Hyperspectral Feature Selection Ensemble for Plant Classification

Ali AlSuwaidi*, Charles Veys, Martyn Hussey, Bruce Grieve, and Hujun Yin*
School of Electrical and Electronic Engineering
The University of Manchester
Manchester, M13 9PL, UK
*E-mail: ali.bghalsuwaidi@postgrad.manchester.ac.uk , hujun.yin@manchester.ac.uk

***Abstract*—Interest in hyperspectral imaging systems has increased recently substantially for studying and monitoring plant properties and conditions. The numerous financial (i.e. improve breeding process) and environmental (i.e. reduce usage of herbicide) advantages of such systems have been a driving force behind the latest surge. This paper aim to differentiate different plant species using hyperspectral image analysis. Main contribution of the work lies in the use of combined output of multiple feature selection algorithms, as compared to the use of single feature selection algorithm. Two independent hyperspectral datasets, captured by different instrumentations, were used in the evaluation. In total, six different feature selection algorithms (relief-f, chi-square, gini index, information gain, FCBF, and CFS) were used in the experiment. Experimental results show significant improvements in classification accuracy with the ensemble version of multiple feature selection algorithms compared to with the individual feature selection algorithms.**

Keywords—ensemble learning; feature selection; hyperspectral imaging; support vector machine

I. INTRODUCTION

The development of hyperspectral imaging system has prompted a great number of innovative scientific quests [1]. Hyperspectral imaging, a branch of multivariate imaging [2], captures spectral and optical properties of an object combining spectroscopy and imaging technologies. These are three main configurations for acquiring hyperspectral images [3]: point, line, and plane scanning. The popularity of this imaging branch has increased recently due to the benefit of sensing a wider range of electromagnetic spectrum as well as gathering a large number of narrow band spectra [3]. Hyperspectral imaging has been used in a growing number of applications such as, medical imaging [4], agricultural monitoring [5, 6], and industrial [7] and chemical processes [8].

The need for an efficient and effective analysis method has been elevated along the advances in hyperspectral imaging system as large amount of data is being generated and it is difficult to analyse the information directly from pixel values. Machine learning is considered as one of the effective analysis tools. Feature selection and ensemble learning are two widely used techniques, designed to achieve better generalisation performance [9].

The process of selecting a relevant subset of features, wavelengths in this case, and removing irrelevant and redundant ones based on certain evaluation criteria is called feature selection [10, 11]. It aims to enhance the performance using an optimal subset (relevant features) compared to the use of the entire feature set. The meaning of each feature needs to be fully understood in order to select appropriate features. A considerable amount of literature has been published on feature selection. These studies cover dependent and independent evaluation criteria, feature selection models, applications, and introduced new algorithms [10-12].

Ensemble learning has been studied for over 20 years with the primary goal of improving prediction performance [13, 14]. Several learning systems are combined, either weighted or unweighted, to obtain improved predictive performances compared to the single learning systems. The main advantage of such an approach is minimising the risk of selecting a least preforming learner. In other words, the performance of combined outputs from multiple learners may or may not outperform the most superior learner, but is usually better than the average, or typical single learner.

This work focuses on classifying different crops based on hyperspectral images using a combination of feature selection algorithms. It compares the results with that of using single feature selection methods. It has been shown that using relevant features improves performance of ensemble learning approach [9]. Only the best feature selection algorithms are used in each stage and passed to the final stage. Support vector machine (SVM) classifier is used as it deals with curse of dimensionality problem effectively [15], hence reducing the risk of overfitting. The results show significant improvements in accuracy over the individual selection algorithms.

The reminder of the paper is organised as follows. An overview of feature selection and ensemble learning are given in section II. Section III presents the imaging system, datasets and corresponding analysis. The experimental results and discussions are given in section IV and followed by the concluding remarks in section V.

II. BACKGROUND

An overview of feature selection and ensemble learning is given in this section. In addition, examples of both techniques are highlighted.

