Hyperspectral Image Classification Using Dynamic Classifier Selection with Multiple Feature Extractions

Chia-Hao Pai(1), Bor-Chen Kuo(1), Tian-Wei Sheu(1), Jinn-Min Yang(2), and Li-Wei Ko(1)
(1) Graduate School of Educational Measurement and Statistics,
National Taichung Teachers College, Taichung, Taiwan
(2) Department of Mathematics Education, National Taichung Teachers College,
Taichung, Taiwan
906118@ms3.ntctc.edu.tw, kbc@mail.ntctc.edu.tw,
ygm@ms3.ntctc.edu.tw, koliwei@pchome.com.tw

Abstract*-Dynamic classifier selection is a strategy in multiple classifier system design. Feature extraction is one of the important procedures for mitigate Hughes phenomenon in hyperspectral image classification. Most papers have discussed the potential discriminatory information between different classifiers. In this paper, we try to exploit the discriminatory information extracted by different feature extractions for improving classification accuracy. Information is then combined by using a dynamic classifier selection strategy based on local information to make a consistency decision. This paper provides another thinking of constructing a multiple classifier system without additional classifier design by using multiple feature extraction.*

Keywords: Feature extraction, Dynamic classifier selection, Multiple classifier system.

1. Introduction

Many researches [4][5][8][9][13][14] show that combined classifier systems can outperform single classifier system. There are three basic combinations strategies: sequential combination [6], parallel combination [8], and dynamic classifier selection [5]. This study focuses on the third approach in hyperspectral data classification problem.

Typically, design of a multiple classifier system considers the potential classification information between different classifiers. However, in hyperspectral data classification, feature extraction is an important factor that influences classification accuracy greatly. In this paper, the effects of three feature extractions, principal component analysis [1], Fisher's linear discriminant analysis [3] and nonparametric weighted feature extraction [12], are explored. It is hard to decide which method is better than others. Therefore, how can we ensure that our classification system will produce optimal or suboptimal result? In this paper, we construct a multiple classifier system for combining different classifiers with different feature extractions. Although there are many different combination strategies [4][5][9][13][14], only dynamic classifier selection based on local accuracy [14] is studied.

2. Multiple Classifier System Design

The Multiple Classifier System (MCS) design cycle can be formulated as shown in Figure 1. In most papers, the ensemble overproduction focuses on the overproduction of classifiers. Since different feature extractions encapsulate complementary discriminatory information between each other. In this paper, the overproduction of both feature extraction and classifiers is considered and the step 2 and 3 are replaced by the dynamic classifier selection. We propose the following algorithm:

Ensemble Overproduction Phase I: Use different feature extraction methods to generate informative feature ensembles.

Ensemble Overproduction Phase II: Apply the ensembles obtained in Phase I to different classifiers and generate classifier ensembles.

Dynamic Classifier Selection: For each point in the testing set, K nearest neighbors in the training set are used to calculate the local accuracies of classifiers. Then the classifier with the highest local accuracy is applied to classify the testing data. Local accuracy using K nearest neighbors is defined as:

$$local\_acc(j) = \frac{\sum_{k=1}^{K} y_{ik}}{\sum_{i=1}^{C}\sum_{k=1}^{K} y_{ik}}$$

K is the number of nearest neighbors surrounding the testing point j. In this study, K is set as 3. C is the number of classifiers. i=1,…,C. y_{ik}=1 if classifier i successfully classifies the neighbor k, otherwise, y_{ik}=0.

Performance Evaluation: Evaluate algorithm performance by holdout accuracy.

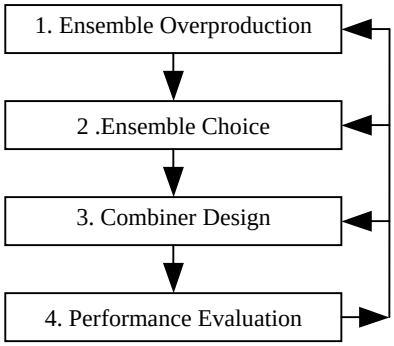

Figure 1. MCS design cycle based on the overproduce and choose paradigm. [13]

3. Feature Extractions and Classifiers

A. Feature Extractions

1. Principal Component Analysis

Principal component analysis (PCA) is defined by the transformation:

$Y = W^T X$

where $X \subseteq R^n$. W is an m-dimensional transformation matrix whose columns are the eigenvectors related to the eigenvalues computed according to the formula:

$\lambda e = Se$

S is the scatter matrix (i.e., the covariance matrix):

$$S = \frac{1}{N-1}(X-M)(X-M)^T, M = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where $x_i \in X$, i=1,…,N, M is the mean vector of X, N is the number of samples.

