Multi-objective Based Sparse Representation Classifier for Hyperspectral Imagery Using Limited Samples

Bin Pan, Zhenwei Shi and Xia Xu

Abstract

Recent studies about hyperspectral imagery (HSI) classification usually focus on extracting more representative features or combining joint spectral-spatial information. However, besides feature extraction, developing more powerful classifiers can also contribute to the accuracies of HSI classification. In this paper, we propose a multi-objective based sparse representation classifier (MSRC) for HSI data, which mainly tries to address two problems: 1) pixel-mixing and 2) lacking abundant labeled samples. MSRC is motivated by the sparse representation classifier (SRC), and further integrating the idea of hyperspectral unmixing. Differently from traditional SRC based methods, the novelty of MSRC consists in the optimization process, *i.e.*, we directly handle the L0-norm problem without any relaxation. The sparse term is not considered as a regularization operation. Instead, we transform the problem of weight vector estimation to subset selection, and propose a multi-objective based method to optimize the L0-norm sparse problem. The residual term and sparse term are regarded as two parallel objective functions that are optimized simultaneously. We further utilize the linear mixing model to represent test pixels based on the selected atoms. The final class labels are determined according to the abundance estimation results by non-negative least squares. Owing to the characteristics of the multi-objective method and the binary property of the sparse solution vector, MSRC does not require too many training samples to build the dictionary. Moreover, theoretically, MSRC can be easily improved to extended version such as combining spatial information.

Index Terms

Hyperspectral imagery classification, multi-objective, sparse representation

The work was supported by the National Key R&D Program of China under the Grant 2017YFC1405605, the National Natural Science Foundation of China under the Grants 61671037, and the Excellence Foundation of BUAA for PhD Students under Grant 2017057. *(Corresponding author: Zhenwei Shi and Xia Xu.)*

Bin Pan, Zhenwei Shi *(Corresponding author)* and Xia Xu *(Corresponding author)* are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China, (e-mail: panbin@buaa.edu.cn; shizhenwei@buaa.edu.cn; xuxia@buaa.edu.cn).

I. INTRODUCTION

Hyperspectral sensors can provide images with hundreds of narrow continuous wavelength channels which have shown promising performance in many applications [1]–[3]. Land-cover classification is a hot theme in hyperspectral analysis [4]–[6]. However, one of the obstacles for the hyperspectral imagery (HSI) classification techniques is lacking abundant labeled samples [7]. Another challenge for HSI classification is the pixel mixing problem [8], *i.e.*, the region one pixel covers may include several materials. The former problem is because of the high costs in acquiring labeled HSI data, and the latter is caused by the low spatial resolution of most hyperspectral images.

A widely used approach to improve the classification accuracies is extracting more representative features from the original spectral data. For example, Wang *et al.* proposed a linear discriminant analysis based method to learn a representative subspace from HSI data [9]. Multiple feature fusion via ensemble methods is also investigated, *e.g.*, [10]–[13]. Some methods tried to extract deep features from HSI data via deep learning [14]–[18].

Based on the feature extraction results, researchers usually classify by means of different classifiers such as support vector machine (SVM) [19], random forest [11], [20], sparse representation classifier (SRC) [21], [22] and extreme learning machine (ELM) [23]. Some works improved the classifiers via kernel based or weighting based methods. In [24] and [25], multiple kernel learning was used for SVM. In [26] and [27], SRC was improved using weighting strategies. However, these methods are extensions of current classifiers. Essentially, the optimization processes and the objective functions are not changed.

In this paper, we develop a new classifier, multi-objective based sparse representation classifier (MSRC), which specially aims at the two problems discussed above: limited samples and mixed pixels. Our target is to achieve better performance than some popular classifiers when the number of training samples is limited. Because we focus on the calculation process in the classifier, MSRC uses the raw spectral vectors as the inputs without further feature representation. However, theoretically, MSRC can be directly applied to classify the extracted features.

The motivations of MSRC are generated from the development of sparse representation in HSI unmixing and classification. Sparse unmixing refers to representing a pixel via a linear combination of several *endmembers* using a semi-supervised manner, and then inverting the *abundance* of each endmember. Note that no labeled pixels are required here, and the semi-supervised manner means that there is an *a priori* spectral library available. To some extent, sparse unmixing can be considered as a soft classification technique [8]. However, just because of the semi-supervised property of sparse unmixing, if it were directly used for classification, the accuracies would be lower than supervised methods. Sparse representation is also investigated in HSI classification task, *i.e.*, SRC. The objective of SRC is to find a weight vector such that representation residual and sparsity error are minimized, and the predicted labels are determined by class-specific residual [22]. A regularization parameter is used to control the sparsity of the solution. The rigorous sparse term is L0-norm and NP-hard. Usually, L1-norm convex relaxation or greedy algorithms are used to solve it [6], [22], [28]–[31]. However, such strategies cannot guarantee that the obtained solution is optimal. Besides, the regularization parameter has to be set manually to a value that is usually hard to determine. Although some greedy methods such as orthogonal matching pursuit [29] can also deal with L0 problem, they are likely to fall into local optimum [32]. Moreover, the pixel mixing problem is not considered.

In this paper, we combine the ideas of HSI unmixing and classification, and develop a multi-objective based sparse representation classifier for hyperspectral data. Multi-objective optimization is a recently proposed approach that tries to overcome NP-hard problems [32]–[35]. For L0 based sparse representation problem, multi-objective methods are more likely to find the global optimal solution. In MSRC, we first build the dictionary by using a limited training set. Then instead of L1 relaxation, we directly optimize the original L0 problem via transforming it to an equivalent subset selection problem. The sparse representation problem is decomposed into two parallel objectives: representation residual and sparsity error. There is no regularization parameter required, and the two objectives are optimized simultaneously. To avoid the decision making problem and guarantee that MSRC can be applied to classification problem, we improve the multi-objective method thanks to the Tchebycheff decomposition. Differently from traditional SRC based methods, in the proposed method the obtained solution is a binary vector corresponding to the selected atoms from the dictionary. Finally, we conduct abundance inversion based on the selected atoms using the non-negative least squares (NNLS) algorithm. The predicted labels are determined by the abundance values. Compared with methods using the estimated probability values by SVM [8], [36], [37], the abundance values obtained by MSRC are not only mathematical meaningful, but also reflect the spatial distribution of materials.

The solution of MSRC is a binary vector and L0 sparse. Differently from L1 sparse, MSRC targets a subset selection problem. It is not necessary to use many atoms to represent a test pixel. Theoretically, the best number of the selected atoms should be the number of materials consisting in the test pixel. In this case, the dimensionality of the dictionary is also not required too much, as long as the overcomplete characteristic is met. Therefore, MSRC is more suitable for small training sets.

MSRC mainly aims at overcoming the following problems in HSI classification:

- Designing a new classifier which targets two typical problems in HSI classification: limited samples and mixed pixels.
- Providing an idea of optimizing the L0-norm problem directly and avoiding the selection of regularization parameters.
- Improving the multi-objective method so as to find a single solution from the Pareto front.

In Section II we describe the objective functions as well as the optimization process of MSRC. In Section III comparison results with some popular classifiers are displayed. The conclusion is drawn in Section IV.

II. The proposed method

A. Background

SRC is a popular classifier and has been applied to HSI classification [21], [22], [26], [27]. SRC assumes that a test sample can be represented by a (sparse) linear combination of atoms from an overcomplete training dictionary, *i.e.*,

$$\hat{\boldsymbol{s}} = \arg\min \|\boldsymbol{s}\|_0, \quad s.t. \quad \mathbf{x} = \mathbf{A}\boldsymbol{s}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{L \times N}$ is an overcomplete dictionary which is obtained by training samples, L and N are the numbers of bands and atoms respectively, $\mathbf{x}$ is a test sample, and $\boldsymbol{s} \in \mathbb{R}^{N \times 1}$ is a sparse weight vector. However, Eq. (1) is NP-hard and very difficult to solve. Usually, this problem is relaxed to L1-norm form with an error tolerance ϵ:

$$\hat{\boldsymbol{s}} = \arg\min \|\boldsymbol{s}\|_1, \quad s.t. \quad \|\mathbf{x} - \mathbf{A}\boldsymbol{s}\|_2^2 \leq \epsilon. \tag{2}$$

Here, the solution $\boldsymbol{s}$ is an L1 sparse vector. Eq. (2) can be transformed to the following convex optimization problem:

$$\arg\min_{\boldsymbol{s}} \|\mathbf{x} - \mathbf{A}\boldsymbol{s}\|_2^2 + \lambda \|\boldsymbol{s}\|_1 \tag{3}$$

where λ is a regularization parameter used to control the tradeoff between sparsity and the reconstruction error. Eq. (3) is the general form of SRC [22].