A. Feature Selection

Large amount of data is being generated with the recent hyperspectral imaging systems. However, the entire collected data does not necessarily contain useful information to the problem investigated. Feature selection minimises the feature space into relevant features only to have better predictive performance than the entire features. There are several methods to identify the relevancy of feature x_i from a feature space $X = \{x_1 \dots x_n\}$ for a specific dataset D with different classes $C = \{c_1 \dots c_n\}$. For example, if two classes $\{c_1, c_2\}$ in D can be distinguished using single feature x_i, then this feature is considered as relevant and vice versa.

Feature selection process can be described in four steps: search organisation, subset evaluation, stopping criteria, and result validation [10-12]. The first two steps are responsible for generating different subsets and evaluate the goodness of the generated subsets individually based on a specific evaluation criterion such as distance, information, dependency, consistency, and accuracy. On the other hand, the last two steps determine when the process should be halted (i.e. reaching threshold) and the significance of the selected subset or the stopping criterion.

Feature selection can be broadly categorized, based on evaluation criteria, into the filter, wrapper, and embedded models [10-12, 16]. The main difference between these three models is their dependence to the classification algorithm; the filter method is independent and the remaining two are dependent. In other words, a filter model depends on data characteristics to rank entire features and then selects the relevant ones among them, while classification algorithms are essential in the last two models to identify a subset of relevant features. It should be noted that performance of feature selection techniques varies due to the ability of individual technique to discard redundant or irrelative features. Moreover, the embedded model was introduced to utilise both filter and wrapper models, i.e. rank features based on their data characteristics and evaluate their goodness through classification algorithms. In addition, the wrapper and embedded models are easy to implement, whilst filter model can produce acceptable to good performances in short time [16]. The pseudo code of generalised feature selection is presented in table I.

There is a large number of published studies introducing various feature selection algorithms such as ReliefF, chi-square, gini index, information gain, fast correlation based filter (FCBF), and correlation based feature selection (CFS) [17-19]. ReliefF uses the distance between the instances to identify features relevancy [20]. It is an extension of relief version to handle multiple classes. ReliefF evaluation criterion $J(x_i)$ is used to update weight vector and is given in:

$$J(x_i) = \sum_{x_i \in NH(C_1)} -d(x_i - x_{NH}) + \sum_{C_r \neq C_1} \frac{P(C_r)}{1 - P(C_1)} \sum_{x_j \in NM(C_r)} d(x_i - x_{NM}) \quad (1)$$

where $d(.), C_1, C_r, P(.), x_{NH}, x_{NM}$ represents distance function, class 1 and the remaining classes, probability, nearest hit, and -miss respectively. Chi-square is another method used to test the independency between the features and class labels. Its evaluation criterion can be described as:

TABLE I. GENERALISED FEATURE SELECTION ALGORITHM

General Filter Algorithm
Requirements:
D: data set
X: feature space with x_i feature
Y: label
J: evaluate criteria
CA: classification algorithm % Wrapper & Embedded
Feature Selection Process:
$S_{initial}$: initialise feature subset (empty or full)
Start:
S_{new}: add or remove feature to the subset based on $S_{initial}$
Evaluate(S_{new}, D, J)
Evaluate(S_{new}, D, CA)
If $J(S_{new})$ is better than $J(S_{initial})$
If CA($\mathbf{S_{new}}$) performance is better than CA($S_{initial}$)
$S_o = S_{new}$: optimal subset
Repeat if stopping criteria == false
end

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{(x_{ij} - \mu_{ij})^2}{\mu_{ij}} \quad (2)$$

where μ_{ij} represents the expected value. Moreover, gini index (3) is a quantifying method to measure the ability of individual feature to distinguish between classes.

$$Gini = 1 - \sum_{i=1}^{C} P[(i|X)]^2 \quad (3)$$

where $P(i|X)$ is conditional propability of class i given set of samples. Information gain is widely used as an individual evaluation criterion or as preliminary measure for other criteria. It is measuring the uncertainty and based on Shannon's entropy:

$$H(X) = \sum_{i=1}^{n} P(x_i) \log_2 (P(x_i)) \quad (4)$$

FCBF uses information gain to measure symmetrical uncertainty (SU), which helps to minimisee feature bias. SU can be measured using:

$$SU = 2 \times \left[\frac{I(X,Y)}{H(X) + H(Y)} \right] \quad (5)$$

where $I(X,Y) = H(X) - H(X|Y)$ and represents information gain. CFS evaluates the output of SU based on (6):

$$Merit_S = \frac{\overline{nr_{cf}}}{\sqrt{n + n(n-1)\overline{r_{ff}}}} \quad (6)$$

where $Merit_S$ represents the heuristic merit of a subset containing n features, $\overline{r_{cf}}$ represents the average feature-class correlation, $\overline{r_{ff}}$ represents the average feature-feature correlation.

