

APPLYING A DYNAMIC SUBSPACE MULTIPLE CLASSIFIER FOR REMOTELY SENSED HYPERSPECTRAL IMAGE CLASSIFICATION

Jinn-Min Yang

Department of Mathematics Education
National Taichung University of Education, Taichung, Taiwan, R.O.C.
jinnmin@gmail.com

ABSTRACT

The multiple classifier system has received remarkable attentions for improving the performance of a single classifier in recent years. The random subspace method (RSM) is one of the multiple classifier systems. In RSM, classifiers are trained by data set with randomly selected and fix-sized feature subsets and are combined using simple majority vote in the final decision rule. The feature subset size of the reduced data set and the fashion to construct the feature subset are two key issues affecting the performance of RSM. The former must be pre-assigned and the latter is randomly generated based on the former assignment. This study applies a dynamic subspace multiple classifier system to the classification of hyperspectral images, and investigates its performance on various conditions. The experimental results demonstrate that the dynamic subspace multiple classifier can achieves better classification results than RSM, and some important results are revealed as well in this study.

Index Terms— multiple classifier system, ensemble learning, random subspace method, curse of dimensionality.

1. INTRODUCTION

Multiple classifier system or ensemble learning [1] has received remarkable attentions for improving the performance of a single classifier in recent years. The idea of ensemble learning is to employ multiple classifiers and combine their predictions. The employed component classifier is generally called a “weak classifier” [2]-[4] because it yields unsatisfactory classification performance individually. Random subspace method (RSM) [5] is one of the well-known methods, in which the component classifiers are constructed in random subspaces of the original data feature space. The resulting classification of the component classifiers is usually combined by simple majority voting in the final decision rule. The remotely sensed hyperspectral data such as AVIRIS (Airborne Visible/InfraRed Imaging Spectrometer) data [6]-[8], generally with hundreds of measured bands, potentially provides more accurate and detailed information for

classification. However, as is well-known, the collection of ground-truth of hyperspectral image can be considerably expensive. Therefore, the so-called small sample size (SSS) problem [9] has been an important issue for hyperspectral image classification. Learning algorithms will then suffer from the problem and yield unsatisfactory classification results. In general, this problem can be tackled by feature extraction or feature selection techniques. Many studies have demonstrated that the RSM is a powerful combining technique for improving classification accuracy [10], [11], particularly for SSS problem in recent years [2].

Lam and Suen [12], [13] found that the majority vote is guaranteed to yield better performance than an individual classifier when each classifier in the ensemble has a classification accuracy great than 0.5. In addition, controlling the classification accuracy of each classifier in an ensemble is conducive to the majority vote accuracy [14], [15]. The higher the classification accuracy of each classifier is, the better the performance of the ensemble will be. Controlling the accuracies of the classifiers in the ensemble is therefore quite an essential part on constructing an ensemble, which implies that these classifiers should be trained by sounder data sets for obtaining better overall classification accuracy. Based on this consideration, we develop an ensemble with systematic mechanism to combine feature subsets automatically. More importantly, the sizes of the constructed feature subsets for classifiers in the proposed ensemble are adaptive. The rest of this paper is organized as follows. The related work of this study will be briefly introduced in Section 2. The proposing algorithm is presented in Section 3, and experimental results are demonstrated in Section 4. Finally, we conclude in Section 5.

2. RANDOM SUBSPACE METHOD

Let $\mathbf{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ denotes a training set, where the training sample $\mathbf{x}_\ell \in \mathbf{X} = \{\mathbf{x} | \mathbf{x} \in \mathcal{R}^d\}$ and its corresponding label $y_\ell \in Y = \{1, 2, \dots, L\}$. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ and $Y = \{y_1, y_2, \dots, y_N\}$ be the set of training data points and their label set, respectively. In RSM, one can randomly

draw a p -dimensional subspace from the original d -dimensional feature space to obtain a dimensionality-reduced training set $\tilde{\mathbf{D}} = \{(\tilde{\mathbf{x}}_1, y_1), (\tilde{\mathbf{x}}_2, y_2), \dots, (\tilde{\mathbf{x}}_N, y_N)\} = \{(\tilde{\mathbf{X}}, Y)\}$ where $\tilde{\mathbf{x}}_\ell \in \mathcal{R}^p$ and $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_N\}$. The procedure will be repeated for B times and then the corresponding B p -dimensional training sets are applied to construct B classifiers. The labels of a test sample $\mathbf{x}^*$ given by the B classifiers, $l_1, l_2, \dots, l_B$, are then combined by simple majority vote in the final decision. The detail of RSM algorithm is described as follows. The RSS procedure denotes the random subspace selector which outputs the indices of the selected features. C denotes the employed base classifier.