There are two problems in Eq. (3) that can be further discussed. Firstly, L1-norm is just a relaxation strategy that can only generate approximate solutions. Secondly, the determination of parameter λ is an open problem.

B. Objective Functions in MSRC

In MSRC, the above two problems are addressed under a multi-objective framework. MSRC tries to optimize $\mathcal{F}(\boldsymbol{s})$ which contains two parallel objective functions:

$$\arg\min_{\boldsymbol{s}} \mathcal{F}(\boldsymbol{s}) = [f_1(\boldsymbol{s}), f_2(\boldsymbol{s})]$$
$$\begin{cases} f_1(\boldsymbol{s}) = \|\mathbf{x} - \mathbf{A}\boldsymbol{s}\|_2^2 \\ f_2(\boldsymbol{s}) = \|\boldsymbol{s}\|_0 \end{cases} \tag{4}$$

$f_1(\cdot)$ is the reconstruction error and $f_2(\cdot)$ is the sparsity, and they are optimized simultaneously. Therefore there is no regularization parameter required. Dictionary $\mathbf{A}$ is built by training samples, and the solution $\boldsymbol{s}$ is a weight vector for each test sample.

However, Eq. (4) is still difficult to solve because of the non-convex L0 problem in $f_2(\cdot)$. On the other hand, this is also the key point that may improve the representative ability of the classifier.

In MSRC, we consider the physical significance of HSI, and propose a two-step solving method. The physical significance of Eq. (4) is explicit. Spectrum $\mathbf{x}$ is composed of several endmembers which construct a subset of spectral library $\mathbf{A}$, and $\boldsymbol{s}$ is the fractional abundance of the endmembers. Obviously, $\boldsymbol{s}$ should be L0 sparse because the region of one pixel cannot include too many materials. Inspired by this physical meaning, in MSRC we transform Eq. (4) to an equivalent form:

$$\arg\min_{\boldsymbol{\alpha}} \mathcal{F}(\boldsymbol{\alpha}) = [f_1(\boldsymbol{\alpha}), f_2(\boldsymbol{\alpha})]$$
$$\begin{cases} f_1(\boldsymbol{\alpha}) = \|\mathbf{x} - \mathbf{A}(\boldsymbol{\alpha} \odot \boldsymbol{\beta})\|_2^2 \\ f_2(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_0 \end{cases} \tag{5}$$

where $\boldsymbol{s} = \boldsymbol{\alpha} \odot \boldsymbol{\beta}$. The most significant difference between Eq. (4) and Eq. (5) is that $\boldsymbol{\alpha}$ is a binary vector, where the locations correspond to the selected atoms are 1 and others are 0. $\boldsymbol{\beta} \in \mathbb{R}^{N \times 1}$ is an augmented abundance vector.

In this case, Eq. (5) is a subset selection problem. More specific, $f_1(\cdot)$ is equivalent to the following form:

$$\begin{aligned} & f_1(\boldsymbol{\alpha}) = \|\mathbf{x} - (\mathbf{AM})\boldsymbol{\beta}\|_2^2 \\ & \text{where} \\ & (\mathbf{M}_{ij})_{N\times N} = \begin{cases} \alpha_i, & i = j \\ 0, & otherwise. \end{cases} \end{aligned} \tag{6}$$

Mathematically, not all the elements in $\boldsymbol{\beta}$ are valuable. Only the elements corresponding to the non-zero columns of $\mathbf{AM}$ (also the non-zero rows of $\boldsymbol{\alpha}$) are the required abundance. So $\boldsymbol{\beta}$ is an augmented abundance vector.

After the above transformation, we can solve Eq. (5) instead of Eq. (4) while keeping the final solution consistent. Because the solution of Eq. (5) is a binary vector rather than a weight vector, the solution space has dramatically shrunk, and thus the solving difficulty is reduced. In this paper, we use a two-step approach to solve Eq. (5): first optimizing $\boldsymbol{\alpha}$ and then calculating $\boldsymbol{\beta}$. To solve the L0 problem in the first step, we propose an improved multi-objective based method. After determining $\boldsymbol{\alpha}$, NNLS is used to invert $\boldsymbol{\beta}$. Note that although we use a two-step approach to obtain $\boldsymbol{\alpha}$, MSRC can still simultaneously optimize $f_1(\cdot)$ and $f_2(\cdot)$ without setting regularization parameters. Actually only $\boldsymbol{\alpha}$ is the optimization variable. Once $\boldsymbol{\alpha}$ is determined, $\boldsymbol{\beta}$ can be directly obtained by NNLS.

Transforming the L1-norm in Eq. (3) to the L0-norm in Eq. (5) is the key operation in handling limited samples problem. L1-norm utilizes quite a few atoms in the dictionary to represent test samples. When the training set shrinks, the available information by L1-norm will synchronously reduce. By comparison, L0-norm, which is a subset selection problem, uses several most representative atoms to reconstruct test samples. An advantage of multi-objective based methods in solving subset selection problem is that they can achieve comparative performance with small dictionary volume [38], [39]. Therefore L0-norm as well as the MO-based method are used to address the limited samples problem. Besides, using NNLS instead of class-specific residual is the strategy for relieving mixed pixel problem, which is widely adopted in unmixing-related literatures [40].

Based on the characteristics of HSI classification, we use a trick to further modify the objectives. In Eq. (5) $\mathbf{A}$ should contain a certain number of samples in each class. Let $\mathbf{A}_c \in \mathbb{R}^{L\times N_c}$ denote a subset of $\mathbf{A}$ with N_c atoms (training samples) from class c. Theoretically, suppose test pixel $\mathbf{x}$ belongs to class c, it may be represented by several atoms in $\mathbf{A}_c$. In this case, to enhance the diversity of the solution, we set a small sparsity k in $f_2(\cdot)$. Then $f_2(\cdot)$ is modified by

$$f_2(\boldsymbol{\alpha}) = |k - \|\boldsymbol{\alpha}\|_0| \,. \tag{7}$$

The physical meaning of k is the number of atoms that are used to represent $\mathbf{x}$. In L0 based sparse representation problem, k should not be ignored [21]. However, literature [21] also indicated that k has little influence as long as it is not very small. In the classification task, the upper bound of k should be N_c. In this paper we simply set k as N_c.

C. Multi-Objective based Optimization Approach

In this section, we introduce the optimization process for Eq. (5) based on the proposed MSRC method. MSRC is developed under the framework of multi-objective evolutionary algorithm based on decomposition (MOEA/D) [41]. MOEA/D is an evolutionary based multi-objective optimization method which is an available approach to solve the subset selection problem.

Multi-objective optimization problems involve several conflicting objectives. It is hard, usually impossible, to find a solution that is optimal for all objectives. Thus a balance among them is preferred, which leads to a set of non-dominated solutions in multi-objective optimization. A general multiple objective problem is expressed by

$$\begin{aligned} &\arg\min_{\mathbf{u}} \, [f_1(\mathbf{u}), f_2(\mathbf{u}), ..., f_n(\mathbf{u})] \\ &s.t. \ \mathbf{u} \in \Omega \subseteq \Re^n \end{aligned} \tag{8}$$

Here $\mathbf{u}$ is the variable, n is the number of objectives, Ω is the feasible region in the decision variable space $\Re^n$.

