Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Hyperspectral band selection for soybean classification based on information measure in FRS theory

Yao Liu [a], Tao Wu [a,*], Junjie Yang [a], Kezhu Tan [b], Shuwen Wang [a]

[a] School of Information Engineering, Lingnan Normal University, Zhanjiang, 524048, China
[b] College of Electrical and Information, Northeast Agricultural University, Harbin, 150030, China

ARTICLE INFO

Article history:
Received 27 November 2018
Accepted 5 December 2018
Published online 18 December 2018

Keywords:
Hyperspectral imaging
Band selection
Fuzzy rough set
Information measure
Soybean classification

Soybeans and soy foods have attracted widespread attention due to their health benefits. Special varieties of soybeans are in demand from soybean processing enterprises. Because of the advantage of rapid measurement with minimal sample preparation, hyperspectral imaging technology is used for classifying soybean varieties. Based on fuzzy rough set (FRS) theory, the research of hyperspectral band selection can provide the foundation for variety classification. The performance of band selection with Gaussian membership functions and triangular membership functions under various parameters were explored. Appropriate ranges of parameters were determined by the number of bands and mutual information of subsets relative to the decision. The effectiveness of the proposed algorithms was validated through experiments on soybean hyperspectral datasets by building two classification methods, namely Extreme Learning Machine and Random Forest. Compared with ranking methods, the proposed algorithm provides a promising improvement in classification accuracy by selecting highly informative bands. To further reduce the size of the subset, a post-pruning design was used. For the Gaussian membership function, a subset containing eight bands achieved an average accuracy of 99.11% after post-pruning. As well as classification accuracy, we explored stability of band selection algorithm under small perturbations. The band selection algorithm of the Gaussian membership function was more stable than that of the triangular membership function. The results showed that the information measure (IM) based band selection algorithm is effective at obtaining satisfactory classification accuracy and providing stable results under perturbations.

1. Introduction

For centuries, soy foods have become increasingly popular in many countries (Messina, 2002). They play an important role in disease prevention owing to their high-quality protein and essential amino-acid content. In particular, experimental studies have shown that the isoflavone compounds, genistein and daidzein, found in soybean and its processed products, prevent the development of osteoporosis, heart disease and

* *Corresponding author.*
E-mail address: wutao@whu.edu.cn (T. Wu).

https://doi.org/10.1016/j.biosystemseng.2018.12.002

cancers (Ishimi, 2009). Soybeans and soy foods have received considerable attention and led to skyrocketing sales of soy foods.

The characteristics of different varieties soybeans vary greatly. Soybeans are classified on the basis of quality parameters such as protein, oil and other contents. For the food processing industry, special varieties of soybeans are needed rather than mixed varieties. For example, high-oil soybeans are needed by edible oil processing plants, whereas high-isoflavone soybeans are needed by pharmaceutical companies that produce health-care products. Traditional methods of variety identification (mass spectrometry and high-performance liquid chromatography) require sample destruction, and are costly and time-consuming (Gowen, O'Donnell, Cullen, Downey, & Frias, 2007). Therefore, it would be beneficial to develop simpler, more rapid and cost-effective methods of quickly and reliably identifying soybean varieties. Hyperspectral imaging technology is a powerful non-destructive food analysis method, which has emerged in recent years. It is a rapid measurement with minimal sample preparation, and has potential for industrial online applications (Cheng & Sun, 2014, 2015).

Hyperspectral images provide better discrimination ability than traditional multispectral images, because they contain hundreds of bands and provide very good spectral resolution (Pantazi, Moshou, & Bravo, 2016). However, the high-dimensional data bring great difficulties in machine learning, data mining and pattern recognition, including degradation in efficiency and accuracy (Ravikanth, Singh, Jayas, & White, 2016). In view of these challenging tasks, many studies have been conducted on dimensionality reduction (Li et al., 2014; Lyashenko & Popov, 2015; Zabalza et al., 2016). Feature (band) selection (Datta, Ghosh, & Ghosh, 2014) and feature (band) extraction (Imani & Ghassemian, 2014) are two main methods of dimensionality reduction. The band selection method retains the original physical information of acquired spectral bands. This paper focuses on this method.

Rough set (RS) theory (Pawlak, 2002) has been widely used in data mining, classification and feature selection. Compared with other approaches to feature selection, the RS models can be used to discover data dependencies by purely structural methods, and remove redundant features. The reduced set of features has the same ability to differences as using all of the features, because it keeps intrinsic semantics of the features (Xu, Miao, & Wei, 2009). One limitation of the RS models is that they are only applicable to nominal features. To deal with numerical hyperspectral data, most existing methods discretise datasets, which may lead to information loss. Introducing the FRS theory is a way to address this problem, and was first proposed by Dubois and Prade (1990). The FRS theory encapsulates concepts of vagueness of fuzzy set and indiscernibility of rough set. These concepts are related and complementary. The properties, axiomatisation and applications of FRS theory have been analysed in detail by Hu, Yu, and Xie (2006) and Wu and Zhang (2004). The lower/upper approximation in the FRS models attempts to give a membership function of each sample (Qian, Wang, Cheng, Liang, & Dang, 2015).

In feature selection algorithms based on the FRS, a function is required to evaluate feature quality (Hu, Zhang, An, Zhang, & Yu, 2012). Consistency, dependency, and mutual information (MI) are employed to determine the reduced set of a fuzzy information system (Zhang, Mei, Chen, & Li, 2016). Since information entropy was introduced to measure uncertainty of random events, a series of IM functions was proposed to calculate the fuzziness (Lin, Liang, & Qian, 2015; Yu, Hu, & Wu, 2007; Zhao, Zhang, Han, & Zhou, 2015). Hu et al. (2012) applied fuzzy information entropy to characterise the uncertainty of fuzzy binary relation, measured the significance of features, and selected a feature subset from a numerical or hybrid dataset. In different application fields, the FRS models have been successfully applied to feature selection. They have seldom been used for hyperspectral band selection. In general, the shape of the membership function in the fuzzy models has an impact on the optimisation result. However, thus far, no work has been done to compare the performance of the models in different membership functions and the influence of the parameters on the models. In this study, we expect to determine the optimal range of parameters and select useful bands in information measure of the FRS theory (IM-FRS). We provide a comprehensive picture of the relationship between the shapes of the membership functions and the performance of the models under different parameters.

For existing feature selection algorithms, to allow appropriate algorithms to be chosen by users, it is necessary to compare algorithms by effective evaluation techniques. One criterion is the classification accuracy, which illustrates distinguishability of selected features (Kalousis, Prados, & Hilario, 2007). For the study, Extreme Learning Machine (ELM) (Heras, Argüello, & Quesada-Barriuso, 2014) and Random Forest (RF) (Breiman, 2001) are taken to assess the performance of feature selection algorithms based on the IM-FRS.

Stability is the sensitivity of an algorithm when the training dataset is disturbed slightly (Liu et al., 2017). It has become a crucial criterion for evaluating feature selection algorithms. Algorithms that present almost identical results even in perturbation are referred to as stable algorithms. Stable feature selection algorithms are preferable to unstable ones because, for unstable algorithms, small perturbations may lead to completely different outputs. An algorithm that is stable, but leads to poor classification performance will not be adopted by domain experts. Thus these two aspects, stability and classification performance, must be investigated together. For hyperspectral datasets, there are limited studies on the stability of feature selection algorithms.

This paper presents an IM for fuzzy equivalence relations. The proposed algorithms compute the relevance between spectral bands and decision by IM and then select informative bands through the greedy forward-search strategy. The effectiveness of the proposed algorithms is validated through experiments on soybean hyperspectral datasets. In addition to classification accuracy of two supervised methods of the ELM and RF classifiers, stabilities of proposed algorithms with different parameters are also compared. The emphases of this study are on (1) selecting informative bands in different fuzzy similarity relations which are computed by the Gaussian and triangular membership functions, respectively; (2) determining optimal parameters that show a significant impact on band selection; (3) employing a post-pruning technique to select the first "k" best bands and acquire minimum subset; (4) discussing the influence of perturbation levels and parameter values on stability.

2. Materials and methods

2.1. Hyperspectral imaging system

A hyperspectral imaging system (Headwall Photonics Company, USA) (Fig. 1(a)) was used to obtain image and spectral information of samples. The spectral resolution is 2.96 nm. The hyperspectral image contains 203 spectral bands in the wavelength range from 400 to 1000 nm. Fifty-five samples of each soybean variety (DongNong42, DongNong51 and DongNong61) were used for hyperspectral analysis. The three varieties of soybeans were obtained from the experimental farm, Northeast Agricultural University, China. Figure 1(b) is the corrected image of DongNong51 samples at 706.15 nm wavelength. The spectral reflectance values of all pixels identified by the region-of-interest image are averaged to represent the spectral information of the sample with a single value.

