curse of dimensionality and dimension reduction," in *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: Wiley, 1992, ch. 7, pp. 195-217.

[2] Isabelle Guyon, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp.1157-1182, Mar. 2003

[3] P.Groves and P. Bajcsy, "Methodology for hyperspectral band and classification model selection ," *IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*, pp.120-128, Oct. 2003

[4] B. Ghaffarzadeh Kermani, Susan S. Schiffman, and H. Troy Nagle, "A novel method for reducing the dimensionality in a Sensor Array," *IEEE Trans. on instrumentation and measurement*, vol.28, pp.728-741,May,1998

[5] Chein-I Chang, Qian Du, Tzu-Lung Sun, and Mark L.G. Althouse, "A Joint Band Prioritization and Band-Decorrelation Approach to Band Selection for Hyperspectral Image Classification," *IEEE Trans. on Geoscience and Remote Sensing*, vol.37, pp.2631-2641, Nov. 1999

[6] J.B.Campbell, *Introduction to Remote Sensing*,2nd ed. New York: The Guilford Press, 1996