In MSRC, $n = 2$ and the variable is $\boldsymbol{\alpha}$. A basic idea of most evolutionary based methods is initializing many solutions and forcing them to search their corresponding local optimums. Then the best local optimum can be considered as global optimum. This idea is adopted by MOEA/D and MSRC. Firstly, we randomly initialize the solution set $\mathfrak{A} = \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_p\}$ which is called *population*, and $\boldsymbol{\alpha}_i$ is called a *individual*. For each $\boldsymbol{\alpha}_i$ its corresponding $\mathcal{F}(\boldsymbol{\alpha}_i)$ is called a *subproblem*. All the subproblems are optimized simultaneously, and the best individual is selected as the final solution. However, in multi-objective problem, it is hard to determine which individual is the best. For example, there may consist the condition that $f_1(\boldsymbol{\alpha}_i) > f_1(\boldsymbol{\alpha}_j)$ while $f_2(\boldsymbol{\alpha}_i) < f_2(\boldsymbol{\alpha}_j)$. Thus the results of multi-objective methods are usually still solution sets, called *Pareto front*. One of the contributions of MSRC is finding a single solution from the Pareto front.

Then there are two challenges for multi-objective problems: 1) How to update $\boldsymbol{\alpha}_i$, and 2) how to evaluate superiority between the original and the updated results. Obviously, gradient descent methods cannot work because $f_2(\cdot)$ is discrete and non-convex. In MOEA/D, random flipping strategy is used to update $\boldsymbol{\alpha}_i$, and the weighted Tchebycheff distance between individuals and an ideal point is used to evaluate the superiority. The ideal point in MOEA/D is virtual and defined by

$$\mathbf{z}_M^* = (\min\{f_1(\boldsymbol{\alpha_i})\}, \min\{f_2(\boldsymbol{\alpha_i})\}). \tag{9}$$

However, the solution by the original MOEA/D is a set, *i.e.*, Pareto front. In the task of classification, it should be guaranteed that there is a unique solution for each test sample. In order to solve this decision making problem, we integrate it into the evolution process of MOEA/D. In MSRC, we define the ideal point $\mathbf{z}^*$ as the individual that has the smallest Euclidean distance to the original point:

$$\begin{aligned} &\mathbf{z}^* = (z_1, z_2) = (f_1(\boldsymbol{\alpha}_{i^*}), f_2(\boldsymbol{\alpha}_{i^*})), \\ &\text{where} \quad i^* = \arg\min_i \|\mathcal{F}(\boldsymbol{\alpha}_i)\|_2. \end{aligned} \tag{10}$$

We will further analyze the advantage of this improvement in the following description. Here we continue focusing on the above two problems: how to update and how to evaluate.

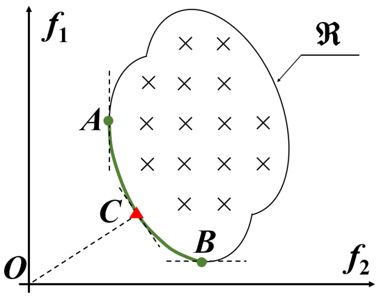


Fig. 1: A simple illustration of the difference between MOEA/D and our improved method. $\overset{\frown}{AB}$ is the Pareto front in MOEA/D, and C is the unique solution of the proposed method.

MSRC uses an iterative approach to optimize each subproblem in parallel. Take an individual $\boldsymbol{\alpha}_i$ for example. In a single iteration, $\boldsymbol{\alpha}_i$ is changed to $\boldsymbol{\alpha}_i'$ based on random flapping [42]. Then we compare the weighted Tchebycheff distances [42] between $\boldsymbol{\alpha}_i/\boldsymbol{\alpha}_i'$ and $\mathbf{z}^*$. If $\boldsymbol{\alpha}_i'$ presents closer distance to $\mathbf{z}^*$, we consider that $\boldsymbol{\alpha}_i'$ is better and it will replace $\boldsymbol{\alpha}_i$ in the next iteration, or $\boldsymbol{\alpha}_i$ remains unchanged. The weighted Tchebycheff distance between $\boldsymbol{\alpha}_i$ and $\mathbf{z}^*$ is expressed as follows

$$g_i^{te}(\boldsymbol{\alpha}_i|\boldsymbol{\lambda}_i, \mathbf{z}^*) = \max_{1\leq j\leq 2}\{\lambda_i^j|f_j(\boldsymbol{\alpha}_i) - z_j^*|\}, \tag{11}$$

where $\boldsymbol{\lambda}_i = [\lambda_i^1, \lambda_i^2]^{\mathrm{T}}$ is the weight vector of individual $\boldsymbol{\alpha}_i$ with $\boldsymbol{\lambda}_i \geq 0$ and $\lambda_1 + \lambda_2 = 1$. For each individual $\boldsymbol{\alpha}_i$, its corresponding $\boldsymbol{\lambda}_i$ is randomly set. The weighted Tchebycheff distance is popular in multi-objective methods because it can avoid the dimensional imbalance problem among different objectives. We recommend readers learn more details about this point from literature [42].

In the next iteration, the ideal point, as well as some individuals, will be updated according to Eq. (10) and Eq. (11). Overall, the population tend to become optimal with the increasing of iterations.

Fig. 1 illustrates the difference between the original MOEA/D and our improved methods. $\mathfrak{R}$ is the range set and it could be either convex or non-convex. The curve $\overset{\frown}{AB}$ is the Pareto front in MOEA/D, while C is the final solution by the proposed method. We can see that C is located on the Pareto front, which guarantees the effectiveness of the solution. Note that Fig. 1 is only an intuitive illustration rather than real data distribution. Actually it is impossible to depict the real distribution of $\mathfrak{R}$. The shape of $\mathfrak{R}$ may be very complex. In MSRC, $\mathfrak{R}$ is discrete (many parallel lines). However, the optimization process in MSRC is not affected by such situations.

It is worth noting that MSRC has significant differences compared with some existing methods such as [32]. The work [32] used NSGA-II [43] framework for hyperspectral unmixing, while MSRC uses MOEA/D for classification. An essential difference between them consists in the determination about the superiority of current solutions. MOEA/D uses Tchebycheff distance among current solutions and ideal point to determine which solutions are optimal, while in NSGA-II non-dominated sorting and crowding distance were used. The term "ideal point" does not consist in NSGA-II. In MSRC, one of the major contributions is finding a single solution, and the above improvement can only conduct on MOEA/D framework.

The detailed process of MSRC is shown in **Algorithm** 1. The training samples $\mathbf{X}_{tr}$ are collected as a dictionary $\mathbf{A}$ (line 2). After initialization, the dictionary $\mathbf{A}$ is represented by binary coding (line 6). Each atom is assigned a binary code, where '1' denotes the corresponding atom is selected and '0' otherwise. Note that each atom in $\mathbf{A}$ corresponds to a training sample. The optimization process begins with generation an initial population (line 7). The individuals in the population can provide a parallel search in the feasible region, which contribute to finding multiple mutually nondominated solutions in a single run. Then a reference point in the current population with the smallest distance to the original point is determined (line 8, $\boldsymbol{\alpha}^*$) which has the smallest distance to the original point. At each iteration, MOEA/D decomposes the atom selection problem into several single-objective optimization subproblems, and divides all individuals into a number of overlapping groups. Specifically, the i-th individual is the current solution of the i-th subproblem. Each subproblem and its neighborhood have relatively close weight vectors. In the evolution process, each subproblem updates the solution according to the weighted Tchebycheff problem in Eq. (11). Firstly, a new individual is generated based on randomly flipping, where each bit is flipped with a probability '1/m' and remains unchanged with '1-1/m' (line 11, $\boldsymbol{\alpha}_i^{'}$). The flipping probability should be small so as to avoid changing too much in a single iteration. Secondly, calculate the Tchebycheff distance between $\boldsymbol{\alpha}_i^{'}$ and $\boldsymbol{\alpha}^*$, denoted by $g_i^{st}(\boldsymbol{\alpha}_i^{'}|\boldsymbol{\lambda}_i, \boldsymbol{\alpha}^*)$. Then compare $g_i^{st}(\boldsymbol{\alpha}_i^{'}|\boldsymbol{\lambda}_i, \boldsymbol{\alpha}^*)$ with the distances between $\boldsymbol{\alpha}^*$ and all the individuals in this group (line 12-13). Individuals that present larger distance are replaced by $\boldsymbol{\alpha}_i^{'}$. This algorithm is conducted by T iterations. After the final solution $\boldsymbol{\alpha}$ is determined, the abundance is estimated based on Eq. (13) and the label is determined using Eq. (14).