2.2. The RS theory and FRS theories

In the early 1980s, Pawlak presented the RS theory, and described the dataset as an information system (IS), which is denoted as $\langle U, A\rangle$. $U = (x_1, x_2, \cdots, x_n)$ is called the universe. A is a set of features. If the feature set $A = C \cup D$, an IS is named as a decision table, and is expressed as $DT = \langle U, C, D\rangle$. C and D are condition and decision feature sets, respectively.

Definition 1. For any $B \subseteq C$, an associated indistinguishable relation IND(B) is defined as

$$\mathrm{IND}(B) = \left\{ (x_i, x_j) \in U \times U \middle| \forall a \in B, f(x_i, a) = f(x_j, a) \right\} \tag{1}$$

According to Definition 1, $(x_i, x_j) \in \mathrm{IND}(B)$ means x_i and x_j are equivalent and indistinguishable by features in B.

Definition 2. The partition of U produced by IND(B) is defined as

$$U/\mathrm{IND}(B) = \left\{ [x_i]_B | x_i \in U \right\} \tag{2}$$

where $[x_i]_B$is the equivalence class, which is also called elemental information granule.

The classical RS model can deal with symbolic-valued datasets by equivalence relations. However, the value of a feature is usually real valued, which restricts applications of the RS model to some extent. The FRS model, combining the RS and fuzzy set, is capable of dealing with real-valued datasets by determining fuzzy similarity relations between samples (Hu, Zhang, Chen, Pedrycz, & Yu, 2010).

Definition 3. Given a fuzzy similarity relation R, the fuzzy equivalence class $[x_i]_R$ is defined as

$$[x_i]_R = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \cdots + \frac{r_{in}}{x_n} \tag{3}$$

The $r_{ik} = R(x_i, x_k)$represents the grade in which the two elements are equivalent or indiscernible. The objects in the class are fuzzy indistinguishable with x_i(Yu et al., 2007).

Definition 4. Fuzzy cardinality of $[x_i]_R$ is defined as

$$\left|[x_i]_R\right| = \sum_{j=1}^{n} r_{ij} \tag{4}$$

2.3. The membership function

In the RS model, the memberships of samples in the positive region are 1, and those in the negative region are 0. The memberships of samples in the boundary region are 0.5. To measure roughness grade in the boundary region, fuzziness is combined with the RS model. In the FRS model, the samples in the boundary region take values between 0 and 1 by a membership function (Wu & Zhang, 2004). Two categories of membership functions used in this study are Gaussian and triangular.

The Gaussian membership function is defined as

$$R(x_{ij}, x_{kj}) = \exp\left(-\frac{\left\| x_{ij} - x_{kj} \right\|^2}{2\sigma^2} \right) \tag{5}$$

The triangular membership function is defined as

$$R(x_{ij}, x_{kj}) = \begin{cases} 1 - \dfrac{\left\| x_{ij} - x_{kj} \right\|}{\lambda}, & \mathit{if} \left\| x_{ij} - x_{kj} \right\| \le \lambda \\ 0, \quad \mathit{if} \left\| x_{ij} - x_{kj} \right\| > \lambda \end{cases} \tag{6}$$

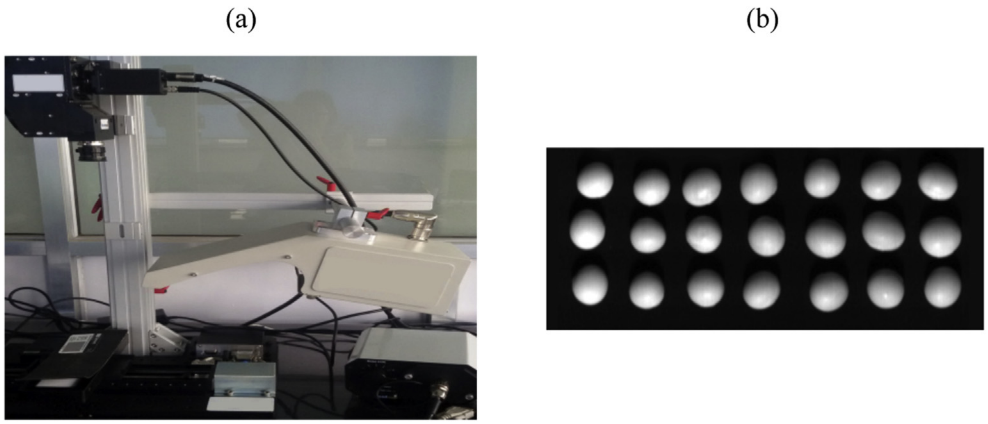

Fig. 1 – (a) Hyperspectral imaging system, and (b) corrected image at 706.15 nm wavelength.

They are used to quantify the similarity degree between samples x_i and x_k in terms of feature j. The parameters σ and λ, which can tune the width of the membership functions, control the granularity of approximation. The higher the values of σ and λ, the greater are the fuzzy information granules. A variety of values for each parameter is tested, ranging from 0 to 1, to investigate the influence of the shapes of the membership functions and the values of the parameters on the result of band selection.

2.4. *Dependency measure and information measure in the FRS model*

A rough set itself is an approximation of a vague concept. A pair of precise concepts, lower and upper approximations, is used for mining association rules (Jensen & Shen, 2004).

Definition 5. Let $X \in U$, lower and upper approximations are described as

$$\underline{P}(X) = \cup\{[x_i]_R \mid [x_i]_R \in X\} \tag{7}$$

$$\overline{P}(X) = \cup\{[x_i]_R \mid [x_i]_R \cap X \neq \phi\} \tag{8}$$

Lower approximation $\underline{P}(X)$ is known as positive region as well, and represented as $POS_P(X)$.

Definition 6. Given a decision system $DT = \langle U, C, D\rangle$, dependency between C and D is described as

$$\gamma_C(D) = \frac{|POS_C(D)|}{|U|} \tag{9}$$

When a feature is removed from the conditional feature set, the dependency will change. Significance of the feature is obtained by calculating the change in dependency.

The IMs, such as mutual information (MI), are capable of assessing the quality of condition features. Features with high MI with the category labels can classify better. By combining the FRS model with information theory, we explore the band selection methods based on the IM-FRS.

Definition 7. Given a decision system $DT = \langle U, C, D\rangle$, R is a fuzzy relation, and $B \subseteq C$ is a subset of bands. Information quantity of fuzzy equivalence relation is

$$H(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|[x_i]_R|}{|U|}. \tag{10}$$

$H(B) = 0$ if and only if $\forall x, y \in U, R(x, y) = 1$. In this case, distinguishing an arbitrary sample from the others is difficult. The granularity of the decision system is largest. $H(B) = \log|U|$ if and only if $\forall x \neq y$, $R(x, y) = 0$. In this case, the fuzzy approximation space is at its finest, and all of the samples are distinguishable.

Definition 8. Given a decision system $DT = \langle U,C,D\rangle$, $[x_i]_B$ and $[x_i]_D$ are fuzzy equivalence classes produced by B ($B \subseteq C$) and D, respectively. The conditional entropy of D conditioned to B is defined as

$$H(D|B) = -\frac{1}{|U|}\sum_{i=1}^{|U|}\log\frac{|[x_i]_D \cap [x_i]_B|}{|[x_i]_B|} \tag{11}$$

The relevance between condition features and decision features is reflected by conditional entropy.

Definition 9. The fuzzy MI of B and D is defined as

$$I(D;B) = H(D) - H(D|B) \tag{12}$$

The MI quantifies the reduction of uncertainty. The increment of discernibility of the system is reflected through the increment of MI. Thus, the significance of a band is described as follows.

Definition 10. Given a decision system $DT = \langle U, C, D\rangle$, $B \subseteq C$, and $a \in C - B$, the significance of band a in feature set B is defined as

$$SIG(a, B, D) == H(D|B) - H(D|B\cup\{a\}) = I(D; B\cup\{a\}) - I(D; B) \tag{13}$$

$SIG(a, B, D)$ is the difference between conditional entropy on the basis of $B\cup\{a\}$ and B. It is the increment of discernibility by adding band a.

Based on the dependency measure and IM, a greedy forward-search strategy is constructed (Liu et al., 2016). First, the fuzzy decision system of hyperspectral data is framed. Then, according to the definition of fuzzy MI, the significance of each band is calculated. The band with maximum significance is selected and added sequentially to the band subset. Depending on different parameters σ and λ, diverse optimal band subsets are generated. The classification performance of band subsets is assessed by using the ELM and RF classifiers.

2.5. *Stability of an algorithm*

Stability of an algorithm is defined as the characteristic of selecting similar or identical results even though perturbation exists in the training dataset. It is a prerequisite to measure similarity between two selected band subsets before evaluating stability of a band selection algorithm. In this study, we used the Jaccard Index, which is a similarity measure based on the index.

