A novel semisupervised SVM classifier based on active learning and context information

Fei Gao · Wenchao Lv · Yaotian Zhang · Jinping Sun · Jun Wang · Erfu Yang

Abstract This paper proposes a novel semisupervised support vector machine classifier (S^3VM) based on active learning (AL) and context information to solve the problem where the number of labeled samples is insufficient. Firstly, a new semisupervised learning (SSL) method is designed using AL to select unlabeled samples as the semilabled samples, then the context information is exploited to further expand the selected samples and relabel them, along with the labeled samples train S^3VM classifier. Next, a new query function is designed to enhance the reliability of the classification results by using the Euclidean distance between the samples. Finally, in order to enhance the robustness of the proposed algorithm, a fusion method is designed. Several experiments on change detection are performed by considering some real remote sensing images. The results show that the proposed algorithm in comparison with other algorithms can significantly improve the detection accuracy and achieve a fast convergence in addition to verify the effectiveness of the fusion method developed in this paper.

Keywords semisupervised support vector machine · active learning · context information · remote sensing

This work was supported by the National Natural Science Foundation of China (61071139; 61471019; 61171122; 61501011), the Aeronautical Science Foundation of China (20142051022), the Pre-research Project(9140A07040515HK01009), the National Natural Science Foundation of China (NNSFC) under the RSE-NNSFC Joint Project (2012-2014) (61211130210) with Beihang University, and the RSE-NNSFC Joint Project (2012-2014) (61211130309) with Anhui University.

Fei Gao • Wenchao Lv • Yaotian Zhang • Jinping Sun • Jun Wang
School of Electronic and Information Engineering, Beihang University, Beijing, China
Fei Gao
e-mail: feigao2000@163.com
Wenchao Lv
e-mail: 13161111047@163.com
Yaotian Zhang(✉)
e-mail: zhangyaotian@buaa.edu.cn
Jinping Sun
e-mail: sunjinping@buaa.edu.cn
Jun Wang
E-mail: wangj203@buaa.edu.cn
Erfu Yang
Space Mechatronic Systems Technology Laboratory, University of Strathclyde, Glasgow G1 1XJ, UK
e-mail: erfu.yang@strath.ac.uk

1 Introduction

With the rapid development of remote sensing (RS) technology, RS data has been expanded constantly and used widely in geologic monitoring, environmental protection and disaster relief, where the distinction of the land-cover change(Hansen and Loveland 2012), classification of the land-use(Gong et al. 2015) and assessment of the earthquake influence(Geiß and Taubenböck 2013)can be attributed to the problem of object detection under certain conditions. In these fields, it is a very difficult task to detect useful objects from a large amount of RS images (RSIs), which presents a high requirement to the information extraction ability of the correlated algorithms and increasingly attracts the research interest.

The traditional technologies for object detection are mainly based on unsupervised learning and supervised learning. Unsupervised method detects the object from its background via grouping the samples into different clusters and does not require prior knowledge, which reduces the time and cost for human-labeled training samples. However, such methods only consider the characteristic information between the different clusters and lack the effective guidance of the labeled samples, which makes it impossible to extract objects of interest(Anand et al. 2014). Therefore, the supervision method based on the labeled samples is adopted to change the detection problem into classification problem and quickly captures the effective RS information with the high accuracy. However, to train the classifier, the supervised method requires enough labeled samples(Ujjwal Maulik and Chakraborty 2012), which are scarce in practice and easily affected by noises. Moreover, it is very difficult to obtain the labeled samples manually because of the limit of the time and cost (Li et al. 2010). In contrast, the unlabeled samples are abundant and contain a wealth of information. By using them, not only the problem of insufficient training samples can be solved, but also the working efficiency is improved(Shahshahani and Landgrebe 1994). Therefore, in recent years the algorithms based on the semisupervised learning (SSL) have been concerned greatly.