Algorithm: Random Subspace Method (RSM)

A. Training Phase

```

Begin
  for  $i = 1$  to  $B$ 
     $I_i = RSS(C, p)$ 
     $\tilde{\mathbf{D}}_i = \mathbf{D}(I_i)$ 
     $h_i = C(\tilde{\mathbf{D}}_i)$ 
  end
End

```

B. Classification Phase

$y^* = \arg \max_{c \in \{1, 2, \dots, L\}} \text{card}(i | h_i(x^*(I_i)) = c)$, where $i = 1, 2, \dots, B$

3. DYNAMIC SUBSPACE ENSEMBLE

The proposed dynamic subspace ensemble (DSE) algorithm is described as follows. There are five parameters, C , $\tilde{A}$, f_0 , Δf and ΔA , for constructing the dimensionality-reduced data set $\tilde{\mathbf{D}}_i$. The parameter C denotes the employed base classifier, $\tilde{A}$ controls the training classification accuracy of each classifier such as 0.95, and is designed to be degradable to prevent DSE from infinite loops, and f_0 denotes the size of the initial feature subset. After the initial subset is determined, one or more random features (Δf) that can contribute to the classification accuracy are sequentially added each time until the classification accuracy surpasses $\tilde{A}$. Intuitively, the DSS generates different feature subsets for the B classifier; in other words, the constructed feature subsets are adaptive. The ΔA denotes the degenerative classification accuracy each time if no feature(s) can be added to the current subset after 25 times tries.

Algorithm: Dynamic Subspace Ensemble (DSE)

A. Training Phase

```

Begin
  for  $i = 1$  to  $B$ 
     $I_i = DSS(C, \tilde{A}, f_0, \Delta A, \Delta f)$ 
     $\tilde{\mathbf{D}}_i = \mathbf{D}(I_i)$ 
     $h_i = C(\tilde{\mathbf{D}}_i)$ 
  end
End

```

B. Classification Phase

$y^* = \arg \max_{c \in \{1, 2, \dots, L\}} \text{card}(i | h_i(x_0(I_i)) = c)$, where $i = 1, 2, \dots, B$

Note that the RSM can be viewed as a special case of DSE if $\tilde{A}$ is set to a arbitrarily small value ϵ or 0, then we can have $RSS(C, p) = DSS(C, \epsilon, p, \Delta A, \Delta f)$. Under the circumstances, the training classification accuracy of the data set with the randomly selected p features can definitely surpass ϵ . Therefore, RSM becomes a special case of DSE.

The main advantages of the proposed DSE are: 1) It is a classifier-oriented and accuracy-based algorithm because the combination of feature subsets is automatically determined and is directly related to the chosen base classifier. We, therefore, do not have to take care about this problem in DSE. In addition, the classification accuracy of DSE is theoretically guaranteed to be better than a single classifier. 2) The number of features for each classifier in DSE needs not to be the same; in other words, the feature sizes of the generated data sets could be different. This is one of the avenues of future research suggested in [10]. The disadvantage of DSE is that it takes more computational time than RSM.

4. EXPERIMENTS AND RESULTS

The hyperspectral image data set, Indian Pines Site [16], is included to explore the usefulness of the proposed DSE. All the Indian Pines data samples are divided into 4757 training samples and 4588 test samples, as listed in Table 1. Experiments with three different sets containing 5%, 10% and 25% of the original training samples are carried out [17]. All experiments are repeated 20 times. The averaged testing classification accuracies and their standard deviations on the testing data sets over 20 repetitions are presented. The classifier applied in this study is the normal density based quadratic discriminant classifier (QDC) [18]. The performance of QDC will be poor when the inverse of the matrix Σ_k is singular. In this study, the value of B is set to 50 as a reasonable trade-off between the computational expense and classification accuracy. The value of $\tilde{A}$ denotes the desired classification accuracy in each classifier in DSE. Although, the higher the value of $\tilde{A}$ is, the better the classification accuracy would be. Considering the possibility of over-fitting in the small sample size cases, the

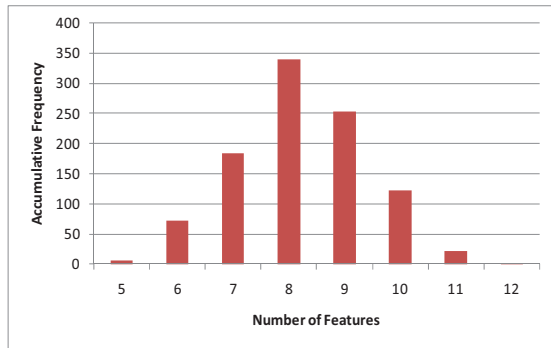
values of $\tilde{A}$ through this study are set to 0.95. The initial feature size f_0 and the degenerative accuracy ΔA is set to 2 and 0.05, respectively. The number of features to be added each time, i.e., Δf , is 1. Two other classifiers, SVM and 1NN, are included for comparisons.