The similarity is calculated by the number of the same bands that occur in both selected subsets. Jaccard Index for two selected subsets R_i and R_j is described as

$$S_J(R_i, R_j) = \frac{|R_i \cap R_j|}{|R_i \cup R_j|} \tag{14}$$

The stability of the algorithm is the average of the similarities which are obtained by comparing all selected subsets from perturbation datasets (Dunne, Cunningham, & Azuaje, 2002). For l subsets, the stability is given by

$$S_J(R) = \frac{2}{l(l-1)}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l} S_J(R_i, R_j) \tag{15}$$

To assess the stability, new datasets with small perturbations in the training set are constructed. Most of methods to construct perturbation datasets are either random removal or cross-validation (Liu, Liu, & Zhang, 2010). The problem of these methods is that they result in an unknown degree of overlap or an invariable degree of overlap. We adopt a method of adjustable degree of overlap (Liu et al., 2017) in this work. The band selection algorithm based on IM-FRS is acted on each of ten perturbation datasets independently. The ten selected band subsets are compared pairwise. According to Jaccard index, similarity of each pairing is computed. The stability is the average of all similarity values.

3. Results and discussion

To illustrate the performance of the IM-FRS-based band selection algorithm, experimental results of hyperspectral datasets derived from the soybean samples are presented. Given a hyperspectral image dataset $DT = \langle U,C,D \rangle$, $U = \{x_1, x_2, \cdots, x_n\}$ is a finite set of samples described with a set of bands $C = \{a_1, a_2, \cdots, a_m\}$. D is the decision feature dividing the samples into categories $d_1, d_2, \cdots, d_N$. R is a fuzzy relation.

3.1. *Spectral analysis*

Figure 2(a) shows the spectral reflectance curves of the region-of-interest obtained from all 165 soybean samples. The samples belong to the same species, therefore, the trends of spectra are similar. The average spectral curves of each variety (DongNong51, DongNong42, and DongNong61) are plotted in Fig. 2(b). We observe that the characteristics and shapes of spectral curves are similar, but differ in the magnitude of reflectance. In the region of 420–520 nm, there are some differences between different varieties. The variety with the highest averaged spectral reflectance is DongNong51. The second highest of the spectral reflectance is DongNong42, and then DongNong61. However, the spectra are seriously overlapping in other wavelength regions. Classifying varieties directly by observed spectral data is difficult. It has become necessary to select informative bands and construct classification models.

3.2. *Variation of the subset size with parameters*

As hyperspectral datasets are comprised of continuous features, a discretisation step is necessary. Standard fuzzification techniques are implemented to decrease the information loss (Yeung, Chen, Tsang, Lee, Xi, 2005). For each feature, fuzzification techniques generate a group of fuzzy sets corresponding to fuzzy similarity relations. Meanwhile, they leave the underlying feature values unchanged. Different fuzzy similarity relations may lead to different techniques of performing feature selection reduction. Here, two types of fuzzy similarity relations, which are computed by the Gaussian and triangular membership functions, are explored in the process of dimensionality reduction. The parameters σ of Gaussian membership function and λ of triangular membership function control the granularity of the granulation space. To explain the impact of parameters on effectiveness of the FRS models, we performed experiments by setting both σ and λ to 0.01 and then increasing them in 0.01 increments to 1. For parameters σ and λ, the number of selected bands is shown in Fig. 3. The number of selected bands shows obvious changes by using different parameters when parameters σ and λ are less than 0.6. For both the Gaussian and triangular membership functions, the trends of variation of the number of selected bands are consistent. The number first increases to a peak, and then decreases with oscillatory variation. The number of selected bands by the Gaussian membership function is more than that by the triangular membership function. When parameters σ and λ are larger than 0.6, the number decreases significantly, and does not change significantly with the variation in the parameters. In some cases, the selected subset includes the same number of bands for different values of σ and λ.

Parameters σ and λ affect the performance of the band selection algorithm. With different parameters, different reduction results are obtained. The setting of parameters relies on the essence of the research object. The appropriate range of parameters cannot be determined only by the number of selected bands; the MI needs to be combined with a decision. Therefore, we carried out a series of experiments to address this concern.

3.3. *Variation of the MI with the number of bands*

The MI reflects the relevance between band subset and decision feature (Hu et al., 2010). More information is provided by the band subset having greater MI with the decision (Liu et al., 2016). Figure 4 illustrates the number of selected bands by Gaussian and triangular membership functions for different parameters (σ and λ are set to 0.1, 0.2, ..., 0.9 and 1). We can observe that when new bands are supplemented to subsets the MI values increase monotonously. At the beginning of the selection process, all MI curves increase fast, and then increase gradually. This indicates that the information provided

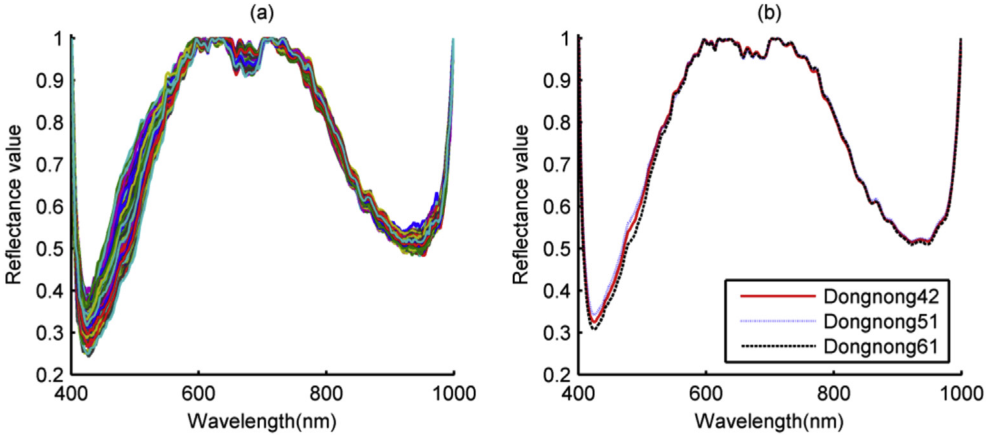

Fig. 2 – (a) Spectra reflectance curves of 165 soybean samples, and (b) average spectra of each variety.

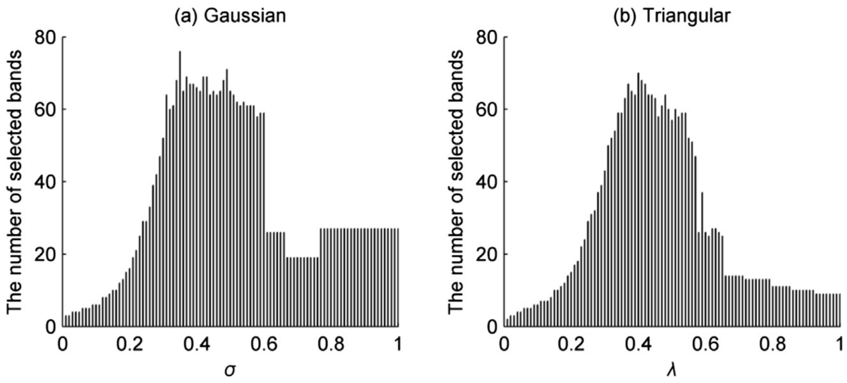

Fig. 3 – The number of bands with different parameters σ and λ.

by the bands selected first is more important than that by the bands added later. Very limited MI increases are added by the last several bands. The experimental results are in good agreement with the design of greedy search algorithm, that is, the band that causes the greatest increase in MI is chosen first and added to the subset.

In Fig. 4, the values of MI of the final subset vary with different parameters. The parameters play the role of containing the granularity. A coarser granulation is induced by using a greater parameter value. In this situation, it is hard to distinguish a sample because of a larger degree of similarity between any pair of samples. As a result, a lower value of MI is obtained. Experiments show that parameters σ and λ have an effect on the performance of band selection. The determination of the parameter range requires through further experiments.

3.4. *Variation of the MI with parameters*

According to the concept of MI, the MI can evaluate the availability of band subsets. Figure 4 presents the variations of MI when new bands are added sequentially to the band subset. Here, we explore the relationship between MI of the final selected subset and different parameters to determine the optimal range of the parameters. Figure 5 presents the curves of MI variation for the values of σ and λ from 0.01 to 1 increased in increments of 0.01. When σ and λ are in the range of 0.01–0.21, the values of MI of selected subset remain largely unchanged at approximately 1.585. If the classification is consistent, the MI between selected subset B and decision D have the same value as decision uncertainty. In our experiments, the variety of soybeans for classification is three. The uncertainty quantity of decision is $H(D) = \log_2 3 = 1.585$. Values $I(B;D)$ and $H(D)$ are identical absolutely. This means if reduced subset B is known, there is no uncertainty in classification. The information provided by reduced subset is the same as that provided by the entire set of bands.