D. Determination of the Predicted Labels

Instead of using minimal class-specific reconstruction residual, in MSRC the test labels are determined based on abundance inversion. The i-th test sample $\mathbf{x}_i$ can be represented by $\mathbf{A}_i \in \mathbb{R}^{L\times|\boldsymbol{\alpha}_i|}$ which contains the selected atoms from $\mathbf{A}$ based on $\boldsymbol{\alpha}_i$. The columns of $\mathbf{A}_i$ are the so-called endmembers in hyperspectral unmixing. Differently from unmixing problem, endmembers in MSRC are composed of labeled samples with higher quality. According to the characteristics of HSI data, the endmember fractions should be greater than zero and sum-to-one. Therefore in MSRC, if one defines $\boldsymbol{\beta}_i \in \mathbb{R}^{|\boldsymbol{\alpha}_i|\times 1}$ as the abundance vector for $\mathbf{A}_i$, then it can be calculated by

$$\arg\min_{\boldsymbol{\beta}_i} \|\mathbf{x}_i - \mathbf{A}_i\boldsymbol{\beta}_i\|_2^2, \quad s.t. \quad \boldsymbol{\beta}_i \geq 0, \quad |\boldsymbol{\beta}| = 1. \tag{12}$$

Note that $\boldsymbol{\beta}_i$ is a subset of the $\boldsymbol{\beta}$ in Eq. (5), where the non-zero indices in $\boldsymbol{\alpha}_i$ are reserved. To solve Eq. (12), a rigorous method is fully constrained least squares (FCLS) [44]. However, FCLS is sensitive to the number of selected atoms and their signature accuracies. According to [45], a more robust approach is using the non-negative constraint alone:

$$\arg\min_{\boldsymbol{\beta}_i} \|\mathbf{x}_i - \mathbf{A}_i\boldsymbol{\beta}_i\|_2^2, \quad s.t. \quad \boldsymbol{\beta}_i \geq 0. \tag{13}$$

Eq. (13) can be solved by NNLS which is adopted in the proposed method. The class label y_i for $\mathbf{x}_i$ is then determined according to the maximal class-specific abundance:

$$y_i = \arg\max_{c} |\boldsymbol{\beta}_i^c|, \quad c \in [1, 2, \cdots, C]. \tag{14}$$

Algorithm 1: MSRC for classification

```
Input: Training samples $\mathbf{X}_{tr}$, training labels $\mathbf{y}_{tr}$, test sample $\mathbf{x}$, $k$
Output: Predicted label $y$
Preprocessing:
Build dictionary $\mathbf{A}$ using $\mathbf{X}_{tr}$;
Initialization:
population size $p$, maximum iteration number $T$, a set of weight vector $\mathbf{\Lambda} = \{\boldsymbol{\lambda}_1, ..., \boldsymbol{\lambda}_p\}$, indexes of each subproblem's neighbors $\{B_1, ..., B_p\}$.
Atoms Selection:
Represent $\mathbf{A}$ by binary coding;
Generate a population $\mathfrak{A} = \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_p\}$;
Calculate current optimal solution $\boldsymbol{\alpha}^*$;
while $t < T$ do
    for $i = 1, ..., p$ do
        Flip each elements in $\boldsymbol{\alpha}_i$ with probability $1/m$, and obtain a new individual $\boldsymbol{\alpha}_i^{'}$;
        if $\|\mathcal{F}(\boldsymbol{\alpha}^*)\|_2^2 > \|\mathcal{F}(\boldsymbol{\alpha}_i^{'})\|_2^2$ then
            Set $\boldsymbol{\alpha}^* = \mathcal{F}(\boldsymbol{\alpha}_i^{'})$
        for $j \in B_i$ do
            if $g_i^{st}(\boldsymbol{\alpha}_i^{'}|\boldsymbol{\lambda}_i, \boldsymbol{\alpha}^*) \leq g_j^{st}(\boldsymbol{\alpha}_j|\boldsymbol{\lambda}_j, \boldsymbol{\alpha}^*)$ then
                Set $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}_i^{'}$ and $F(\boldsymbol{\alpha}_j) = F(\boldsymbol{\alpha}_i^{'})$
    Update the population;
    $t = t + 1$
Return the solution $\boldsymbol{\alpha}^*$ and record the corresponding spectral signatures.
Label Determination:
Abundance estimation based on Eq. (13);
Label determination based on Eq. (14)
```

$\boldsymbol{\beta}_i^c$ is a subset of $\boldsymbol{\beta}_i$, which denotes the elements corresponds to class c. C is the total number of classes.

An extra advantage of the proposed labels determination approach is that it is robust to different training samples, *i.e.*, changing training samples has little influence on the final results. This phenomenon is easy to explain: It is caused by the L0-based optimization process. Any pixel $[\mathbf{x}_i, y_i = c]$ is represented by several atoms from the dictionary with a binary weight vector. Because the weight vector is binary, the solution space is much smaller than that in L1 problem. This situation is more apparent when the dimensionality of $\mathbf{A}$ is low. After sufficient iterations the results of MSRC tend to shrink to a global optimal solution. Therefore, in MSRC the results by repeated experiments have little variation.

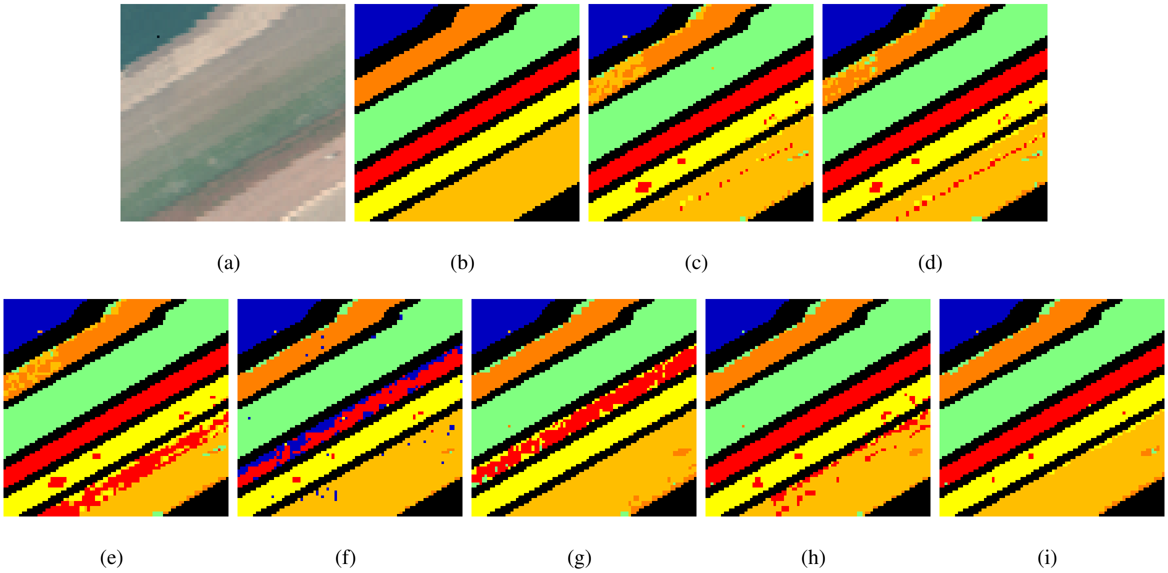


Fig. 2: SalinasA data set. (a) The false-color composite image. (b) The ground truth where each color corresponds to one land-cover material. Classification maps by (c) SVM, (d) ELM, (e) KNN, (f) SRC, (g) CRC, (h) CRT and (i) MSRC.