This transformation W is called Karuhnen-Loeve transform. It defines the m-dimensional space in which the covariance among the components is zero. In this way, it is possible to consider a small number of "principal" components exhibiting the highest variance (the most expressive features).

2. Linear Discriminant Analysis

The purpose of LDA is to find a transformation matrix A such that the class separability of transformed data (Y) is maximized. A linear transformation A from an n-dimensional X to an m-dimensional Y ($m < n$) is expressed by

$Y = A^T X$

In LDA of statistics, within-class, between-class, and mixture scatter matrices are used to formulate criteria of class separability. LDA uses the mean vector and covariance matrix of each class. A within-class scatter matrix for L classes is expressed by (Fukunaga, 1990):

$$S_w^{DA} = \sum_{i=1}^{L} P_i E\{(X-M_i)(X-M_i)^T \mid \omega_i\} = \sum_{i=1}^{L} P_i \Sigma_i$$

where P_i means the prior probability of class i, M_i is the class mean and Σ_i is the class i covariance matrix. A between-class scatter matrix is expressed as:

$$S_b^{DA} = \sum_{i=1}^{L} P_i (M_i - M_0)(M_i - M_0)^T$$

where $M_0 = E\{X\} = \sum_{i=1}^{L} P_i M_i$

The optimal criterion of LDA algorithm is to find the first m eigenvectors corresponding to the largest m eigenvalues of $(S_w^{DA})^{-1} S_b^{DA}$.

3. Nonparametric Weighted Feature Extraction

One of limitations of LDA is that it works well when data is normally distributed. A different between-class scatter matrix and a within-class scatter matrix were proposed in nonparametric weighted feature extraction (Kuo and Landgrebe, 2001; Kuo and Landgrebe, 2004) for improving this limitation. The optimal criterion of NWFE is also by optimizing the Fisher criteria, the between-class scatter matrix and the within-class scatter matrix are expressed respectively by

$$S_b^{NW} = \sum_{i=1}^{L} P_i \sum_{\substack{j=1 \\ j \neq i}}^{L} \sum_{k=1}^{n_i} \frac{\lambda_k^{(i,j)}}{n_i} (x_k^{(i)} - M_j(x_k^{(i)}))(x_k^{(i)} - M_j(x_k^{(i)}))^T$$

$$S_w^{NW} = \sum_{i=1}^{L} P_i \sum_{k=1}^{n_i} \frac{\lambda_k^{(i,i)}}{n_i} (x_k^{(i)} - M_i(x_k^{(i)}))(x_k^{(i)} - M_i(x_k^{(i)}))^T$$

In the formula, $x_k^{(i)}$ refers to the k-th sample from class i. The scatter matrix weight $\lambda_k^{(i,j)}$ is a function of $x_k^{(i)}$ and local mean $M_j(x_k^{(i)})$, and defined as:

$$\lambda_k^{(i,j)} = \frac{dist(x_k^{(i)}, M_j(x_k^{(i)}))^{-1}}{\sum_{l=1}^{n_i} dist(x_l^{(i)}, M_j(x_l^{(i)}))^{-1}}$$

where $dist(a,b)$ means the distance from a to b. If the distance between $x_k^{(i)}$ and $M_j(x_k^{(i)})$ is small then its weight $\lambda_k^{(i,j)}$ will be close to 1; otherwise, $\lambda_k^{(i,j)}$ will be close to 0 and sum of total $\lambda_k^{(i,j)}$ for class i is 1. $M_j(x_k^{(i)})$ is the local mean of $x_k^{(i)}$ in the class j and defined as:

$$M_j(x_k^{(i)}) = \sum_{l=1}^{n_j} w_{kl}^{(i,j)} x_l^{(j)}, \text{ where } w_{kl}^{(i,j)} = \frac{dist(x_k^{(i)}, x_l^{(j)})^{-1}}{\sum_{l=1}^{n_j} dist(x_k^{(i)}, x_l^{(j)})^{-1}}$$

The weight $w_{kl}^{(i,j)}$ for computing local means is a function of $x_k^{(i)}$ and $x_l^{(j)}$. If the distance between $x_k^{(i)}$ and $x_l^{(j)}$ is small then its weight $w_{kl}^{(i,j)}$ will be close to 1; otherwise, $w_{kl}^{(i,j)}$ will be close to 0 and sum of total $w_{kl}^{(i,j)}$ for $M_j(x_k^{(i)})$ is 1.