When σ and λ are larger than 0.3, the values of MI decrease drastically. For the Gaussian membership function, the value decreases to approximately 0.7 when σ = 0.5. Afterwards, it only changes slightly. For the triangular membership function, the curve of MI continues to decrease gradually after

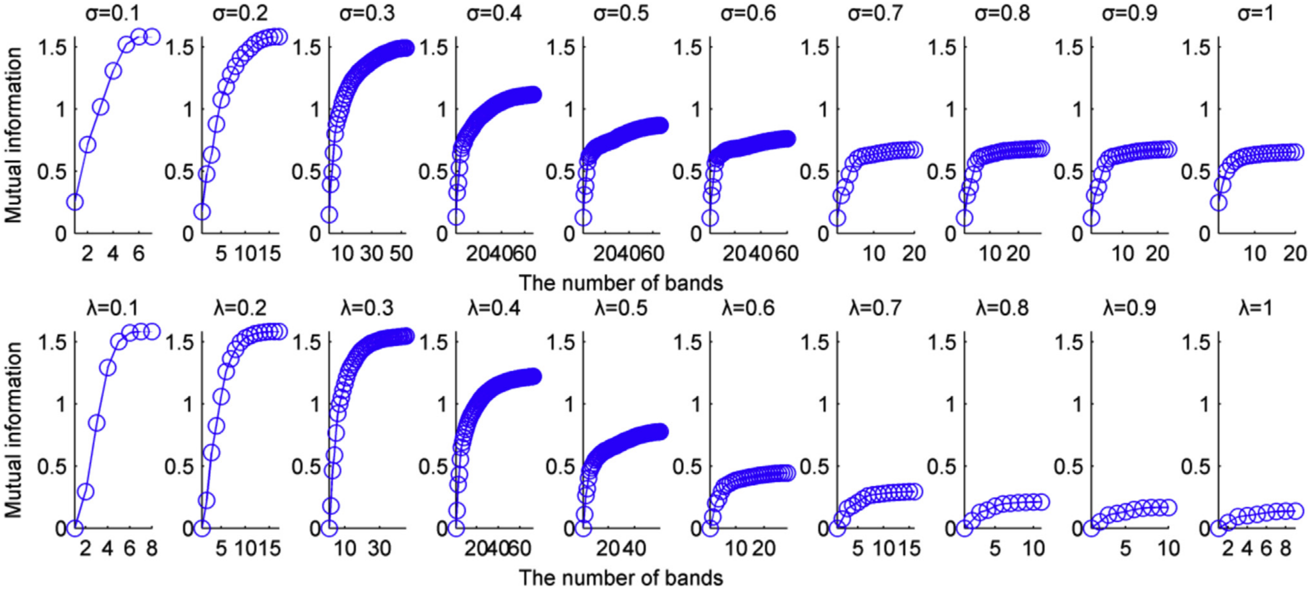

Fig. 4 – Variations of MI value for different parameters.

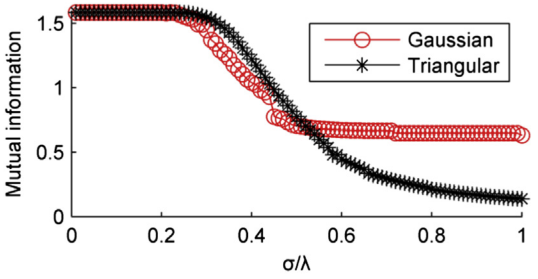

Fig. 5 – MI of the band subset as a function of σ (Gaussian) or λ (triangular).

decreasing drastically. As σ and λ increase, the values of MI decrease. This conclusion is consistent with our previous observations in Fig. 4. We can see from these results that 0.01–0.21 is a proper interval for σ and λ. However, we have found that the values of MI do not decrease substantially when σ and λ are within the interval [0.21, 0.25]. Therefore, we extend the range of the parameters σ and λ to 0.01–0.25 and can further explore the size of the selected subset and the corresponding classification performance in this range. In addition, for different membership functions, the variations of MI differ. The membership functions also affect the results of band selection. This is to be further explored.

3.5. *Classification performance of band selection algorithm*

The ELM and RF classifiers were applied to assess the performance of the algorithm based on the IM-FRS. For each soybean variety, forty samples were used for training and the other fifteen samples for testing. After the samples of training set and testing set have been selected randomly, the experiments were conducted 100 times. The parameters σ and λ are set from 0.01 to 0.25 in increment of 0.01. The maximum and average accuracy was adopted as measures of classification performance. Figures 6 and 7 show that the classification accuracy of Gaussian and triangular membership functions vary with σ and λ, respectively. The performance of the IM-FRS band selection algorithms using the ELM classifier is better than that using the RF classifier. For the Gaussian membership function, when $\sigma > 0.15$, the average classification accuracy using the ELM classifier achieves 95%, whereas the average classification accuracy with the RF classifier is only approximately 80%. For discriminating soybean varieties, the ELM classifier can achieve satisfactory results.

The ELM is an innovative machine learning algorithm. Only the number of hidden neurons needs to be tuned when using the ELM network. The input weights and hidden-layer bias are chosen at random. To determine the optimal number of neurons, we conducted experiments by changing the number from 10 to 200. The number of neurons is increased by 10 each time. In our experiments, the ELM network with sigmoidal activation function (Suresh, Saraswathi, & Sundararajan, 2010) $g(x) = 1/[1 + \exp(x)]$ was chosen. Take several cases of the Gaussian membership function as an example. In Fig. 8, the variations of classification accuracy with the number of hidden neurons are given. The experimental results show that the number of neurons has a great influence on output of network. When the number of neurons varies from 10 to 40, the classification accuracy increases gradually and achieves a maximum at 40 neurons. When the number of neurons increases to more than 40, the classification accuracy begins to decrease. The accuracy reaches a minimum at approximately 120 neurons. It can be seen that a greater number of neurons is not necessarily better. The ELM network with too large a number of neurons could suffer from over-fitting. For the soybean hyperspectral dataset, we adopt the ELM classifier with 40 neurons.

The RF is an ensemble method. To classify a new sample, many decision trees are constructed. From the whole original set, a subset of features selected randomly is used as the decision tree nodes. However, there is no uniform standard regarding how many trees should be used to compose an RF (Oshiro, Perez, & Baranauskas, 2012). In general, the number of trees is set on a trial-and-error basis. For the RF classifier, to determine the relationship between the classification performance and the number of trees, we increased the number from 10 to 200 in increments of 10. For the Gaussian membership function, Fig. 9 illustrates the variations of accuracy with the number of trees. When the number of trees reaches 50, the classification accuracy reaches a maximum. When the number of trees is greater than 50, the classification accuracy fluctuates as the number of trees increases. The accuracy also can reach the maximum value when the number of trees reaches certain values greater than 50. However, the excessive number of trees increases the model complexity and leads to a longer prediction-response time. Therefore, we use 50 trees in the experiment.

Overall, the band selection algorithm with the Gaussian membership function produces higher classification accuracy than with the triangular membership function. We can easily conclude that, even on the same hyperspectral dataset, the results of band selection with different membership functions differ. The membership functions, affect the performance of band selection. However, except for classification accuracy, the number of selected bands and stability of the algorithm need to be considered in the comparison. The comprehensive effectiveness of the IM-FRS algorithm for the hyperspectral dataset is discussed in a later section.

3.6. *Performance comparison between IM-FRS algorithm and ranking algorithm*

The ranking algorithms are auxiliary band selection mechanisms. For ranking algorithms, candidate bands are sorted by employing an evaluation function, such as correlation criteria, inter-class distance, and MI. The top ranked bands are selected. Here the MI between each band and category labels is calculated as the significance of an individual band. The band with greater significance provides more information for classification. The significance of an individual band by Gaussian and triangular membership functions at $\sigma = 0.15$ ($\lambda = 0.15$) are shown in Fig. 10. For the Gaussian membership function, the band at the 863.2 nm wavelength has the highest significance of 0.2198, and for the triangular membership function, the band at the 916.6 nm wavelength has the highest significance of 0.2556. We rank the bands in descending order

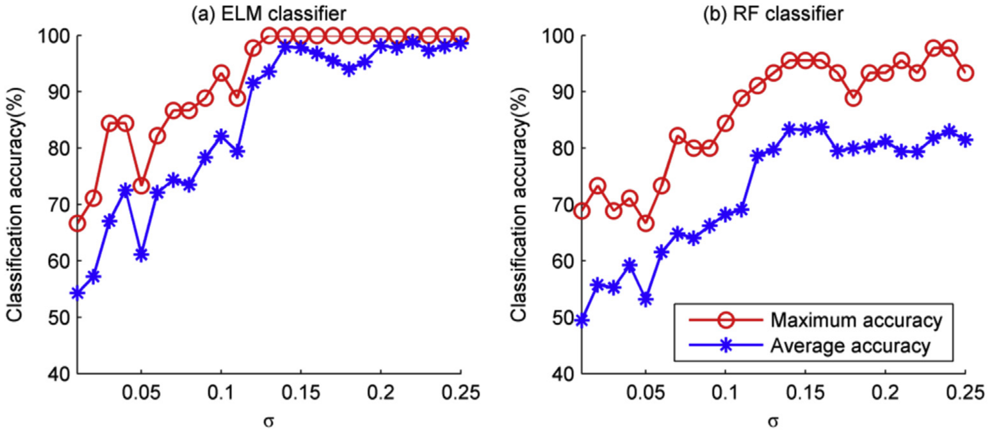

Fig. 6 – Classification accuracy of Gaussian membership function.

