# Research on forage hyperspectral image recognition based on F-SVD and XGBoost

*Xuanhe* Zhao[1] , *Xin* Pan[1] , *Yubao* Ma[2] , and *Weihong* Yan[2,*]

[1]College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot, 010018, China
[2]Institute of Grassland Research of CAAS, Hohhot, 010010, China

**Abstract.** Aiming at the high time complexity and poor accuracy of traditional SVD in hyperspectral recognition. we proposed F-SVD, which introduces the latent factors(F) into the SVD decomposition strategy and uses the correlation between the latent variable and the original variable to improve the singular matrix. Firstly, we used F-SVD to reduce the dimension of visible-near infrared hyperspectral image, and consequently designed a forage recognition model based on XGBoost. When the test set sets 40%, the OA of F-SVD-XGBoost is 91.67%, which takes 0.601s. Compared with the traditional FA-XGBoost and SVD-XGBoost, OA increases 1.98% and 1.67%, and the time consumption decreases 1.369s and 0.522s, respectively. The results show that our model not only effectively extracts the essential features of forage hyperspectral and improves the accuracy of classification, but also has a faster processing speed, so that can efficiently and quickly realize the identification of forage hyperspectral images.

## 1 Introduction

Forage identification is great significance for dynamic monitoring of grassland health evaluation and grass yield estimation. Recently, with the continuous innovation of hyperspectral technology and related theories, it has widely used in the grassland ecological and other fields [1]. The hyperspectral image has the characteristics of unification of map, high dimension and multiple bands [2]. However, the great number of hyperspectral data which increases the calculation and reduces the generalization ability of the classifier. So in the classification of hyperspectral images, data reduction is a very eventful link.

Feature extraction is based on the correlation between features and targets, mapping high-dimensional data into low-dimensional space [3]. Among them, factor analysis (FA) and singular value decomposition (SVD) are commonly used methods. Shen et al [4] proposed FA-BPNN, but requires more sensitive spectral bands to reconstruct the original data. Huang et al [5] used the SVD to directly extract the singular values of bands to achieve the classification, but the accuracy needs to be further improved. Guo et al [6] proposed that based on multi-label shared subspace learning. They used SVD iterative,

* Corresponding author: yanweihong@163.com

which improved the classification accuracy at a low sample rate. But with the expansion of training samples, the anti-noise ability is not high.

SVD can operate on arbitrary matrices, but requires a lot of calculation and time. Therefore, we proposed the fusion of FA and SVD (F-SVD), which can reduce time consumption and enhance feature expression to improve the performance of dimensionality reduction. In this paper, we used F-SVD for feature extraction of hyperspectral images, and then used XGBoost to complete accurate and rapid identification of pasture.

## 2 Preliminary knowledge

### 2.1 FA algorithm

FA discovers a few independent latent variables to reflect many original factors [7]. Suppose the number of samples and bands are *n* and *q*, then the original variable $X$ is

$$X = (x_1, x_2, \cdots, x_q)' \tag{1}$$

and the common factor is

$$F = (F_1, F_2, \cdots, F_h)(h < q) \tag{2}$$

Group $X$ and represent the original data with *h* common factors, the form is

$$X = AF + \varepsilon \tag{3}$$

where *F* is a common factor and $\varepsilon$ is a special factor. *A* is the factor load matrix. Rotate *A* to find more obvious common factors.

### 2.2 SVD algorithm

SVD decomposes the hyperspectral matrix to obtain the singular value matrix and the vector matrix, arranges them according to the importance of the singularity. Selecting a part of the best values and vectors to reconstruct the image, which are reduced to the feature subsets of smaller and more relevant [8]. SVD of hyperspectral matrix $X_{n\times q}$:

$$X_{n\times q} = U_{n\times n} \sum_{n\times q} V_{q\times q}^{T} \tag{4}$$

where *U* and *V* are $n\times n$ and $q\times q$ dimensional orthogonal matrices respectively, *T* is the transpose of the matrix; $\Sigma$ is the singular value of the diagonal matrix.

### 2.3 XGBoost

Exterme Gradient Boosting (XGBoost) can automatically perform parallel computing. Its characteristic for hyperspectral classification is that can intuitively classified according to features, and explained physically. It expands the objective function Taylor to the second order, and adds an L2 regularization term to the loss function to avoid model overfitting [9].

## 3 Proposed model

This paper proposed a grassland hyperspectral image recognition method based on F-SVD and XGBoost. It includes the following steps:

Step 1 Use FA to extract the latent variable features of the original data and the internal common factors between the variables. Select factors whose contribution rate are greater than 95% to obtain d feature value, and use that as the output data Y.

Step 2 we perform SVD decomposition on the matrix of d dimensional feature vectors, and select the first l principal components to obtain low-dimensional data T.

Step 3 T is used as the input matrix of the XGBoost, the booster is gbtree, and finally obtained the classification result with the best generalization ability via OA(Overall Accuracy), Kappa coefficient, test time, and F1-score to evaluate.