Table 1: Number of training and test samples used in the Indian Pines subset scene.

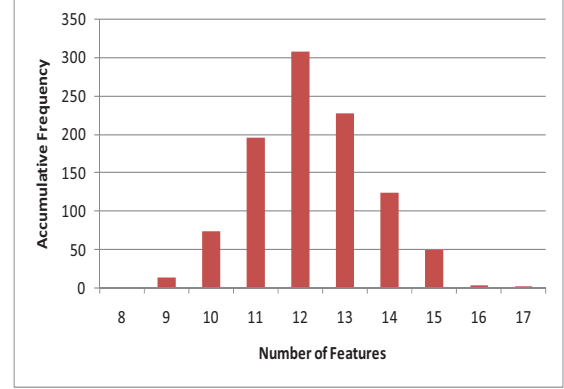
Class		Samples	
No	Name	Training	Test
1	Corn-notill	742	692
2	Corn-min	442	392
3	Grass/Pasture	260	237
4	Grass/Trees	389	358
5	Hay-windrowed	236	253
6	Soybeans-notill	487	481
7	Soybeans-min	1245	1223
8	Soybeans-clean	305	309
9	Woods	651	643
Total		4757	4588

The bar charts in Fig. 1 demonstrate the accumulative frequencies of feature subspace sizes in DSE over 20 trials on the experimental data sets. As has been stated, the value of B is 50 in each trial, the overall frequency is therefore 1000 in each plot. The x coordinate denotes the number of features in the classifiers and the y coordinate represents the accumulative frequencies with respect to the number of features. The results reveal:

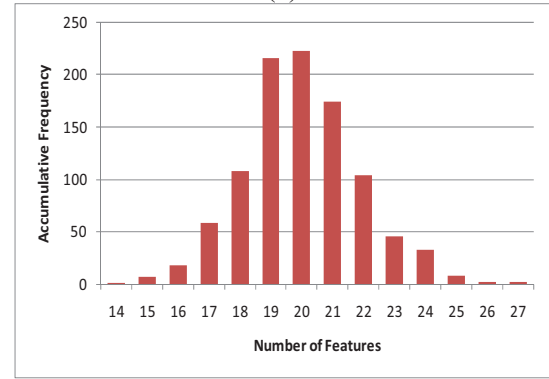
- 1) The sizes of the feature subsets are various. Fig. 1(a), for example, only few subsets with 5 or 11 features can reach the setting accuracy on the training data set, while most of the feature subset sizes vary from 7 to 10. In addition, the range is also various with respect to the training sample sizes.
- 2) Most of the frequencies are evidently concentration on the vicinity of the highest one.
- 3) The singularity problem of QDC becomes a natural shield to keep the feature size in a reasonable range.



(a)



(b)



(c)

Fig. 1 Accumulative frequencies of dimensionality of QDC classifiers in DSE over 20 trials on the experimental data sets. (a), (b) and (c) demonstrate the results of containing 5%, 10% and 25% training data, respectively.

Table 2 demonstrates the performances for SVM, 1NN, RSM (p_h) and DSE classifiers on the Indian Pines data sets in terms of overall accuracy (OA), respectively. The p_h denotes the highest frequency of features in Fig. 1. From the results, we have the following findings. 1) DSE obtains better results as compared with other methods. 2) The results of RSM with p_h features show the values of p_h given from DSE is constructive and reliable 3) Allowing different sizes of feature subsets can obtain better results than using a fixed one.

Table 2. The classification accuracies (in %) for SVM, 1NN, RSM and DSE classifiers on Indian Pines data sets.

Percentage of training samples	Overall accuracy (in %)			
	SVM	1NN	RSM(p_h)	DSE
5%	73	69.7	72.4±0.7	73.5±0.8
10%	78	71.4	78.4±0.6	79.2±0.6
25%	83.1	77	84.7±0.3	85.7±0.3

Figure 2 shows interesting findings as we give more insight into the difference of the producer's accuracy (PA)

[15] between DSE and SVM. Evidently, SVM significantly outperforms DSE on class 3 (Grass/Pasture) in all cases, achieving particularly 46.6% of the testing samples of this class in 5% case. The number of test samples of Grass/Pasture class is 237, so there are about 110 samples correctly classified by SVM, as compared with DSE. Therefore, if we integrate the two classifiers and introduce the output of SVM on the Grass/Pasture class to replace that of DSE, then the OA can be increased about 2% in 5% case. To some extent, DSE and SVM can be referred to be complementary on this data set.