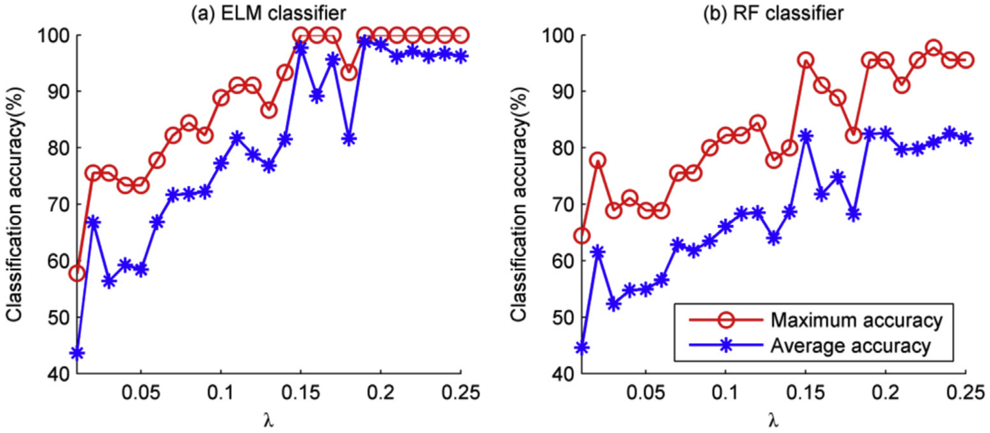

Fig. 7 – Classification accuracy of triangular membership function.

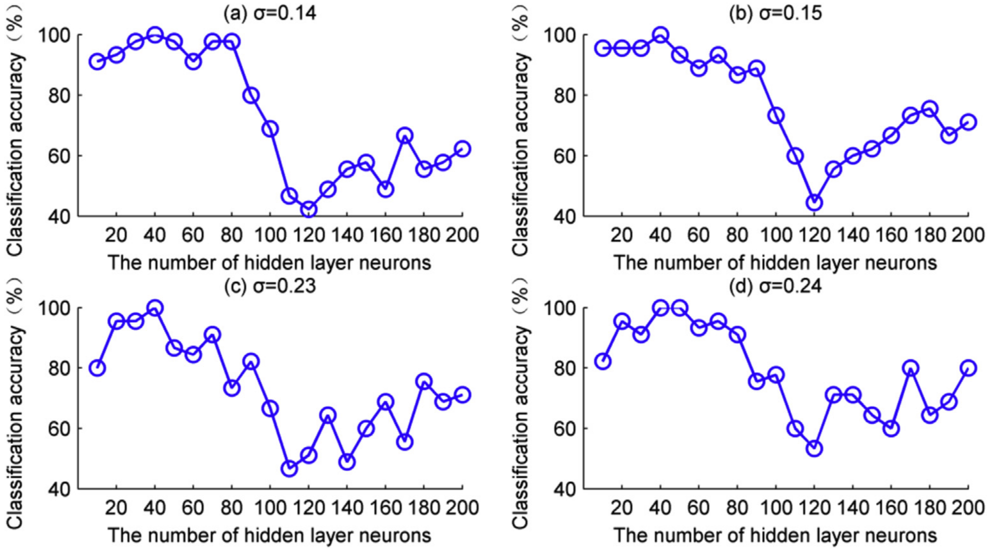

Fig. 8 – Impact of the number of hidden neurons on classification performance for ELM classifier.

of significance. For the Gaussian membership function, the top ten bands distribute in sequence at wavelengths of 863.2, 916.6, 839.5, 842.5, 884.0, 869.1, 617.3, 878.0, 895.8 and 845.4 nm. For the triangular membership function, the top ten bands distribute in sequence at wavelengths of 916.6, 863.2, 916.6, 869.1, 845.4, 884.0, 889.9, 842.5, 617.3 and 895.8 nm. For

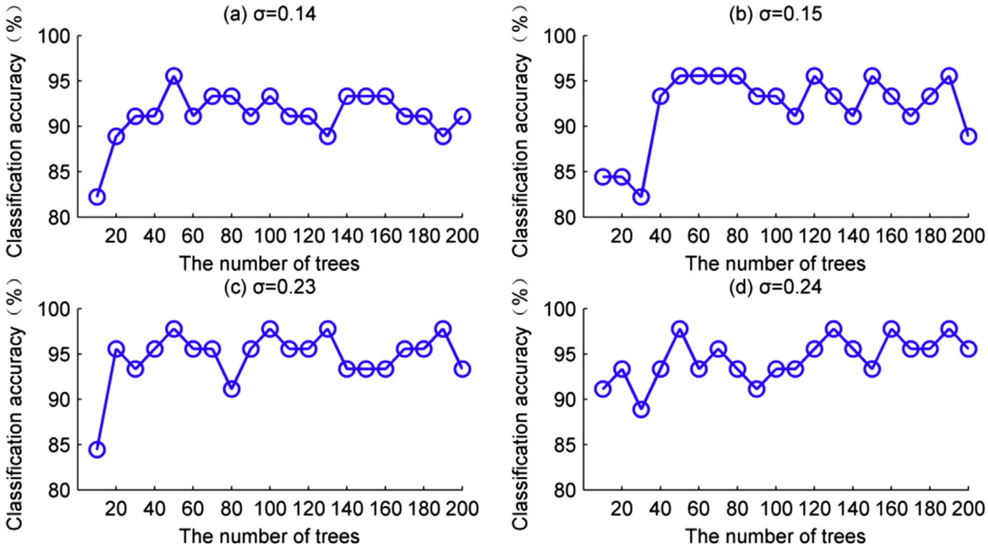

Fig. 9 – Impact of the number of trees on classification performance for RF classifier.

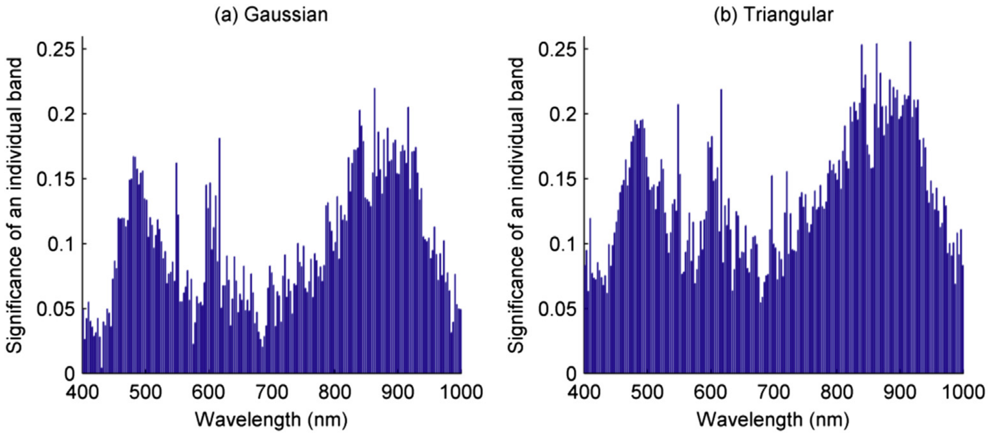

Fig. 10 – Significance of an individual band.

different membership functions, the selected bands are slightly different. The selected bands are concentrated around 872.1 nm and some of them are adjacent to each other. As is well known, the correlation between adjacent bands of hyperspectral images is very strong. However, ranking algorithms only compute the dependency between bands and category labels, while neglecting the redundancy or correlation between bands.

To compare the performance of the ranking algorithm and the IM-FRS algorithm, the same number of bands are taken for each algorithm. The ELM classifier is used to evaluate the performance. Figure 11 illustrates the classification performance (maximum and average classification accuracy) by the IM-FRS band selection algorithm and the ranking algorithm when σ and λ range from 0.15 to 0.2. For both Gaussian and triangular membership functions, the subsets selected by the IM-FRS band selection algorithm have much better classification ability than those selected by the ranking algorithm. For instance, with respect to the Gaussian membership function, when $\sigma = 0.15$, the average classification accuracy of the proposed algorithm with ten bands achieves 98.80%, while the average accuracy is only 71.47% with the ranking algorithm. The top ranking bands selected by the ranking algorithm are bands with great individual power, but their combinations are weak at discernibility.

3.7. *Performance comparison between IM-FRS algorithm and dependency measure of the FRS theory*

We compare performance of the IM-FRS algorithm with the dependency measure of the FRS theory (DM-FRS). Tables 1 and 2 illustrate the accuracy of the IM-FRS and DM-FRS algorithms for different size subsets by ELM and RF classifiers. The IM-FRS algorithm exhibits superior classification performance compared to the DM-FRS algorithm in terms of subsets of different sizes, for both Gaussian and triangular membership functions. For the average values for subsets of different sizes, there is 9.78% performance improvement in terms of maximum accuracy compared with the DM-FRS algorithm by the Gaussian membership function and the ELM classifier.