In the NWFE criterion, we regularize S_w^{NW} to reduce the effect of the cross products of between-class distances and prevent singularity by

$$0.5S_w^{NW} + 0.5diag(S_w^{NW})$$

Hence, the features extracted by NWFE are the first m eigenvectors corresponding to the largest m eigenvalues of $(S_w^{NW})^{-1} S_b^{NW}$.

B. Classifiers

Ten classifiers described in Table 1. are used to construct the multiple classifier system. All classifiers are implemented in a Matlab toolbox for pattern recognition, called PR-tools. [2].

Table 1. Classifiers used in this study.

Notation	Classifiers
qdc	Normal densities based quadratic classifier
bpxnc	Train feed forward neural network classifier by backpropagation
parzenc	Parzen density based classifier
svc	Support vector classifier
pfsvc	Pseudo-Fisher support vector classifier
loglc	Logistic linear classifier
knnc1	k-nearest neighbor classifier.(k=1)
knnc20	k-nearest neighbor classifier.(k=20)
neurc	Automatic neural network classifier
treec	Construct binary decision tree classifier

4. Data Set and Experiment Design

A. Training and Testing Data

Training and testing data sets are selected from a small segment of a 191 bands hyperspectral image data. It was collected over the DC Mall maps which have seven classes (Roof, Street, Path, Grass, Trees, Water and Shadow) are selected to form training and testing data sets. There are 100 training samples and testing samples in each class.

B. Experiment Design

The experiment design is showed in Figure 1.

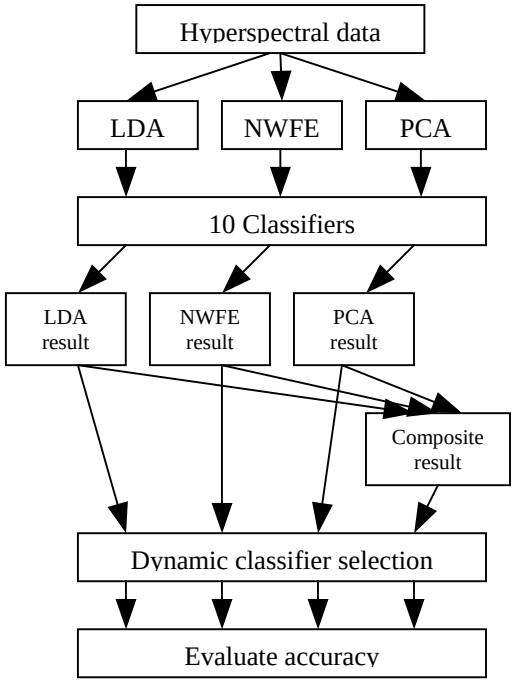

Figure 2. Experiment design in this study.

5. Results and Findings

To simply result graphs, only the performances of the top 5 single classifiers using 2 to 6 features are shown in Figure 3, 4, and 5. Figure 6 shows the classification accuracy obtained by using dynamic classifier selection strategy. Table 2 shows single classifier accuracy using different feature space, multiple classifier accuracy using different feature space, and multiple classifier accuracy using composite feature space. The best single classifier is an arbitrary choice by authors because each single classifier has different performance at different number of features.

The experimental result shows that the NWFE feature space produces better classification accuracy than LDA and PCA ones. The highest accuracy (0.939) occurs in the combination of NWFE and pfsvc. (number of features = 5)