B. *Ensemble Learning*

Ensemble learning can be defined as the process of combining different learning algorithms, either weighted or unweighted, to obtain enhanced predictive performances when compared to single learning systems. The diversity of the learning algorithms that generate uncorrelated error patterns need to be promoted in the ensemble learning system in order to improve the performances [14, 21]. The diversity can be ensured either using resampling techniques or using different features. The former includes varying the parameters of the learning techniques or varying learning algorithms, while different sets of features are used to train each learner in the latter.

Ensemble diversity can be measured by several quantitative methods [14, 21]. Correlation and Q-statistics are two examples to measure the diversity through the probabilities.

$$\rho = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \qquad (7)$$

$$Q = \frac{ad - bc}{(ad + bc)}$$

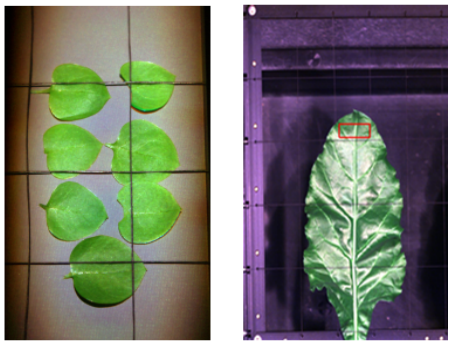
where *a,d* represent the ratios of the samples that are correctly and wrongly classified by both learning algorithm to the total number of samples respectively. *b,c* represent the ratios of the samples that are correctly classified by at least one learning algorithm to the entire samples, and the summation of them is unity, i.e. $a+b+c+d=1$.

Different methods can be used to construct ensembles [14]. Some of these methods are general, using any learning algorithm, while the others are limited to particular learning algorithms. Bagging is one of the simplest ensembles to construct, where different input subsets are randomly replicated from the original input dataset with the goal of increasing diversity. Boosting is a general ensemble that boosts the performance of the weak learning algorithms. Stacking is a very old method and thus less preferable to bagging and boosting, as there is no standard procedure to implement it. This method is divided into two stages: ensemble and learning algorithm stages. It should be mentioned that the final result is achieved by combining the entire outputs of each leaner (averaging or majority voting).

III. Materials and Method

The datasets were taken by two hyperspectral imaging systems; University of Manchester (UoM) system and University of Bonn (Bonn) system. The former system generates effective image size of 5184×3456 pixels and operates over both the visible (VIS), 380 – 720 nm, and near infrared (NIR), 730 – 1000 nm, regions with spectral resolution of 5 nm. The system operates in a controlled environment (dark box) in order to minimize the effect of unwanted noise. The latter system is a line scanning system with 2.8 nm spectral resolution and a spatial resolution of 0.29 mm [22]. It operates over the range of 400 to 1000 nm and maximum effective line that can be achieved with system is 1600 pixels. More details about the orientation and pre-processing of this system can be found in [22].

Two datasets (UoM and Bonn) with different acquisition dates and exposure times were used for the analysis. The UoM dataset consists of weeds and non-weed species (potato, wheat, and oil seed), while the Bonn dataset contains healthy and unhealthy (Cercospora and rust) sugar species. Both datasets were prevented from saturation (dynamic range management) and were spectrally normalized using the reference grey tile included in the image scene and the reference white (barium sulphate) tile, respectively. Fig. 1 shows samples of the UoM and Bonn datasets.