The SSL based method iteratively selects the unlabeled samples with the labeled samples together to train the classifier(Kawakita and Kanamori 2013), and can reduce human intervention and classify the unlabeled samples more effectively. In general, the SSL can be grouped into the following five categories, i.e.: self-training, cooperative training, generative probability model, semisupervised support vector machine(S^{3}VM) and graph based methods(Chapelle et al. 2006; Zhu 2010), where the key to their effective functioning is the way for screening the unlabeled samples. The progressive S^{3}VM with diversity (PS^{3}VM-D) takes γ samples which are within and closer to the margin band, to define the candidate semilabeled samples, then incrementally selects ρ diverse samples among the γ candidates by considering the kernel cosine-angular similarity(Persello and Bruzzone 2014). The context-sensitive S^{3}VM (CS^{4}VM) selects the context patterns of training samples as the semilabeled samples, then weights them depending on their consistency with the center sample by SVM(Bruzzone and Persello 2009).Junwei et al. (2015a)built a high-level object feature representation using a deep Boltzmann machine(DBM) and applies the proposed Bayesian principle to characterize the objects information, then classifies the objects by the liner SVM to select the semilabeled samples.

The performance of the SSL algorithms can still be improved in the following aspects. First, the SSL based methods can be deteriorated when exploiting huge amount of the unlabeled samples, which would seriously hinder the adjustment of the relevant parameters(Munoz-Mari et al. 2012; Gomez-Chova et al. 2008), and the effective strategy for the selection of the semilabeled samples is necessary. The SSL based methods can be greatly affected by initial training set(Didaci et al. 2012), where the validity of samples can't be guaranteed sometimes, leading to random classification result.

For the selection strategy of the semilabeled samples, Pasolli et al. (2014) applied the active learning (AL) and the spatial information, including Parzen window, spatial entropy, etc., to process the unlabeled samples, where the semilabeled samples were selected by estimating the probability density distribution function of the random variables and the entropy of the discrete random variables. Demir et al. (2011) applied the AL based on kernel clustering to deal with the unlabeled samples, then designed a new strategy to select the most representative semilabeled samples. Inspired by the above algorithms, the selection strategy was designed utilizing AL in the proposed SSL based algorithm. The AL iteratively extracts the most informative unlabeled samples to be labeled manually and can obtain a higher classification accuracy than other SSL algorithms, reducing the dependence on the labeled samples (Tuia et al. 2011; Persello and Bruzzone 2014).However, the existing feature descriptors are insufficiently powerful to characterize the information of objects, which inevitably leads samples mislabeling and seriously reduces the classification accuracy. To fully describe the sample characteristics, Junwei et al. (2015b)proposed a novel framework via the deep learning methods for the salient object detection, which extracts four neighborhood image boundaries of the object as a reference, then applied the stacked de-noising auto-encoder(SDAE) with the deep learning architectures to model image background. The robustness of the mislabeled patterns is improved by exploiting spatial information. Inspired by the above method, the structure of the AL is modified by using the context information, which can expand the amount of semilabeled samples and improve the accuracy to a certain extent when labeling them.

The reason why the S^{3}VM is chosen to process the selected samples is detailed in the following. The support vectors of S^{3}VM make the decision rather than many redundant samples to improve the operational efficiency and the robustness to the polluted samples. The nonlinear mapping based S^{3}VM can map the samples into a high dimensional characteristic space to make them linearly separable. The functional margin that is the distance between the sample and the optimal hyper plane decided by the S^{3}VM, can be used as the confidence-based decision marking to make the S^{3}VM be effectively exploited by the SSL based methods(Hearst et al. 1998; U. Maulik and Chakraborty 2014).

For the uncertainty of the initial training set, the fusion is an effective method. The downsampling is commonly used to obtain different data models by generating the images at different resolutions(Bazi et al. 2010). However, the progressively downsampling is not suitable to deal with the image with large difference in areas, because it can lose them with scanty pixels. The initial sets in the SSL does not require many labeled samples, so they can be directly selected from the image to reduce the complexity of the preprocessing. Their SSL results as different models were fused to decrease the influence of the random distribution of the initial set and obtain stable classification result.

In this paper, a novel S^{3}VM algorithm is proposed to address the classification problems of the RSIs when the available labeled samples are not sufficient. To guarantee the decision-making ability of the proposed method, the AL and context information are applied to select the informative unlabeled samples (semilabeled samples), which are combined with the labeled samples together to train the S^{3}VM. The novelty of this paper lies in the following aspects. First, a new query function is designed based on the Euclidean distance among samples. Second, the amount of the semilabeled samples are increased by exploiting their contextual patterns. Last, the selected samples are relabeled by exploiting the correlation between the sample and context. Hence, the comprehensive utilization of the context information ensures the reliability of the selected samples and increases the classification credibility of the S^{3}VM.