III. Experiments and Discussion

In this section, we compare the proposed method with some state-of-the-art classifiers. Since MSRC is developed under the framework of sparse representation, three representation based classifiers are compared, *i.e.*, SRC (L1-norm, single objective) [22], [46], collaborative representation-based classification (CRC) [22], [47] and collaborative representation with Tikhonov regularization (CRT) [48]. In addition, three popular classifiers, SVM, ELM and KNN are also used for comparison.

Three public hyperspectral data sets are analysed in the experiments, and the spectral vectors are directly used as the features. Because we focus on the limited-training-samples conditions, we only randomly select 5 pixels from each class for training, and the rests for testing. Correspondingly, the sparsity k in Eq. 7 is set as 5. All the methods are conducted 30 times, and the average results are reported. Three widely used metrics, overall accuracy (OA), average accuracy (AA) and kappa coefficient (κ), are used for evaluation. As a general classifier, MSRC has few parameters for tuning. This is one of its advantages. Here we mainly discuss the influence of iterations.

A. Data Sets

Three public HSI data sets are used for comparison, namely SalinasA, Kennedy Space Center (KSC) and Botswana. All of them are online available[1].

[1] Available online: http://www.ehu.eus/ccwintco/index.php?title=

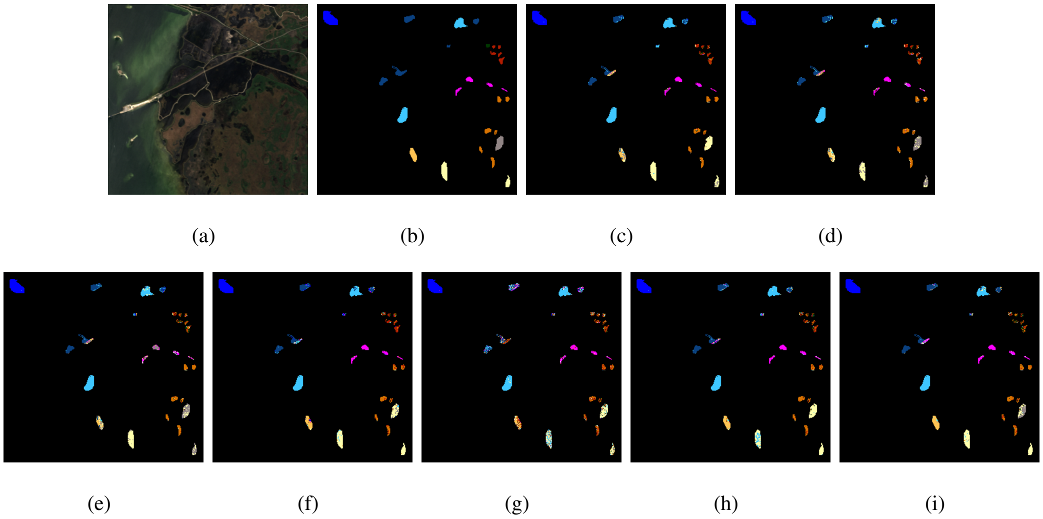


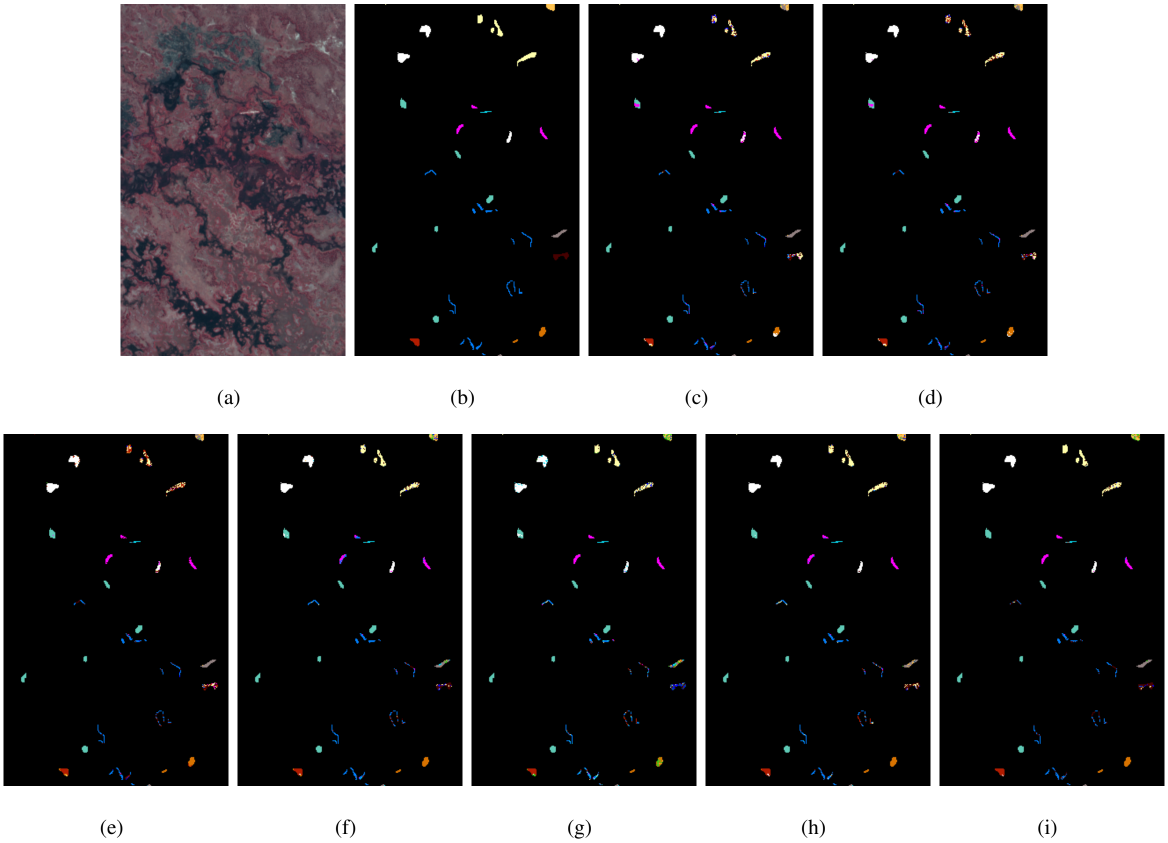
Fig. 3: Part of the KSC data set. (a) The false-color composite image. (b) The ground truth where each color corresponds to one land-cover material. Classification maps by (c) SVM, (d) ELM, (e) KNN, (f) SRC, (g) CRC, (h) CRT and (i) MSRC.

1) SalinasA: This data set is part of the Salinas scene with 86×83 pixels size, which is collected by the AVIRIS sensor over Salinas valley, California. This data contain 224 bands with 3.7m spatial resolution. 20 water absorption bands are removed, namely [108-112], [154-167], 224. Totally 5348 labeled pixels belonging to 6 land-cover classes are observed in the ground truth image. Fig. 2(a)(b) show the false color (R-G-B=bands 37-17-11) and the ground truth images of this data set. Considering the computational cost, here we do not use the whole scene of Salinas data.

2) KSC: KSC data set is collected by the NASA AVIRIS instrument over the Kennedy Space Center in 1996. 224 bands are acquired with wavelengths from 400 to 2500nm. After removing water absorption and noisy bands, 176 bands were used for the analysis. Compared with SalinasA, the spatial resolution of KSC is much lower (18m). Thus the pixel mixing problem is more serious. Totally 5211 pixels are labeled which are separated into 13 classes. Fig. 3(a)(b) display the false color (R-G-B=bands 28-9-10) and the ground truth images of this data set. Since the size of this data is relatively large (512 × 614), Fig. 3 only shows part of the whole scene.