(a) (b)

Fig. 1. Plant leaves sample. Non-standard RGB representation of (a) UoM steme weed captured on 2015 (b) Bonn healthy sugar capture on 2013

Fig. 2 illustrates the schematic overview of the proposed method. An N-fold validation is applied to feature selection (FS), resulting N feature pools. Each FS pool consists of six feature selection algorithms (discussed in section II). In each pool only the best performing algorithm is retained and passed on to the ensemble. That means each pool is working as a switch, and selecting a single algorithm only. Thereafter, the selected methods, deemed to be the best from each pool, are used to produce the combined performance (i.e. majority voting in this case).

IV. Experimental Results and Discussion

The experiment assessed the utility of the proposed method. The utility was determined by comparing the result of the proposed ensemble method with the result of the existing feature selection algorithms. The procedure of this experiment was divided into four steps.

First, leaf reflectance extraction: averaging twenty pixels from leaf area to reduce the variation in pixels intensities and get the spectral signature. Second, selecting relevant wavelength: employing feature selection algorithms. Third, using the proposed method to select the best performing feature selection algorithms in each pool then combining them and last, comparing the classification performance. Table II shows the classification results of the proposed method as well as the individual feature selection algorithms. The percentage of training and testing set was 1:2 to the entire datasets. In addition, 10-fold cross validation was used, and the classification rates were the average of 100 runs.

The classification rate of the proposed method has been shown an improvement of 3.3% on UoM dataset and 0.46% on Bonn dataset compared to the best individual feature algorithms. In addition, a statistical test (p-value) was performed to determine the significance of the proposed method. P-value at 1% significance level is 2.36. The calculated p-value of the pro-

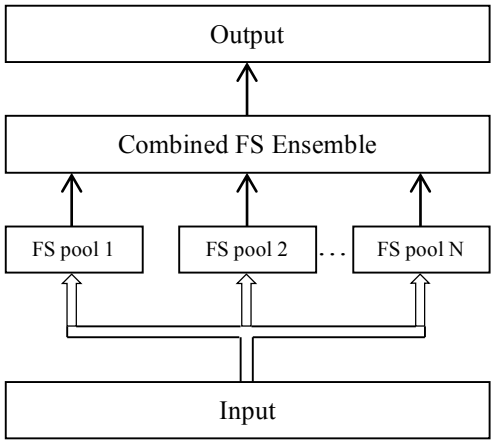


Fig. 2. Schematic diagram of the proposed method. Single arrow represents selection of the best feature selection algorithm at each pool.

-posed method was 7.07 and 3.29 for UoM and Bonn datasets, respectively, compared to the best performing feature selection algorithm (gini), thus the improvement is significant

V. CONCLUSION

The results of our experiment agree with previous studies on using ensembles for improving the prediction performance [9]. This finding assisted us in proposing a feature selection ensemble where the performance of different feature selection algorithms was combined. The prediction performance of the proposed method has shown an improvement of 3.3% and 0.46% on hyperspectral datasets (i.e. UoM and Bonn dataset respectively) compared to the best individual feature algorithm. Moreover, the statistical test has shown the significance of the proposed method with 1% significance level. The results of the proposed method indicate that combining the feature selection methods tends to be more robust and generally provides better performances than any individual algorithm. Future studies will explore weighted version of the proposed ensemble to investigate further improvement in prediction performance.

ACKNOWLEDGMENT

The authors would like to thank Anne-Katrin Mahlein and University of Bonn for provision of the Bonn dataset.

TABLE II. PERFORMANCE AND CAMPARISON OF PROPOSED METHOD

Method	Average Classification rate (%) (std)	
	UoM dataset	Bonn dataset
All	60.56 (0.97)	73.18 (4.31)
CFS	62.06 (5.80)	88.23 (2.74)
FCBF	59.71 (6.27)	72.79 (2.38)
χ^2	67.40 (5.68)	88.11 (1.48)
Gini	73.42 (6.97)	90.30 (1.77)
Info Gain	67.37 (5.68)	88.35 (1.64)
Relief-F	72.88 (6.17)	89.37 (1.55)
Proposed ensemble	**76.75 (4.71)**	**90.76 (1.40)**

REFERENCES

[1] A. F. H. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, “Imaging spectrometry for earth remote sensing,” Science, vol. 228, no. 4704, pp. 1147-1153, 1985.