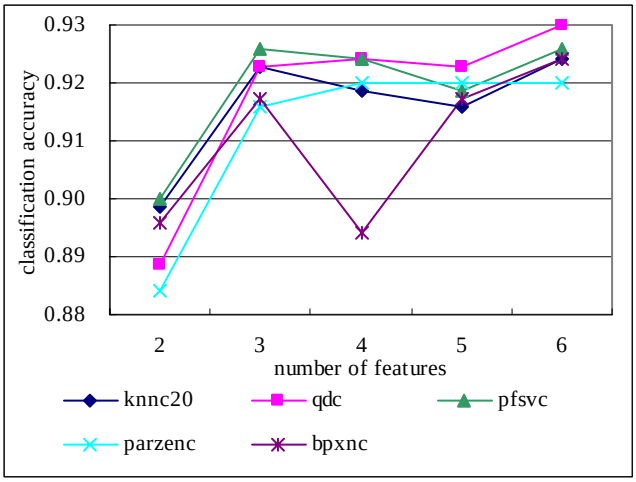

Figure 3. The performances of top 5 classifiers among 10 classifiers with LDA feature extraction.

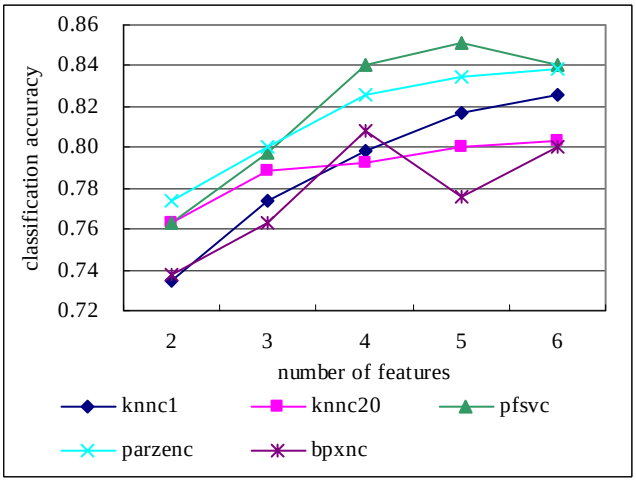

Figure 5. The performances of top 5 classifiers among 10 classifiers with PCA feature extraction.

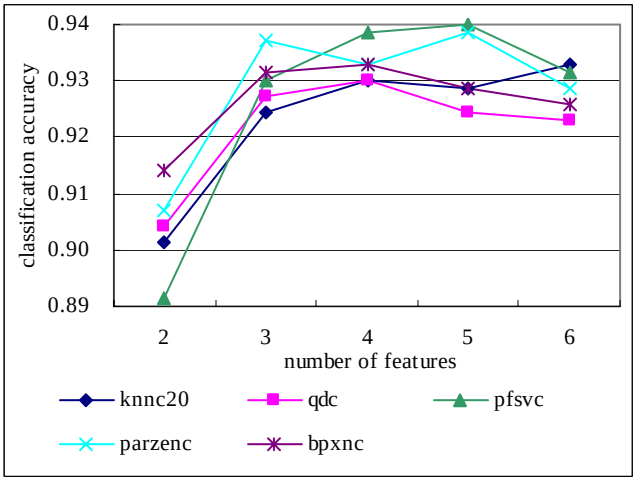

Figure 4. The performances of top 5 classifiers among 10 classifiers with NWFE feature extraction.

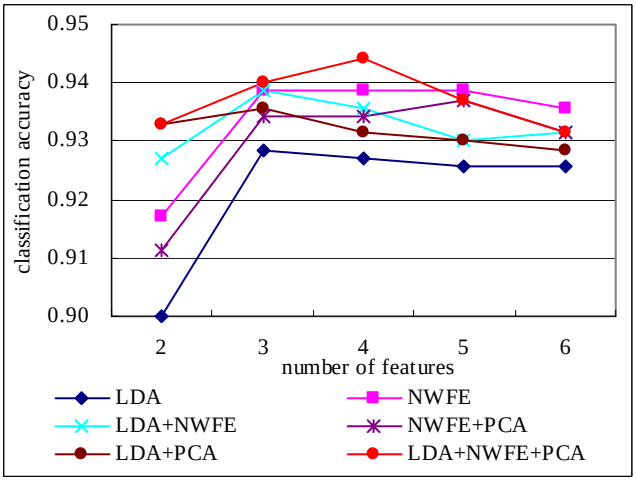

Figure 6. Dynamic classifier selection using local accuracy and 4 different feature extractions.

Table 2. Classification accuracy from dimension 1 to 6 in this study.