The rest of this paper is organized as follows. Section 2 describes the proposed SSL based method in detail, and the necessity of the fusion is also analyzed to design new fusion rules. Section 3 presents the three data sets and experiments for the RSIs change detection. Section 4 draws the conclusion of this paper.

2 Proposed method

The algorithm consists of two parts: the proposed S^{3}VM based on AL and context information, and the fusion method.

2.1 Proposed SSL-based method

Assuming that the training set **L**is composed of *n* pairs (x_i,y_i) available, where x_i denotes the training samples and $y_i \in \{+1, -1\}$denotes the label for the binary classifier. We assume that a learning set **U** with *m* unlabeled samples (x'_j). They are defined as

$$\mathbf{L} = \left\{(x_i, y_i) \mid i = 1......n\right\} \tag{1}$$

$$\mathbf{U} = \{x'_j \mid j = 1......m\} \tag{2}$$

In order to expand **L**, the semilabeled samples are selected from **U** iteratively. Fig. 3 shows the flowchart of the SSL method based on the S^{3}VM.

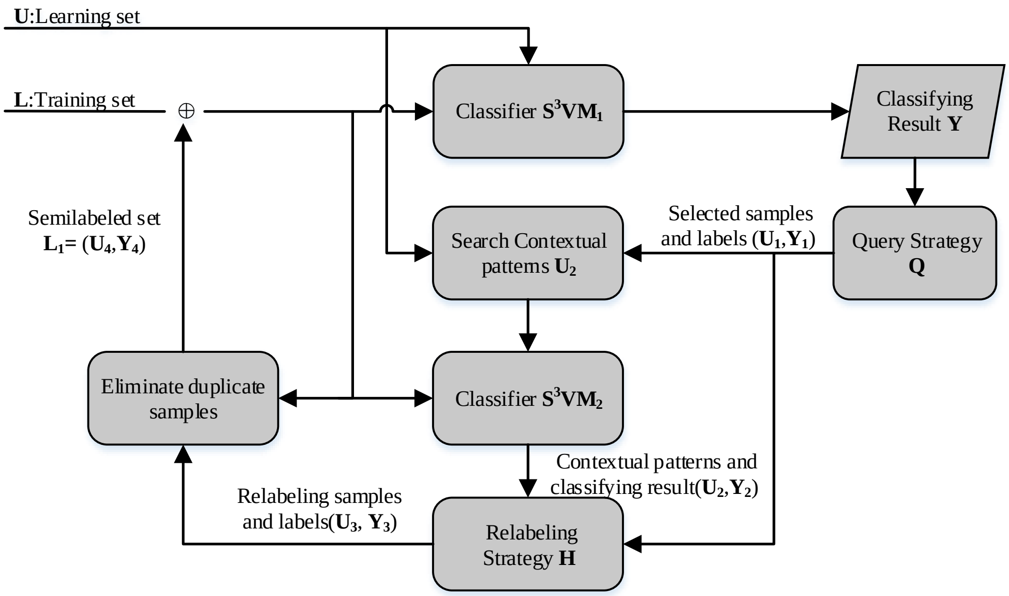

U:Learning set
L:Training set
⊕
Classifier S3VM1
Classifying Result Y
Semilabeled set L1= (U4,Y4)
Selected samples and labels (U1,Y1)
Search Contextual patterns U2
Query Strategy Q
Eliminate duplicate samples
Classifier S3VM2
Contextual patterns and classifying result(U2,Y2)
Relabeling samples and labels(U3, Y3)
Relabeling Strategy H

Fig. 1 Flowchart of the proposed method

First, **L** is processed by the S^{3}VM$_1$,during which the label **Y** of the **U** is obtained. The subset ($\mathbf{U}_1$, $\mathbf{Y}_1$) is selected by the query function Q. Then, the context patterns $\mathbf{U}_2$of the $\mathbf{U}_1$ is found from the **U** and classified by the S^{3}VM$_2$.$\mathbf{U}_2$along with the central pattern ($\mathbf{U}_1$,$\mathbf{Y}_1$) are relabeled by H to get the semilabeled set ($\mathbf{U}_3$,$\mathbf{Y}_3$). Last, the repeated samples between **L**and ($\mathbf{U}_3$,$\mathbf{Y}_3$) are eliminated to obtain the ultimate semilabeled set $\mathbf{L}_1$= ($\mathbf{U}_4$, $\mathbf{Y}_4$)to be added into **L**. The entire process is iterated until the predefined convergence condition is satisfied, e.g., the total number of the semilabeled samples or the classification accuracy is reached (Pasolli et al. 2014).