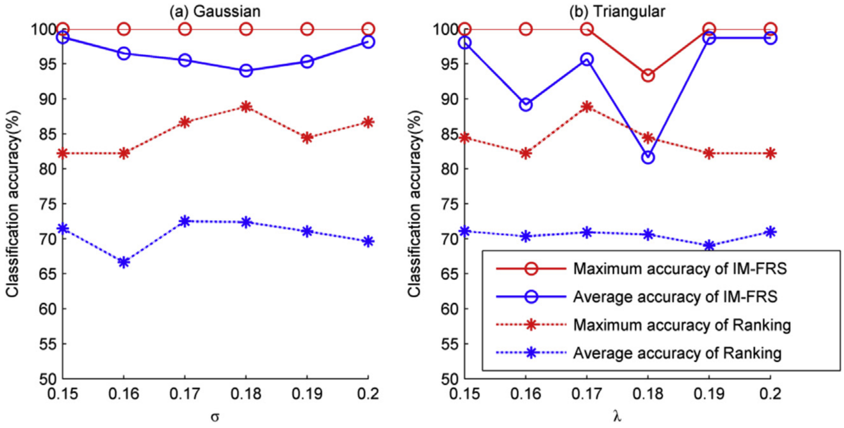

Fig. 11 – Classification accuracy of the IM-FRS band selection algorithm and the ranking method.

Table 1 – Classification performance of IM-FRS and DM-FRS algorithms for different size of subsets by ELM classifier.

Size of subset	Gaussian membership function				Triangular membership function			
	MA (%)		AA (%)		MA (%)		AA (%)	
	IM-FRS	DM-FRS	IM-FRS	DM-FRS	IM-FRS	DM-FRS	IM-FRS	DM-FRS
6	93.33	80.00	82.13	63.53	88.88	84.44	77.24	70.71
7	97.78	84.44	91.60	73.35	91.11	84.44	81.76	72.40
8	100.00	86.67	93.58	74.18	93.33	86.67	81.51	74.60
9	100.00	93.33	97.96	79.20	100.00	88.89	89.18	77.20
10	100.00	91.11	96.80	79.13	100.00	88.89	97.71	79.38
11	100.00	93.33	97.80	80.93	100.00	77.78	95.67	65.96
12	100.00	88.88	95.53	72.42	93.33	88.89	81.62	79.93
13	100.00	88.88	94.02	74.67	100.00	91.11	98.76	78.62
14	100.00	91.11	95.31	83.73	100.00	95.56	98.29	86.22
15	100.00	95.56	98.16	84.22	100.00	97.78	96.16	91.76
Average	99.11	89.33	94.29	76.54	96.67	88.45	89.79	77.68

Note: MA = maximum accuracy, AA = average accuracy.

Table 2 – Classification performance of IM-FRS and DM-FRS algorithms for different size of subsets by RF classifier.

Size of subset	Gaussian membership function				Triangular membership function			
	MA (%)		AA (%)		MA (%)		AA (%)	
	IM-FRS	DM-FRS	IM-FRS	DM-FRS	IM-FRS	DM-FRS	IM-FRS	DM-FRS
6	82.22	68.89	69.82	52.36	80.00	71.11	61.73	57.78
7	95.56	75.56	80.29	63.38	86.67	75.56	67.84	61.29
8	97.78	82.22	80.73	66.96	82.22	77.78	68.80	65.49
9	93.33	84.44	82.40	69.38	84.44	80.00	71.96	61.40
10	95.56	84.44	82.58	69.16	95.56	82.22	82.44	69.11
11	97.77	80.00	85.44	68.64	93.33	75.56	75.29	57.80
12	95.56	80.00	79.51	62.27	84.44	77.78	70.00	65.78
13	93.33	77.78	80.24	63.44	97.78	82.22	82.89	68.13
14	93.33	80.00	80.53	67.78	97.78	88.89	84.51	72.02
15	95.56	84.44	82.84	68.76	95.56	86.67	81.56	73.36
Average	94.00	79.78	80.44	65.21	89.78	79.78	74.70	65.22

Moreover, the IM-FRS algorithm improves the average accuracy by 17.75%.

In terms of the overall scale, the IM-FRS algorithm using the Gaussian membership function is superior to that using the triangular membership function. In Table 1, for the Gaussian membership function, the average accuracy is 94.29%, and for the triangular membership function, 89.79%.

3.8. *Post-pruning strategy*

Over-fitting is an important challenge for classification of hyperspectral datasets. To overcome over-fitting, pre-pruning and post-pruning methods are adopted. Pre-pruning method is essentially the same as the filtering method, which is preferred if the accuracy improves monotonically with the

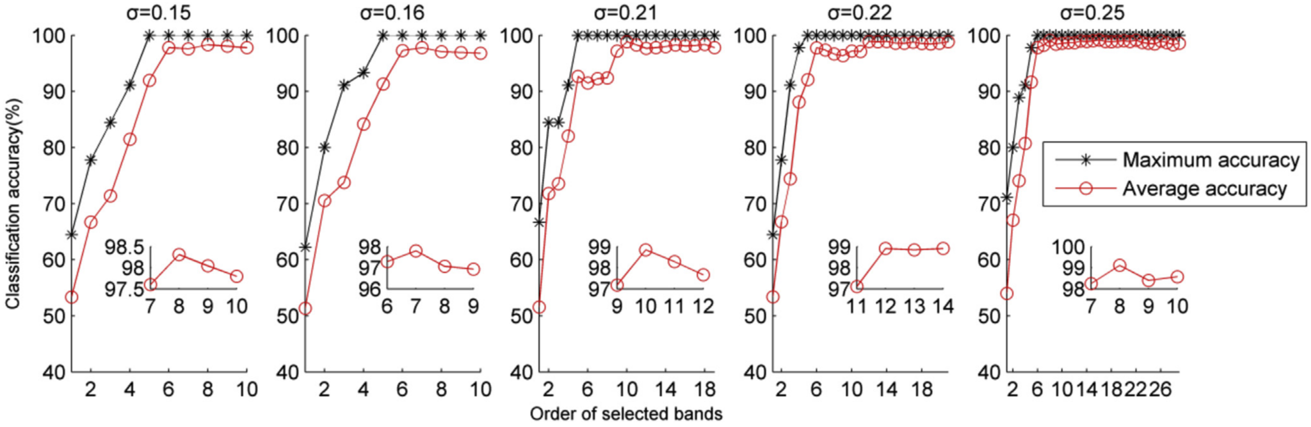

Fig. 12 – Variation of accuracy with the order of selected bands using Gaussian membership function.

number of bands. However, for many algorithms used in machine learning, the monotonicity assumption is not valid. Acquiring a minimal subset with excellent classification performance is the target of band selection. A more appropriate strategy would be to use a filtering method first, followed by a post-pruning method which will detect and delete those bands that have become useless and possibly interfere in the prediction task.

By applying the greedy forward-search strategy, the bands with maximum significance were added to the subset individually. A group of subsets was produced ($B_1 \subset B_2 \subset \cdots \subset B_K$). The ELM classifier was applied to evaluate capability of $B_i(i = 1, 2, \cdots, K)$ sequentially in the order of selection. The subset obtaining the highest accuracy was selected.

Experiments were conducted to evaluate how the order of selected bands affects classification performance. The phenomenon of over-fitting does not occur under all parameters, but only when the parameters σ and λ take certain values. Figures 12 and 13 visualise the relation between classification performance and the order of selected bands. The results show that the average and maximum accuracy increase sharply initially. Afterwards, the classification accuracy is not significantly improved. Once the average accuracy reaches maximum, it decreases slightly. The bands selected after the peak augment the MI, but they are superfluous for classification. They do not exhibit any significant improvement in the classification performance, but rather degenerate performance. These bands should be removed from the subsets.

To demonstrate effectiveness of the post-pruning, Table 3 displays a comparison of classification performance and the size of subset with and without post-pruning. After post-pruning, in the final results of band selection, a large number of bands have been excluded. However, the average classification accuracy is simultaneously enhanced. For instance, with respect to the Gaussian membership function, when $\sigma = 0.25$, a subset containing eight bands achieves an average accuracy of 99.11% after post-pruning, compared to a 29-band subset with an average accuracy of 98.58% obtained without post-pruning. In other words, a 72% reduction in the number of bands increases the accuracy by 0.5%. With respect to the triangular membership function, when $\lambda = 0.25$, the number of bands is reduced by 72%, while the average accuracy increases by more than 2.3%. These results reveal that the post-pruning strategy can delete redundant bands from band subsets. It is essential to band selection.

A trade-off between subset suitability and subset minimality is always necessary to solve the problem of determining subset optimality (Jensen & Shen, 2007). A higher accuracy is of course desirable, but so is a smaller subset size.