3) Botswana: It is acquired by the Hyperion sensor which is carried by the NASA EO-1 satellite, over the Okavango Delta, Botswana, in 2001. This data set has 30m spatial resolution and 10nm spectral resolution. The original data contains 242 bands covering the wavelength 400-2500nm. After removing uncalibrated and noisy bands, 145 channels remain. There are 3248 labeled pixels available which consist of 14 identified classes. A false color scene (R-G-B=bands 45-26-17) and the ground truth images are shown in Fig. 4(a)(b). Due to the large size of this data (1476 × 256), Fig. 4 also shows part of the whole scene.

Hyperspectral_Remote_Sensing_Scenes



(a) (b) (c) (d)

(e) (f) (g) (h) (i)

Fig. 4: Part of the Botswana data set. (a) The false-color composite image. (b) The ground truth where each color corresponds to one land-cover material. Classification maps by (c) SVM, (d) ELM, (e) KNN, (f) SRC, (g) CRC, (h) CRT and (i) MSRC.

TABLE I: CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON SALINASA DATA SET (%).

Class	Samples		Methods						
	Train	Test	SVM	ELM	KNN	SRC	CRC	CRT	MSRC
C1	5	386	99.48±0.01	100.0±0.00	100.0±0.00	99.71±0.08	99.84±0.13	99.74±0.01	100.0
C2	5	1338	86.34±20.1	80.89±18.8	87.69±11.9	90.98±10.8	93.39±4.55	91.12±13.4	97.31
C3	5	611	90.09±7.89	92.57±4.44	91.84±4.64	94.54±3.29	92.52±1.38	92.32±3.27	94.43
C4	5	1520	98.32±2.01	99.37±0.77	98.34±4.65	97.06±6.08	97.41±4.65	94.92±11.9	98.94
C5	5	669	99.65±0.21	99.79±0.22	99.47±1.16	99.35±0.28	51.12±24.3	99.76±0.14	93.27
C6	5	794	96.17±2.41	97.16±0.17	95.84±2.75	95.65±5.18	99.43±0.36	98.86±0.51	94.45
OA			94.29±4.58	93.71±4.51	94.41±3.55	95.51±3.35	90.49±4.37	95.21±5.76	**96.71**
AA			95.01±3.01	94.96±2.93	95.19±2.42	96.22±2.26	88.95±4.66	96.12±3.33	**96.40**
κ			92.92±5.59	92.20±5.50	93.05±4.36	94.43±4.11	88.11±5.42	94.10±7.00	**95.87**

TABLE II: CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON KSC DATA SET (%).

Class	Samples		Methods						
	Train	Test	SVM	ELM	KNN	SRC	CRC	CRT	MSRC
C1	5	756	73.92±20.6	75.93±9.17	83.30±10.5	81.28±8.12	61.82±18.9	77.60±14.2	82.12
C2	5	238	80.84±9.00	79.36±7.72	70.96±12.1	76.59±6.06	54.36±10.2	79.28±5.81	70.81
C3	5	251	80.71±15.1	82.98±8.01	81.07±6.71	86.45±9.99	55.65±13.6	81.11±11.0	87.39
C4	5	247	40.61±22.7	45.38±15.7	45.99±10.7	39.67±13.6	36.39±11.3	42.10±14.6	66.94
C5	5	156	47.56±21.8	68.07±9.42	53.39±8.53	54.23±13.8	44.93±11.2	58.33±15.5	64.97
C6	5	224	50.17±14.7	47.81±10.4	44.24±9.37	47.09±8.42	44.15±12.4	50.00±10.3	74.59
C7	5	100	72.00±21.4	82.90±10.5	77.50±6.25	92.00±7.90	83.90±8.94	93.80±4.41	95.00
C8	5	426	44.67±18.5	64.53±9.77	57.04±9.30	75.02±9.80	43.84±10.5	75.96±10.8	69.46
C9	5	515	80.77±8.89	80.25±11.4	79.45±7.19	94.60±5.87	57.10±13.2	93.78±4.43	74.35
C10	5	399	68.49±19.6	56.89±11.5	60.78±9.58	83.59±7.52	57.79±9.90	86.84±6.00	83.96
C11	5	414	83.16±11.8	92.43±4.87	92.15±5.11	90.12±4.14	85.14±5.89	89.27±4.05	95.08
C12	5	488	74.89±8.11	67.18±10.9	71.14±4.33	66.74±5.15	41.68±11.3	72.31±4.25	82.73
C13	5	922	95.76±5.34	97.89±1.95	98.24±0.51	99.11±0.73	98.85±0.78	98.10±0.51	99.57
OA			73.70±3.46	76.07±1.83	76.02±1.95	80.74±1.61	63.55±3.68	80.97±2.47	**83.08**
AA			68.73±2.83	72.43±1.60	70.40±1.86	75.91±1.49	59.05±2.69	76.80±1.72	**80.53**
κ			70.83±3.74	73.41±2.00	73.34±2.12	78.56±1.77	59.72±3.92	78.87±2.69	**81.19**

TABLE III: CLASSIFICATION ACCURACIES OF DIFFERENT METHODS ON BOTSWANA DATA SET (%).

Class	Samples		Methods						
	Train	Test	SVM	ELM	KNN	SRC	CRC	CRT	MSRC
C1	5	265	98.67±1.39	98.90±1.27	99.32±1.21	98.64±1.09	94.07±3.01	99.24±1.15	98.49
C2	5	96	90.20±16.1	92.50±8.68	85.41±13.6	93.43±10.4	95.62±4.35	95.10±7.65	93.75
C3	5	246	90.48±7.83	93.65±4.73	88.17±6.60	94.59±2.44	70.81±12.7	94.30±3.11	75.20
C4	5	210	86.28±8.63	93.52±2.36	89.61±3.77	95.38±4.91	89.80±9.01	96.00±5.76	97.14
C5	5	264	71.62±11.2	66.74±8.66	65.71±7.26	76.36±10.5	59.77±17.2	71.74±9.66	66.28
C6	5	264	53.71±16.2	51.59±10.2	52.42±8.50	59.50±16.2	34.65±12.4	55.64±13.2	64.77
C7	5	254	86.69±12.3	87.67±9.44	94.44±2.96	97.79±1.84	80.43±9.00	98.36±1.31	96.06
C8	5	198	85.30±14.7	94.69±5.36	85.75±8.14	88.28±9.63	77.62±8.73	93.53±8.10	98.99
C9	5	309	60.12±13.6	72.65±13.3	68.89±9.17	79.12±8.55	62.97±11.5	67.34±12.5	83.49
C10	5	243	81.93±9.96	75.59±13.9	67.32±9.16	68.55±9.25	36.21±12.3	69.25±6.43	90.53
C11	5	300	84.46±10.8	85.00±7.35	77.50±7.52	86.00±8.28	84.80±9.39	85.16±6.95	93.33
C12	5	176	93.46±5.78	92.67±5.41	91.07±5.39	96.36±2.04	91.70±12.5	95.39±3.16	97.16
C13	5	263	75.77±9.34	76.76±9.30	70.34±10.0	77.03±10.5	37.41±15.7	75.81±5.80	80.99
C14	5	90	95.89±6.98	97.66±3.07	98.77±1.33	94.88±6.28	99.66±1.05	99.00±0.81	93.33
OA			80.56±2.46	82.32±1.87	79.14±1.96	84.72±1.86	69.15±2.68	83.30±1.78	**86.60**
AA			82.47±2.69	84.26±1.82	81.05±2.12	86.14±1.62	72.54±2.14	85.41±1.35	**87.82**
κ			78.96±2.67	80.86±2.03	77.40±2.13	83.46±2.01	66.72±2.86	81.92±1.92	**85.49**

B. Classification Results

The classification results on the three data sets are shown in Fig. 2-4 and Table I-III. To further validate that the improvements by MSRC are statistically significant, we conduct two-sample t-test (t-test for short) on OA, AA and κ. T-test is a popular approach which is able to validate the significance between two groups of data [49]. It is defined by

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}} > t_{1-\alpha}[n_1+n_2-2], \tag{15}$$

where $\bar{x}_1$ and $\bar{x}_2$ are mean accuracy values obtained by different methods, s_1 and s_2 are the corresponding standard variations, n_1 and n_2 are the numbers of repeated experiments, and α is the significance level.