[2] H. F. Grahn and P. Geladi, Techniques and Applications of Hyperspectral Image Analysis. West Sussex, England: John Wiley & Sons Ltd, 2007.

[3] P. L. M. Geladi, H. F. Grahn, and J. E. Burger, "Multivariate images, hyperspectral imaging: Background and equipment," Techniques and Applications of Hyperspectral Image Analysis, pp. 1-15: John Wiley & Sons, 2007

[4] Ali B Gh Alsuwaidi, “De-Noising Hyperspectral Images,” MSc. dissertation, Dept. Elect. Eng., University of Manchester, Manchester, UK, 2011.

[5] Y. Zhang, D. C. Slaughter, and E. S. Staab, “Robust hyperspectral vision-based classification for multi-season weed mapping,” ISPRS J. Photogramm. Remote Sens., vol. 69, pp. 65–73, 2012.

[6] H. Liu, C. Saunders, and S. H. Lee, “Development of a proximal machine vision system for off-season weed mapping in broadacre no-tillage fallows,” J. Comput. Sci., vol. 9, no. 12, pp. 1803–1821, 2013.

[7] C. Duchesne, J. J. Liu, and J. F. MacGregor, “Multivariate image analysis in the process industries: A review,” Chemom. Intell. Lab. Syst., vol. 117, pp. 116–128, 2012.

[8] P. Geladi, E. Bengtsson, K. Esbensen, and H. Grahn, “Image analysis in chemistry I. Properties of images, greylevel operations, the multivariate image,” Trends Anal. Chem., vol. 11, no. 1, pp. 41–53, 1992.

[9] G.-Z. Li, and J. Y. Yang, "Feature selection for ensemble learning and its application," Machine Learning in Bioinformatics, Y.-Q. Zhang and J. C. Rajapakse, eds., pp. 135-155: John Wiley & Sons, 2008.

[10] H. Liu, and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining: Kluwer Academic Publishers, 1998.

[11] L.C. Molina, L. Belanche, and A. Nebot, “Feature selection algorithms: a survey and experimental evaluation,” in Proc of IEEE Int. Conf. on Data Mining (ICDM-03), 2002, pp. 306–313.

[12] H. Liu, and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Trans. on Knowl. and Data Eng., vol. 17, no. 4, pp. 491-502, 2005.

[13] I. H. Witten, E. Frank, and M. A. Hall, "Chapter 8 – Ensemble learning," Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, pp. 351-373, Boston: Morgan Kaufmann, 2011.

[14] R. Polikar, “Ensemble based systems in decision making,” IEEE Circuits Syst. Mag., vol. 6, no. 3, pp. 21-45, 2006

[15] G. Camps-Valls and L. Bruzzone, “Kernel-based methods for hyperspectral image classification,” IEEE Trans. on Geosci. Remote Sens., vol. 43, no. 6, pp. 1351–1362, 2005.

[16] I. Guyon, and A. Elisseeff, “An introduction to variable and feature selection,” J. of Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.

[17] Z. Zhao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand, H. Liu, “Advancing feature selection research – a feature selection repository”. School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tec. Rep., 2010.

[18] H. Liu and L. Yu, “Feature selection for high-dimensional data: A fast correlation-based filter solution.” in Proc. of 12th Int. Conf. on Machine Leaning (ICML-03), 2003, pp. 856–863, Washington, D.C.

[19] M. A. Hall, and L. A. Smith, “Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper,” in Proc. of 12th Int. Florida Artificial Intelligence Research Society Conference, 1999, pp. 235-239.

[20] M. Robnik-Šikonja, and I. Kononenko, “Theoretical and empirical analysis of reliefF and rreliefF,” Mach. Learn., vol. 53, no. 1, pp. 23-69.

[21] G. Brown, J. Wyatt, R. Harris, and X. Yao, “Diversity creation methods: a survey and categorisation,” J. Info. Fusion, vol. 6, no. 1, pp. 5-20, 2005.

[22] A.-K. Mahlein, T. Rumpf, P. Welke, H.-W. Dehne, L. Plmer, U. Steiner, and E.-C. Oerke, “Development of spectral indices for detecting and identifying plant diseases,” Remote Sensing of Environment, vol. 128, pp. 21–30, 2013.