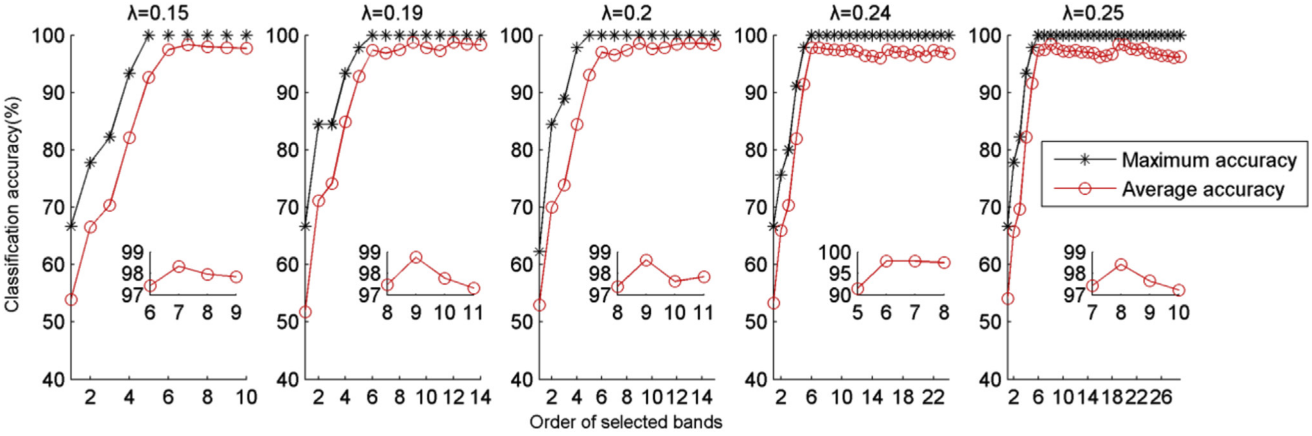

Fig. 13 – Variation of accuracy with the order of selected bands using triangular membership function.

Table 3 – Performance comparison with and without post-pruning.

Gaussian membership function							Triangular membership function						
σ	No post-pruning			Post-pruning			λ	No post-pruning			Post-pruning		
	NB	MA (%)	AA (%)	NB	MA (%)	AA (%)		NB	MA (%)	AA (%)	NB	MA (%)	AA (%)
0.15	10	100.00	97.80	8	100.00	98.31	0.15	10	100.00	97.71	7	100.00	98.33
0.16	10	100.00	96.80	7	100.00	97.80	0.19	14	100.00	98.33	9	100.00	98.75
0.21	19	100.00	97.82	10	100.00	98.84	0.2	15	100.00	98.29	9	100.00	98.62
0.22	21	100.00	98.85	12	100.00	98.91	0.24	24	100.00	96.73	6	100.00	97.84
0.25	29	100.00	98.58	8	100.00	99.11	0.25	29	100.00	96.24	8	100.00	98.42

Note: NB = the number of bands, MA = maximum accuracy, AA = average accuracy.

As can be seen from Table 3, most of the bands are deleted from the original set, and the reduction rate is as high as 97% for certain parameters. For example, when λ = 0.24, the number of bands is 6 after post-pruning, whereas the number of original hyperspectral dataset is 203. The results also verified that many different subsets performed equally well in classification accuracy. This observation prompts other evaluation methods to be explored to supplement the comprehensive evaluation. Therefore, we explore stability of algorithm based on the IM-FRS for different membership functions and different parameters.

3.9. *Stability of algorithm based on the IM-FRS*

To describe the stability of the proposed algorithm in detail and more intuitively, Fig. 14 provides a frequency histogram of the selection results from ten perturbation datasets by the Gaussian membership function when $\sigma = 0.16$ and 0.21. The l denotes the number of deleted samples of each variety. A greater value of l means a larger perturbation. As is well known, when the selected bands are concentrated, the band selection algorithm is more stable. When the variety of selected bands is lower, the band selection algorithm is more stable. From Fig. 14, we can observe that the selection result when $l = 1$ is more stable than that when $l = 3$ and 5. Overall, some bands have a big opportunity to be selected for different parameters, because they are greatly relevant to the category labels. For example, the wavelengths of 549.1, 617.3 and 863.2 nm are selected in most cases.

When $l = 1$, the results of the proposed algorithm with $\sigma = 0.16$ are more stable than with $\sigma = 0.21$. However, when $l = 3$, by comparing the histogram of frequency selection, it is difficult to judge the algorithm in terms of which of two parameters is more stable. The reason is that the number of bands differs according to parameters. To quantitatively characterise and precisely compare the stability, Fig. 15 describes the change of stability in relation to the perturbation levels of the dataset and the parameters that are listed in Table 3. As the perturbation level increases, the stability decreases. The conclusion is consistent with the result of the histogram of frequency. Moreover, the difference in stability between $l = 3$ and $l = 5$ is slight, except for $\sigma = 0.16$ with the Gaussian membership function.

From Fig. 15, the stability is directly affected by the membership functions and their parameters. Overall, the algorithm with Gaussian membership function is always more stable than with triangular membership function. For example, considering the same size of subset, when $\sigma = 0.25$ and $\lambda = 0.25$, the stabilities using the Gaussian membership function are 0.75, 0.58, and 0.57 for three levels of

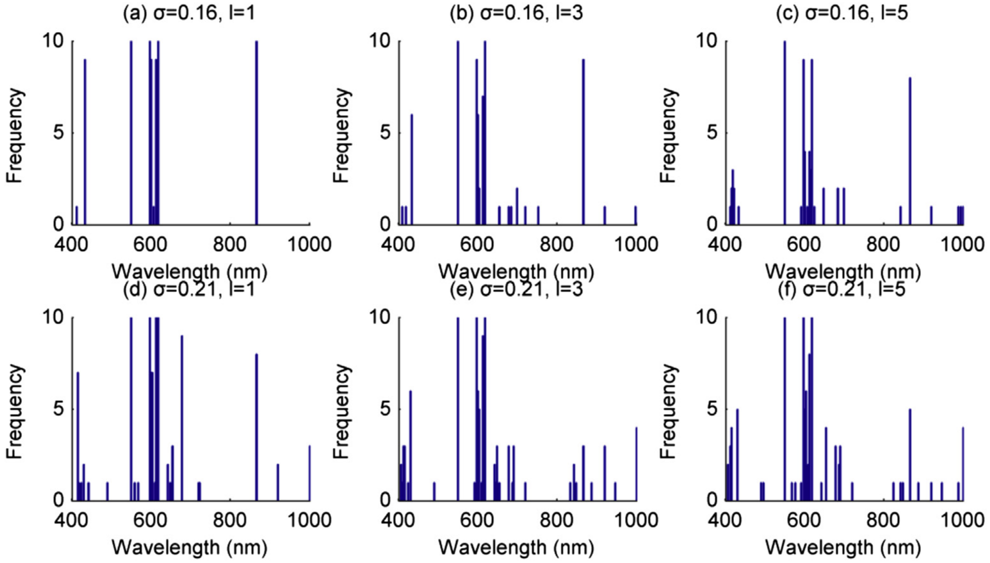

Fig. 14 – Frequency histograms of selection results by Gaussian membership function when σ = 0.16 and 0.21.

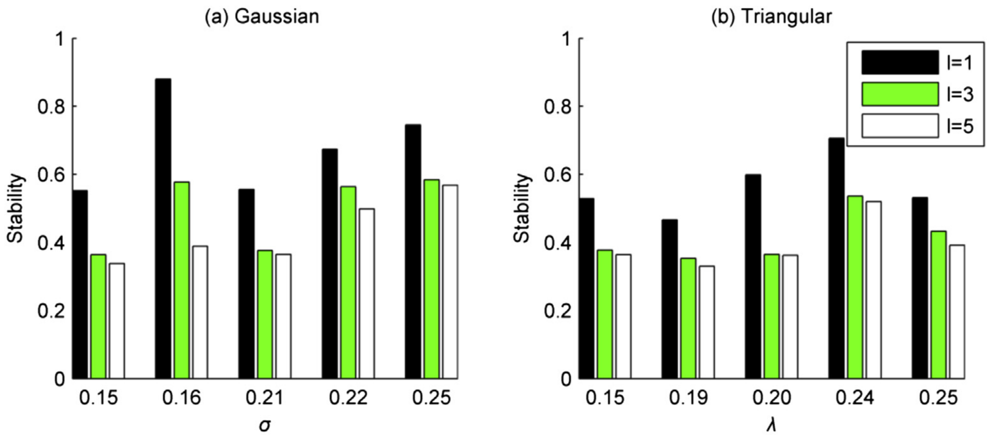

Fig. 15 – Variation of stability with the perturbation levels and the parameters.

perturbation, while the stabilities using the triangular membership function are 0.53, 0.43, and 0.39. Even for the same membership function and the same subset size, different stability results are obtained by different parameters.