1) Results on SalinasA: Because no spatial information is used in the feature extraction process, there are many isolated points observed in Fig. 2(c)-(i). The objective evaluation results are shown in Table I. As discussed in Section II-D, results of MSRC are almost impervious to repeated experiments. Hence we report the mode after 30 runs. Since MSRC is developed under the basis of SRC, the comparison with SRC is the most meaningful. MSRC achieves about 1% advantages in OA and κ, and has a small decrease in the AA. The advantages are not very significant, because the accuracy values for this data set are relatively high. In this case a small increase also requires many efforts. In addition, there are only 6 classes in this data set, the accuracy in single class has considerable influence on AA. MSRC presents much lower accuracy than SRC in class 5 with a substantial decrease of the final AA. CRC is an efficient method which has a L2-norm objective and closed-form solutions. However, limited samples may aggravate the underdetermined phenomenon in solving CRC and thus harm the accuracies. Popular classifiers such as SVM and ELM present similar accuracies with about 2% gaps against MSRC. KNN uses spectral distance without optimization process to selected atoms, and then votes for decision. To some extent, KNN could be considered as a much simplified MSRC. If we replace the optimization process in MSRC by a simple spectral distance calculation, and use max voting instead of abundance inversion, MSRC is transformed into KNN. However, it is observed that MSRC presents significant advantages over KNN. However, efficiency is not among these advantages. Compared with SRC, MSRC requires about 20 times computational cost. How to improve the efficiency is our next priority.

2) Results on KSC: The accuracies in this data set is much lower than those in SalinasA. SalinasA has relatively smaller size, which means materials' intra-class similarities are higher. While in KSC, samples are selected from different locations of the whole scene, as shown in Fig. 3(b). Fig. 3(c)-(i) and Table II show the evaluation results of all the 7 methods. Approximately 2-4% gaps are observed between MSRC and the compared methods. The advantages in this data are more apparent maybe because the baseline in KSC is much lower than that in SalinasA. CRT is a weighted version of CRC, and the results in Table II has demonstrated that this improvement really works in the case of limited samples. SVM, ELM and KNN have shown very close accuracies, but MSRC outperforms them by nearly 7%. T-test by Eq. (15) indicates that the advantages obtained by MSRC are statistically significant under 95% level.

3) Results on Botswana: Fig. 4(c)-(i) show the classification maps of different methods. Note that because of the large size of this data set, Fig. 4 displays only part of the whole scene. We can see that the distribution of labeled pixels is scattered, which may weaken the effect of spatial information. Table III presents the objective evaluation on this data set. Most of the methods achieve over 80% OA, and MSRC slightly outperforms others by 2-6%. According to the results by t-test, these gaps are also statistically significant. Due to the lack of training samples, CRC still does not perform well. In most classes, MSRC presents over 80% accuracies. We also note that the spatial resolution in Botswana data is lower (30m), which means the mixing pixels problem is more serious. However, MSRC still works well.

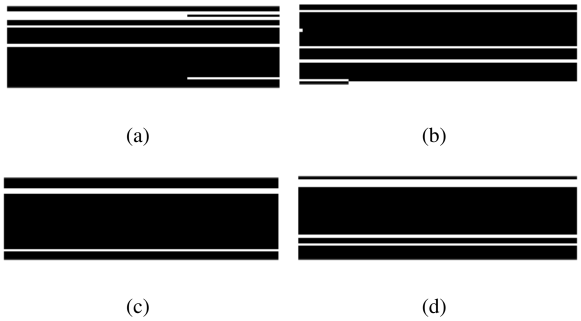
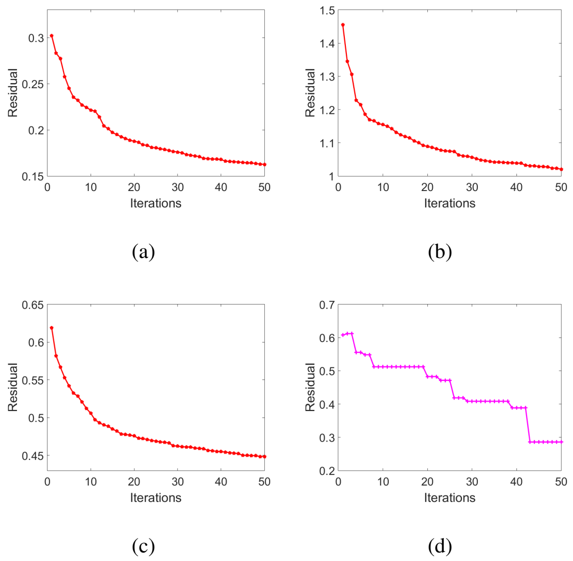


Fig. 5: Pareto solution set by different iterations for a randomly selected samples in SalinasA data set (label=1). The horizontal axis denotes individuals in Pareto solution set, and the ordinate axis denotes atoms in the dictionary. (a) Iterations=1, k=3. (b) Iterations=10, k=3. (c) Iterations=100, k=3. (d) Iterations=100, k=5.

C. Analysis and Discussion

As a non-convex optimization method, a common doubt may be whether MSRC could find the global optimal solution. However, theoretical proof is almost impossible. In this section, we discuss the convergence of MSRC via experimental analysis.

Fig. 5 shows the influence of iterations on the concentration of solutions. To make the evaluation result visible, we randomly select a test pixel with label 1 from SalinasA. The horizontal axis denotes individuals in Pareto solution set, and the ordinate axis denotes atoms in the dictionary. To be specific, the horizontal axis has 101 bins, each of which corresponds to an individual in Pareto solution set; the ordinate axis has 30 bins, because the dictionary in SalinasA contains 30 atoms (6 classes with 5 pixels in each). Class 1-6 are arranged from top to bottom. A white point denotes that this atom is selected by an individual. Correspondingly, a horizontal white line denotes that this atom is selected by all the individuals. In a nutshell, each column in Fig. 5(a)-(d) corresponds to a single solution. We can find that in Fig. 5(a)(b) the columns are not consistent, but after 100 iterations the columns in Fig 5(c)(d) tend to become uniform. Take Fig. 5(c) for example, all the 101 individuals have selected the same atoms (the white lines). Owing to our improvement on the decision-making process in MSRC, the Pareto solution set tends to



(a)

(b)

(c)

(d)

Fig. 6: Influence of iterations on residuals. (a), (b) and (c) are the average results on SalinasA, KSC and Botswana, respectively. (d) is an example for a randomly selected test pixel from SalinasA data.

concentrate on a single solution, *i.e.*, the white points tend to compose lines. Another reason for the concentration is that the range of Eq. 8 is discrete which leads to very few solutions in the Pareto front. In Fig. 5(a) and Fig. 5(b), because the iterations are not enough, we can see that the solutions are scattered and even wrong. After 100 iterations the Pareto set converges to a single solution, as shown in Fig. 5(c)(d). More importantly, several atoms in $\mathbf{A}_c$ are selected, which is very beneficial to the final classification result. Although the sparsity k is different in Fig. 5(c)(d), they both find the atoms from $\mathbf{A}_c$. In this case, they tend to present the same predicted label.

Fig. 6 illustrates the influence of iterations on residuals. Curves in Fig. 6(a)-(c) are the average results on the whole data sets, where we find that the residuals continue declining with the increase of iterations. In Fig. 6(d), we randomly select a test pixel for example. The curve is step-down, *i.e.*, the residual keeps unchanged during some iterations while plummets for some iterations. The experimental result in Fig. 6(d) is consistent with the theoretical analysis. Because MSRC is an L0 based method, the residuals are reduced only if the selected atoms are changed. However, due to the random flipping strategy in MSRC, the optimal solution is not updated in every iteration. Therefore, several flat lines can be observed in the iteration process. The iteration stops when a long line appears. According to Fig. 6 only 50 iterations are required.