	Number of Features	1	2	3	4	5	6
Best single classifier	LDA (qdc)	0.616	0.889	0.923	0.924	0.923	0.930
	NWFE (parzenc)	0.821	0.907	0.937	0.933	0.939	0.929
	PCA (pfsvc)	0.596	0.763	0.797	0.840	0.851	0.840
Dynamic classifier selection	LDA	0.613	0.900	0.929	0.927	0.926	0.926
	NWFE	0.834	0.917	0.939	0.939	0.939	0.936
	PCA	0.701	0.817	0.820	0.856	0.859	0.854
	LDA+ NWFE	0.840	0.927	0.939	0.936	0.930	0.931
	NWFE+ PCA	0.837	0.911	0.934	0.934	0.937	0.931
	LDA+ PCA	0.819	0.933	0.936	0.931	0.930	0.929
	LDA+ NWFE +PCA	0.869	0.933	0.940	0.944	0.937	0.931

6. Conclusions

According to the experimental results, the conclusions can be drawn:

1. Dynamic classifier selection strategy does not guarantee of producing better accuracy than a single classifier. But it is worth noting that the dynamic classifier selection strategy ensures to produce optimal or suboptimal classification accuracy. If we are not sure about which classifier is the best, dynamic classifier selection can be used for "stabilizing" classification accuracy.

2. Experimental results show that combining feature extraction methods slightly improve classification accuracy when the number of used features is smaller than 5 (see Figure 6). When the number of features is larger than 5, the classification accuracy of proposed algorithm in this study is not as good as single classifier with single feature extraction or dynamic classifier selection with single feature extraction. In our opinion, there exists

potential discriminatory information between different feature extractions, but if the number of features is larger, the increasing noise may influence this information.

7. Acknowledgements

Authors would like to thank National Science Council for partially supporting this work under grant NSC-91-2520-S-142-001 and NSC-92-2521-S-142-003.

References

[1] P.N. Belhumeur, J.P. Hespanha and D.J. Kriegman, ” Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection, ” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(7), 711-720, 1997.

[2] R.P.W. Duin, *PRTools, a Matlab Toolbox for Pattern Recognition*, (Available for download from http://www.ph.tn.tudelft.nl/prtools/), 2002.

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego: Academic Press Inc., ch9-10, 1990.

[4] G. Giacinto and F. Roli, ”An approach to automatic design of multiple classifier systems. ” *Pattern Recognition Letters*, 22, 25–33, 2001.

[5] G. Giacinto and F. Roli, ”Dynamic classifier selection based on multiple classifier behaviour.” *Pattern Recognition*, 34(9), 179–181, 2001.

[6] N. Giusti, F. Masulli, and A. Sperduti, ”Theoretical and Experimental Analysis of a Two-Stage System for Classification, ” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(7):893–904, 1998.

[7] G.F. Hughes, ”On the mean accuracy of statistical pattern recognition,” *IEEE Trans. Information Theory*, 14(1), 55-63, 1968.

[8] J. Kittler, M. Hatef, R.P.W. Duin and J. Matas, ”On Combining Classifiers,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3), 226–239, 1998.

[9] L.I. Kuncheva, J.C. Bezdek and R.P.W. Duin, ”Decision templates for multiple classifier fusion: an experimental comparison.” *Pattern Recognition*, 34(2), 299–314, 2001.

[10] B-C. Kuo and D.A. Landgrebe, ”Improved statistics estimation and feature extraction for hyperspectral data classification,” *Technical Report*, Purdue University, West Lafayette, IN., TR-ECE 01-6, December, 2001.

[11] B-C. Kuo, D.A. Landgrebe, L-W. Ko, and C-H. Pai, ”Regularized Feature Extractions for Hyperspectral Data Classification,” *International Geoscience and Remote Sensing Symposium*, Toulouse, France, 2003.

[12] B-C. Kuo, and D.A. Landgrebe, ”Nonparametric Weighted Feature Extraction for Classification,” *IEEE Trans. on Geoscience and Remote Sensing*, 42(5), 1096-1105, 2004..

[13] F. Roli, G. Giacinto, ”Design of Multiple Classifier Systems,” *H. Bunke and A. Kandel (Eds.) Hybrid Methods in Pattern Recognition*, World Scientific Publishing, 2002.

[14] K. Woods, W.P. Kegelmeyer, K. Bowyer, ”Combination of Multiple Classifiers Using Local Accuracy Estimates, ” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 19(4), 405-410, 1997.