## 4 Experiments and results

### 4.1 Dataset

We constructed a forage hyperspectral database to testify the model, and all samples are from the Institute of Grassland Research of CAAS and the artificial grassland in the College of Grassland and Resources Environment of Inner Mongolia Agricultural University was grown by a team of professor Han. Using HyperSpec © PTU-D48E developed by Opter to collect data from 10:00 to 13:00. When there is sufficient sunlight and no wind, as shown in Fig.1. The spectrum ranges from 400 to 1000nm, 750 channels, line scanning, and including 125 bands, that images resolution is 1166×1004 pixels. It is necessary to perform radiation correction every ten minutes to avoid the influence of external sun, wind and other factors [10].


**Fig. 1.** HyperSpec©PTU-D48E.

We use ENVI5.3 to extracte the ROI with clear imaging and a large reflection spectrum, then calculate the average value of spectral reflectance. 16 species of grassland with 5 hyperspectral images per specie, 50 ROI/specie, total 800 samples. Species C1-C16 are .Melissilus ruthenicus(L.) Peschkova(TrigonellaruthenicaL.), Medicago Sativa Linn, Hordeum brevisubulatum(Trin.) Link, Medicago ruthenica(L.) Trautv., Melissilus ruthenicus(L.) Peschkova(TrigonellaruthenicaL.), Agropyron michnoi, Trifolium repens L., Agropyron cristatum(L.) Gaertn. var. pectiniforme(Roem. et Schult.) H. L. Yang, Bromus ciliatus L., Lespedeza bicolor Turcz, Medicago falcata L., Elymus sibiricus Linn, Bromus inermis Leyss., Agropyron mongolicum Keng, Avena sativa L. and Festuca rubra L..

## 4.2 Analysis of Results

### *4.2.1 Data preprocessing*

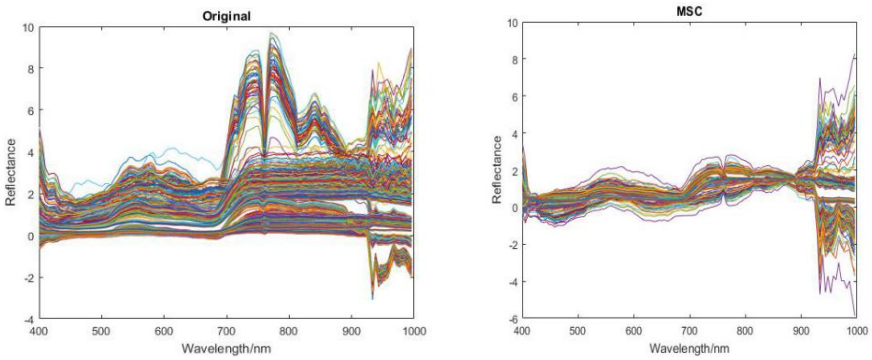
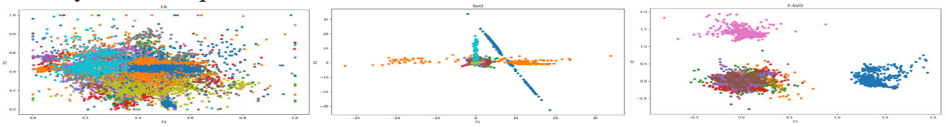



**Fig. 2.** The average reflectance spectrum before and after MSC treatment.

Due to the uneven distribution of the grass sample surface, a certain scattering effect covers the spectral changes. The multivariate scattering correction (MSC) is used to correct the baseline shift and offset phenomena of the spectral data through the standard spectrum for denoising, as shown in Fig.2. After processed by MSC, the convergence degree of the average reflection spectrum curve is enhanced, and the correlation between the spectrum and the data is increased, which can effectively eliminated the spectral difference caused by different scattering levels and improve the signal-to-noise ratio.

### *4.2.2 Feature extraction*

As shown in Fig.3, FA only a few data are linearly distributed, most are discretely distributed, with obvious overlaps and no feature clustering. It takes 1.970s. During SVD, the range of components is 1~30.When set to 10 is the optimal value. The linear distribution of data increases, has a certain aggregation ability. It takes 1.123s. For F-SVD, we used FA to select the potential factor with a cumulative contribution rate of 95%. Then, performed SVD. After 30 different parameter combinations, which had best effect when extracted the first 12 potential factors and components seted to 7. That the distance within the specie is small and between species is large, the clustering effect is obvious, and the separability between species is increased. It takes 0.601s.



(a) (b) (c)

**Fig. 3.** Visualization results of FA(a), SVD(b) and F-SVD(c) dimensionality reduction.

Above, FA and SVD have the flaw of large memory usage and long consumption time; F-SVD can reduce the consumption time, and get the more comprehensive feature. Compared with SVD , the running time decreased 0.522s.