As a method with no physical explanation, one of the disadvantages of MSRC is that it cannot handle the problems of “same materials with different spectra” and “different materials with same spectra”. These problems may get a solution by integrating the physical meaning of materials in the hyperspectral images.

IV. Conclusion

HSI classification task includes two crucial aspects: extracting more representative features and developing more powerful classifiers. In this paper, we focus on the latter, and propose a multi-objective based sparse representation classifier. MSRC is developed based on SRC, and especially targets two problems: limited samples and mixed pixels. The novelty of our work consists in the objectives as well as the optimization process, and can be summarized by two points: 1) Instead of using L1-norm sparse regularization, in MSRC the L0 problem is directly calculated without any relaxation; 2) An improved multi-objective based method is proposed to solve the non-convex problem, which is able to concentrate on a unique solution. The predicted labels are determined by abundance inversion. In the experiments, MSRC is compared with several popular classifiers on three public HSI data sets. The results prove that MSRC presents better performance when the training samples are limited.

As a non-convex optimization method, the efficiency of MSRC should be further improved. In our following work, we will focus on reducing the computational cost of the multi-objective based methods.

V. Acknowledgment

The authors would like to thank Prof. Wei Li and Prof. Qian Du for sharing the code of L1 based SRC for comparison. The authors would also like to thank the associate editor and four anonymous reviewers for the very insightful comments and instructions which have significantly improved the quality of this work.

References

[1] R. Duca and F. Del Frate, “Hyperspectral and multiangle chris-proba images for the generation of land cover maps,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 10, pp. 2857–2866, 2008.

[2] C. McCann, K. S. Repasky, M. Morin, R. L. Lawrence, and S. Powell, “Using landsat surface reflectance data as a reference target for multiswath hyperspectral data collected over mixed agricultural rangeland areas,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5002–5014, 2017.

[3] B. Pan, Z. Shi, and X. Xu, “Mugnet: Deep learning for hyperspectral image classification using limited samples,” *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271617303416

[4] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, “Hyperspectral remote sensing data analysis and future challenges,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 1, no. 2, pp. 6–36, 2013.

[5] Q. Wang, J. Lin, and Y. Yuan, “Salient band selection for hyperspectral image classification via manifold ranking,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, pp. 1279–1289, 2016.

[6] L. He, J. Li, C. Liu, and S. Li, “Recent advances on spectral–spatial hyperspectral image classification: An overview and new guidelines,” *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

[7] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, “Advanced spectral classifiers for hyperspectral images: A review,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8–32, 2017.

[8] A. Villa, J. Chanussot, J. A. Benediktsson, and C. Jutten, “Spectral unmixing for the classification of hyperspectral images at a finer spatial resolution,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 521–533, 2011.

[9] Q. Wang, Z. Meng, and X. Li, “Locality adaptive discriminant analysis for spectral–spatial classification of hyperspectral images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2077–2081, 2017.

[10] Z. Zhong, B. Fan, K. Ding, H. Li, S. Xiang, and C. Pan, “Efficient multiple feature fusion with hashing for hyperspectral imagery classification: A comparative study,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4461–4478, 2016.

[11] J. Xia, L. Bombrun, T. Adali, Y. Berthoumieu, and C. Germain, "Spectral–spatial classification of hyperspectral images using ica and edge-preserving filter via an ensemble strategy," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4971–4982, 2016.

[12] L. Zhang, X. Zhu, L. Zhang, and B. Du, "Multidomain subspace classification for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6138–6150, 2016.

[13] B. Pan, Z. Shi, and X. Xu, "Hierarchical guidance filtering-based ensemble classification for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4177–4189, 2017.

[14] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2014.

[15] B. Pan, Z. Shi, N. Zhang, and S. Xie, "Hyperspectral image classification based on nonlinear spectral–spatial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1782–1786, 2016.

[16] B. Pan, Z. Shi, and X. Xu, "R-VCANet: A new deep-learning-based hyperspectral image classification method," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 5, pp. 1975–1986, 2017.

[17] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 391–406, Jan 2018.

[18] X. Kang, C. Li, S. Li, and H. Lin, "Classification of hyperspectral images by gabor filtering based deep network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 4, pp. 1166–1178, April 2018.

[19] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[20] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2017.

[21] Y. Chen, N. M. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3973–3985, 2011.

[22] W. Li and Q. Du, "A survey on representation-based classification and detection in hyperspectral remote sensing imagery ," *Pattern Recognition Letters*, vol. 83, pp. 115–123, 2016.

[23] D. B. Heras, F. Argello, and P. Quesadabarriuso, "Exploring ELM-based spatial–spectral classification of hyperspectral images," *International Journal of Remote Sensing*, vol. 35, no. 2, pp. 401–423, 2014.

[24] Y. Gu, C. Wang, D. You, Y. Zhang, S. Wang, and Y. Zhang, "Representative multiple kernel learning for classification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 7, pp. 2852–2865, 2012.

[25] Y. Gu and H. Liu, "Sample-screening MKL method via boosting strategy for hyperspectral image classification," *Neurocomputing*, vol. 173, pp. 1630–1639, 2016.

[26] W. Li and Q. Du, "Adaptive sparse representation for hyperspectral image classification," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2015, pp. 4955–4958.

[27] H. Yu, L. Gao, W. Li, Q. Du, and B. Zhang, "Locality sensitive discriminant analysis for group sparse representation-based hyperspectral imagery classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 8, pp. 1358–1362, 2017.

[28] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *Siam Review*, vol. 43, no. 1, pp. 129–159, 2001.

[29] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[30] Q. Wang and X. Li, "Shrink image by feature matrix decomposition," *Neurocomputing*, vol. 140, pp. 162–171, 2014.

[31] W. Sun, G. Yang, B. Du, L. Zhang, and L. Zhang, "A sparse and low-rank near-isometric linear embedding method for feature extraction in hyperspectral imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4032 – 4046, 2017.

[32] X. Xu and Z. Shi, "Multi-objective based spectral unmixing for hyperspectral images ," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 124, pp. 54–69, 2017.

[33] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 4, pp. 577–601, 2014.

[34] H. Jain and K. Deb, "An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach,

part ii: Handling constraints and extending to an adaptive approach," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 4, pp. 602–622, 2014.

[35] X. Xu, Z. Shi, and B. Pan, "A new unsupervised hyperspectral band selection method based on multio-bjective optimization," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2112–2116, 2017.

[36] F. A. Mianji and Y. Zhang, "SVM-based unmixing-to-classification conversion for hyperspectral abundance quantification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 11, pp. 4318–4327, 2011.

[37] I. Dopido, A. Villa, A. Plaza, and P. Gamba, "A quantitative and comparative assessment of unmixing-based feature extraction techniques for hyperspectral image classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 421–435, 2012.

[38] M. Gong, H. Li, E. Luo, J. Liu, and J. Liu, "A multiobjective cooperative coevolutionary algorithm for hyperspectral sparse unmixing," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 2, pp. 234–248, April 2017.

[39] F. van den Bergh and A. P. Engelbrecht, "A cooperative approach to particle swarm optimization," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 225–239, June 2004.

[40] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[41] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 6, pp. 712–731, Dec 2007.

[42] C. Hillermeier, *Nonlinear multiobjective optimization*. Birkhauser Verlag, 2001.

[43] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[44] D. C. Heinz and C. Chang, "Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 3, pp. 529–545, 2001.

[45] M. D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 6, pp. 2014–2039, 2011.

[46] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[47] L. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *IEEE International Conference on Computer Vision*, 2012, pp. 471–478.

[48] W. Li, Q. Du, and M. Xiong, "Kernel collaborative representation with tikhonov regularization for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 1, pp. 48–52, 2015.

[49] A. Papoulis and S. Pillai, *Probability, Random Variables and Stochastic Processes (4nd Edition)*. McGraw-Hill, 2012.