Finding the smallest subset of bands with excellent classification accuracy and stability is the main goal of band selection. However, finding an algorithm that could optimise all of the aforementioned aspects is difficult. According to the performance that one aims to optimise, the value of σ and λ could be chosen flexibly. For the least number of bands, the triangular membership function is chosen and λ is set to 0.24. For the best stability, the Gaussian membership function is chosen and σ is set to 0.16. For the highest classification accuracy, the Gaussian membership function is chosen and σ is 0.25. In this case, the stability is also the best when $l = 5$, and the number of bands is only eight.

From these results, the IM-FRS-based band selection algorithm can effectively provide stable results under perturbation of the training dataset. For application, the selected hyperspectral bands can best explain the differences between varieties.

4. Conclusion

For classification tasks, high-dimensional hyperspectral data cause problems in accuracy, efficiency and model interpretability. Effective dimensionality reduction methods are needed. In this paper, the IM-FRS algorithm was proposed to select informative bands.

The performance of the algorithm is affected by the shapes of fuzzy membership functions and the values of the parameters. When the parameters are larger than 0.3, the values of MI decrease drastically as the parameters increase. The range of 0.01–0.25 is a suitable candidate interval for the parameters. For both the Gaussian and triangular membership functions, the performance of the proposed algorithms by the ELM classifier is superior to that by the RF classifier. To solve the over-fitting problem, a post-pruning strategy was employed. The size of the subset is further downsized. The experimental results show its validity. The number of bands in the subset was reduced from 29 to 8 after post-pruning, while the average classification accuracy rose from 96.24% to 98.42% when $\lambda = 0.29$ for the triangular membership function.

For high-dimensional datasets, when conducting band selection, the stability is another significant aspect to be considered. In this paper, we constructed perturbation datasets by using an adjustable degree of overlap method instead of removing samples randomly or choosing samples by the cross-validation method. In this way, the same degree of overlap between the datasets is ensured. According to need (best classification performance, best stability or smallest subset), the values of the parameters σ and λ can be set flexibly. In other words, regardless of the membership function used, the IM-FRS-based band selection algorithm is able to effectively reduce the dimensionality.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant no. 61402399), Natural Science Foundation of Heilongjiang Province (Grant no. ZD201303), Zhanjiang Science and Technology Project (Grant no. 2017B01143), and Lingnan Normal University Nature Science Research Project (Grant no. ZL1902).

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Cheng, J. H., & Sun, D. W. (2014). Hyperspectral imaging as an effective tool for quality analysis and control of fish and other seafoods: Current research and potential applications. *Trends in Food Science & Technology, 37*(2), 78–91.

Cheng, J. H., & Sun, D. W. (2015). Recent applications of spectroscopic and hyperspectral imaging techniques with chemometric analysis for rapid inspection of microbial spoilage in muscle foods. *Comprehensive Reviews in Food Science and Food Safety, 14*(4), 478–490.

Datta, A., Ghosh, S., & Ghosh, A. (2014). Band elimination of hyperspectral imagery using partitioned band image correlation and capacitory discrimination. *International Journal of Remote Sensing, 35*(2), 554–577.

Dubois, D., & Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. *International Journal of General Systems, 17*(2–3), 191–209.

Dunne, K., Cunningham, P., & Azuaje, F. (2002). Solutions to instability problems with sequential wrapper-based approaches to feature selection. *Journal of Machine Learning Research*, 1–22.

Gowen, A. A., O'Donnell, C., Cullen, P. J., Downey, G., & Frias, J. M. (2007). Hyperspectral imaging-an emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology, 18*(12), 590–598.

Heras, D. B., Argüello, F., & Quesada-Barriuso, P. (2014). Exploring ELM-based spatial–spectral classification of hyperspectral images. *International Journal of Remote Sensing*, 35(2), 401–423.

Hu, Q., Yu, D., & Xie, Z. (2006). Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition Letters, 27*(5), 414–423.

Hu, Q., Zhang, L., An, S., Zhang, D., & Yu, D. (2012). On robust fuzzy rough set models. *IEEE Transactions on Fuzzy Systems,* 20(4), 636–651.

Hu, Q., Zhang, L., Chen, D., Pedrycz, W., & Yu, D. (2010). Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications. *International Journal of Approximate Reasoning,* 51(4), 453–471.

Imani, M., & Ghassemian, H. (2014). Band clustering-based feature extraction for classification of hyperspectral images using limited training samples. *IEEE Geoscience and Remote Sensing Letters, 11*(8), 1325–1329.

Ishimi, Y. (2009). Soybean isoflavones in bone health. In *Food factors for health promotion* (Vol. 61, pp. 104–116). Karger Publishers.

Jensen, R., & Shen, Q. (2004). Fuzzy–rough attribute reduction with application to web categorization. *Fuzzy Sets and Systems, 141*(3), 469–485.

Jensen, R., & Shen, Q. (2007). Fuzzy-rough sets assisted attribute selection. *IEEE Transactions on Fuzzy Systems*, 15(1), 73–89.

Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: A study on high-dimensional spaces. *Knowledge and Information Systems, 12*(1), 95–116.

Lin, G., Liang, J., & Qian, Y. (2015). Uncertainty measures for multigranulation approximation space. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 23*(03), 443–457.

Li, S., Qiu, J., Yang, X., Liu, H., Wan, D., & Zhu, Y. (2014). A novel approach to hyperspectral band selection based on spectral shape similarity analysis and fast branch and bound search. *Engineering Applications of Artificial Intelligence, 27*, 241–250.

Liu, H., Liu, L., & Zhang, H. (2010). Ensemble gene selection by grouping for microarray data classification. *Journal of Biomedical Informatics, 43*(1), 81–87.

Liu, Y., Xie, H., Chen, Y., Tan, K., Wang, L., & Xie, W. (2016). Neighborhood mutual information and its application on hyperspectral band selection for classification. *Chemometrics and Intelligent Laboratory Systems, 157*, 140–151.

Liu, Y., Xie, H., Tan, K., Chen, Y., Xu, Z., & Wang, L. (2016). Hyperspectral band selection based on consistency-measure of neighborhood rough set theory. *Measurement Science and Technology, 27*(5), 055501.

Liu, Y., Yang, J., Chen, Y., Tan, K., Wang, L., & Yan, X. (2017). Stability analysis of hyperspectral band selection algorithms based on neighborhood rough set theory for classification. *Chemometrics and Intelligent Laboratory Systems, 169*, 35–44.

Lyashenko, I. A., & Popov, V. L. (2015). Impact of an elastic sphere with an elastic half space revisited: Numerical analysis based on the method of dimensionality reduction. *Scientific Reports, 5*, 8479.

Messina, M. J. (2002). Soy foods and soybean isoflavones and menopausal health. *Nutrition in Clinical Care*, 5(6), 272–282.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). *How many trees in a random forest? In international workshop on machine learning and data mining in pattern recognition* (pp. 154–168). Berlin, Heidelberg: Springer.

Pantazi, X. E., Moshou, D., & Bravo, C. (2016). Active learning system for weed species recognition based on hyperspectral sensing. *Biosystems Engineering*, *146*, 193–202.

Pawlak, Z. (2002). Rough sets and intelligent data analysis. *Information Sciences, 147*(1–4), 1–12.

Qian, Y., Wang, Q., Cheng, H., Liang, J., & Dang, C. (2015). Fuzzy-rough feature selection accelerator. *Fuzzy Sets and Systems,* 258(C), 61–78.

Ravikanth, L., Singh, C. B., Jayas, D. S., & White, N. D. (2016). Performance evaluation of a model for the classification of contaminants from wheat using near-infrared hyperspectral imaging. *Biosystems Engineering, 147*, 248–258.

Suresh, S., Saraswathi, S., & Sundararajan, N. (2010). Performance enhancement of extreme learning machine for multi-category sparse data classification problems. *Engineering Applications of Artificial Intelligence, 23*(7), 1149–1157.

Wu, W. Z., & Zhang, W. X. (2004). Constructive and axiomatic approaches of fuzzy approximation operators. *Information Sciences, 159*(3–4), 233–254.

Xu, F. F., Miao, D. Q., & Wei, L. (2009). Fuzzy-rough attribute reduction via mutual information with an application to cancer classification. *Computers & Mathematics with Applications, 57*(6), 1010–1017.

Yeung, D. S., Chen, D., Tsang, E. C., Lee, J. W., & Xizhao, W. (2005). On the generalization of fuzzy rough sets. *IEEE Transactions on Fuzzy Systems, 13*(3), 343–361.

Yu, D., Hu, Q., & Wu, C. (2007). Uncertainty measures for fuzzy relations and their applications. *Applied Soft Computing, 7*(3), 1135–1143.

Zabalza, J., Ren, J., Zheng, J., Zhao, H., Qing, C., Yang, Z., et al. (2016). Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging. *Neurocomputing, 185*, 1–10.

Zhang, X., Mei, C., Chen, D., & Li, J. (2016). Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy. *Pattern Recognition, 56*(1), 1–15.

Zhao, J., Zhang, Z., Han, C., & Zhou, Z. (2015). Complement information entropy for uncertainty measure in fuzzy rough set and its applications. *Soft Computing, 19*(7), 1997–2010.