### *4.2.2 Recognition model*

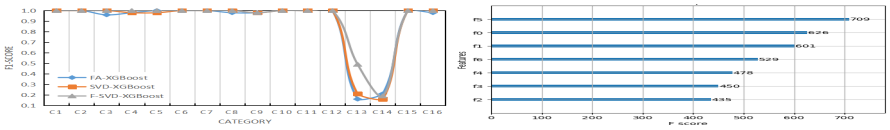
We recorded labels of samples and then randomly shuffle. The dataset was reasonably divided according to cross-validation, and the test set was set to 30%, 40% and 50%. Using XGBoost to establish recognition model, which employs the XGBClassifier function, and optimizes parameters by GridSearch. The optimal parameters are objective: multi softmax, max_depth: 3, learning_rate: 0.1, subsample: 0.6, n_estimators: 100, reg_alpha: 1e-05, nthread: 4, colsample_bytree: 0.6, seed: 27 and num_class: 16.

Table 1 shows models achieved best results when the test set was set to 40%. F-SVD-XGBoost is superior others, that OA increased by 1.98% and 1.67%, the running time reduced by 1.369s and 0.522s, respectively. And it can effectively solve the problem of data redundancy by decreasing the misclassification phenomenon caused by foreign objects of the same spectrum and reducing the number of misclassification samples.

**Table 1.** Recognition results of different algorithms.

| Classification model | Test size=0.3 | | | Test size=0.4 | | | Test size=0.5 | | |
|---|---|---|---|---|---|---|---|---|---|
| | OA /% | Kappa | Testing time /s | OA /% | Kappa | Testing time /s | OA /% | Kappa | Testing time /s |
| FA-XGBoost | 89.17 | 0.88 | 1.045 | 89.69 | 0.89 | 1.970 | 89.00 | 0.88 | 1.209 |
| SVD-XGBoost | 89.58 | 0.89 | 0.753 | 90.00 | 0.89 | 1.123 | 89.00 | 0.88 | 0.620 |
| F-SVD-XGBoost | 90.00 | 0.89 | 0.657 | 91.67 | 0.91 | 0.601 | 90.25 | 0.90 | 0.576 |

Fig. 6(a) shows F1-score of $C_i$. The F1-score of others are all above 0.95; C13 and C14 in FA/SVD-XGBoost are 0.16, 0.21 , in F-SVD-XGBoost is 0.19, 0.49, because the internal structure spectra of them are similar and easy to confus. In contrast, the proposed method is relatively robust to the differences in spectral features. The important scores for extracting 7 features by the proposed method are shown in (b), where f5 is the most great feature.




(a) (b)

**Fig. 6.** F1-score recognized by various Ci (a) and importance ranking of new features (b).

## 5 Conclusion

In this paper, we proposed F-SVD to reduce the dimensionality of hyperspectral images, and established a forage recognition model based on XGBoost. The result shows that F-SVD-XGBoost not only can effectively reduce the number of end-element features, make better classification accuracy, but also maintain a high processing speed. Compared with FA-XGBoost and SVD-XGBoost, that our model can improve performance effectively.

This work was supported by National Natural Science Foundation of China under Grant 61962048 and 61562067, Central Public-interest Scientific Institution Basal Research Found under Grant 1610332020020.

## References

1. F. F. GUO, J. J. Xiao, J. R. Fan, Detection and classification of grassland with stellera chamaejasme through HJ-1A hyperspectral image in northern Tibet, ICEESD, 6 (2017)
2. H. Huang, Z. Y. Li, Y. S. Pan, Multi-Feature Manifold Discriminant Analysis for Hyperspectral Image Classifification, Remote Sens, **11**, 24 (2019)
3. B. DU, L. F. Zhang, L. P. Zhang, W.B. Hu, Discriminant Manifold Learning Approach for Hyperspectral Image Dimension Reduction, Acta Photonica Sinica, **42**, 5 (2013)
4. W. Shen, Y. Li, W. Feng, H. Zhang, Y. Zhang, Y. Xie, Inversion model for severity of powdery mildew in wheat leaves based on factoranalysis-BP neural network, Transactions of the CSAE, **31**, 8 (2015)
5. M. Huang, Q. B. Zhu, Feature Extraction of Hyperspectral Scattering Image for AppleMealiness Based on Singular Value Decomposition, Spectroscopy & Spectral Analysis, **31**, 4 (2011)
6. L. Q. Guo, Q. C. Meng, Space Spectrum Classification Algorithm Based on Multi-label Shared Subspace Learning and Kernel Ridge Regression.Acta Photonica Sinica, **49**, 12 (2020)
7. W. N. Ismail, M. M. Hassan, H. A. Alsalamah, G. Fortino, CNN-Based Health Model for Regular Health Factors Analysis in Internet-of-Medical Things Environment, IEEE Access, **8**, 8 (2020)
8. W. H. Alshoura, Z. Zainol, J. S. Teh, M. Alawida, A New Chaotic Image Watermarking Scheme Based on SVD and IWT, IEEE Access, **8**, 16 (2020)
9. R. Song, T. Li, Y. Wang, Mammographic Classification Based on XGBoost and DCNN With Multi Features, IEEE Access, **8**, 10 (2020)
10. S. D. Fabiyi, H. Vu, C. Tachtatzis, P. Murray, S. Marshall, Varietal Classification of Rice Seeds Using RGB and Hyperspectral Images, IEEE Access, **8,** 12 (2020)