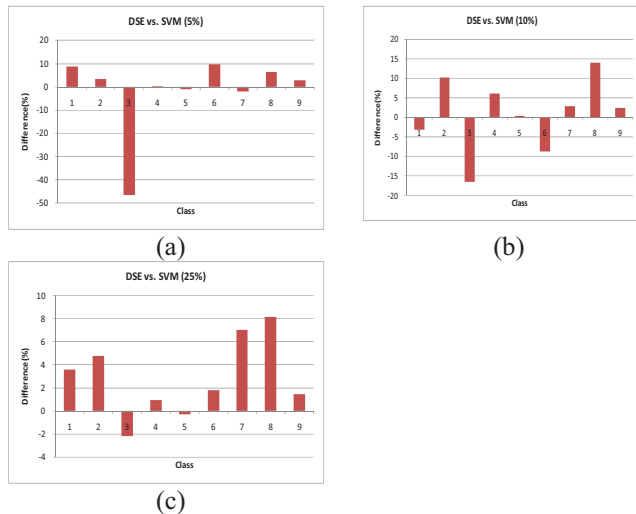


Fig. 2 Difference of the producer's accuracy (PA) between DSE and SVM on Indian Pines data sets. (a)-(c) are the results in 5%, 10% and 25%, respectively.

5. CONCLUSION

In this study, an accuracy-based and classifier-oriented multiple classifier system DSE is applied for hyperspectral image classification, in which the component classifiers are trained by the data sets with automatically combined feature subsets. Unlike RSM, one has to determine a suitable dimensionality of the random feature subsets by his experience and the randomly selected feature subsets might not provide sound discriminant information. Importantly, DSE allows different-sized feature subsets and obtains better results than RSM which is with a fixed size. In addition, when the QDC classifier is employed as the base classifier, the singularity problem of the QDC becomes the natural shield to keep DSE away from selecting too many features.

REFERENCES

- [1] T. G. Dietterich, "Ensemble learning," *The Handbook of Brain Theory and Neural Networks*, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, 2002. 405-408.
- [2] M. Skruichina and R.P.W. Duin, "Bagging, boosting and random subspace method for linear classifiers," *Pattern Analysis & Applications*, vol. 5, no.2, pp.121-135, 2002.

- [3] T.G. Dietterich, M. Kearns, and Y. Mansour, "Applying the weak learning framework to understand the improve c4.5," *Proceedings 13th International Conference on Machine Learning*. Morgan Kaufmann Publishers, 1996, pp. 96-104.
- [4] C. Ji, and S. Ma, "Combinations of weak classifiers," *IEEE Transactions on Neural Networks*, vol.8 no.1, pp.32-42, 1997.
- [5] T.K. Ho, "The random subspace method for constructing decision forests," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.20, no.8, pp. 832-844, 1998.
- [6] R. O. Green, et al., "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sensing of Environment*, vol. 65, pp.227-248.
- [7] P. J. Curran, and J.L. Dungan, "Estimation of signal-to-noise: A new procedure applied to AVIRIS data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 27, no. 5, pp. 620-628, Sep. 1989.
- [8] J. A. Benediktsson, J.R. Sveinsson and K. Arnason, "Classification and feature extraction of AVIRIS data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 33, no. 5, pp. 1194-628, Sep. 1995.
- [9] S.J. Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol.13 no.3, pp. 252-264, 1991.
- [10] S.D. Bay, "Nearest neighbor classification from multiple feature subsets," *Intelligent Data Analysis*, vol.3 no.3 pp.191-209, 1999.
- [11] L. Breiman, "Random forests-random features," Technical Report 567, Statistics Department, University of California, Berkeley, CA 94720, Sep. 1999.
- [12] L. Lam and C.Y. Suen, "Optimal combination of pattern classifiers," *Pattern Recognition Letter*, vol. 16, pp.945-954, 1995.
- [13] L. Lam and C.Y. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 27, no.5, pp.553-568, 1997.
- [14] L.I. Kuncheva, *Combining Pattern Classifiers*, John Wiley & Sons, New Jersey, 2004.
- [15] D.A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: Wiley, 2003.
- [16] AVIRIS NW Indiana's Indian Pines 1992 data set. [Online]. Available: <ftp://ftp.ecn.purdue.edu/biehl/MultiSpec/92AV3C> (original files) and ftp://ftp.ecn.purdue.edu/biehl/PC_Multispec/ThyFiles.zip (ground truth).
- [17] A. Plaza, J. A. Benediktsson, J. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, J. C. Tilton, and G. Triani, "Recent advances in techniques for hyperspectral image processing," *Remote Sensing of Environment*, supplement 1, vol. 113, pp.110-122, 2009.
- [18] F. van der Heiden, R. P. W. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*, Chichester: John Wiley & Sons, 2004.

ACKNOWLEDGEMENT

This study is supported by National Science Council, R.O.C., under the contract number of NSC 100-2218-E-142 -002 -MY2.