Next, the implementation of the S^{3}VM, the query function Q, the search contextual patterns and relabeling strategy H are described.

2.2S^{3}VM method

The S^{3}VM is the expansion of the SVM. A standard SVM is based on the structural risk minimization to classify the learning set by extracting the support vectors from the training set to find the optimal hyper plane(Scholkopf et al. 1997). In case of the binary SVM, given the training set **L**and the testing set **U**, it is limited to the following constrained optimization problem (Izquierdo-Verdiguier et al. 2013):

$$\begin{aligned} &\min \Psi(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w}) + C\sum_{i=1}^{n}\xi_i \\ &\text{Subject to :}\quad y_i[\mathbf{w}^T\Phi(\boldsymbol{x}_i)+b]\ge 1-\xi_i \quad i=1,\ldots,n \\ &\qquad\qquad \xi_i \ge 0 \end{aligned} \tag{3}$$

Where x_i is the training sample and y_i is the corresponding label,(x_i,y_i)∈ **L**;Φ(·) maps the data into the feature space; $\boldsymbol{w}$ is the orthogonal vector between x_i and the hyper plane; b is the bias to measure the distance between **L** and the hyper plane; ξ_i is the slack variable to represent offset of x_i; $C$is the cost factor to measure the weight between the optimal hyper plane and the minimum offset; n is the number of the training samples.

After the initialization, the iterative process is operated and the semilabeled samples (selected from **U** in the previous step) are added to **L**. Their confidence is diverse in different iterative steps and they are given different cost factors(Bovolo et al. 2008), which leads to the following cost function for the classifier learning :

$$\begin{aligned} &\min \Psi(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w}^T\boldsymbol{w}) + C\sum_{i=1}^{n}\xi_i + \sum_{j=1}^{m}C_j\square\varepsilon_j \\ &\text{Subject to :}\quad y_i[\mathbf{w}^T\Phi(x_i)+b]\ge 1-\xi_i \quad i=1,\ldots,n \\ &\qquad\qquad \hat{y}_j[\mathbf{w}^T\Phi(\hat{x}_j)+b]\ge 1-\varepsilon_j \quad j=1,\ldots,n' \\ &\qquad\qquad \xi_i \ge 0, \varepsilon_j \ge 0 \end{aligned} \tag{4}$$

where$\hat{x}_j$is the semilabled sample selected from **U**, with the slack variable(ε_j),cost factor (C_j) and semilabel($\hat{y}_j$), and $n'$is the number of the semilabeled samples.

Applying the Lagrange Duality, the equation (4) can be transformed into the dual problem, which can be solved by applying the Karush–Kuhn–Tucker(KKT)(Yi et al. 2011)optimality conditions:

$$\max_{\alpha,\beta}\left(\sum_{i=1}^{n}\alpha_i+\sum_{j=1}^{m}\beta_j-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\frac{1}{2}\sum_{i,j=1}^{m}\beta_i\beta_j\hat{y}_i\hat{y}_j K(\hat{x}_i,\hat{x}_j)-\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j y_i\hat{y}_j K(x_i,\hat{x}_j)\right)$$

$$\text{Subject to: } \sum_{i=1}^{n}\alpha_i y_i+\sum_{j=1}^{m}\beta_j\hat{y}_j=0 \qquad (5)$$

$$0\le\alpha_i\le C \qquad i=1,\dots,n$$

$$0\le\beta_j\le C_j \qquad j=1,\dots,m$$

where K=(x_i,$\hat{x}_{j\cancel{j}}$)is a kernel function to calculates the inner product<Φ(x_i)·Φ($\hat{x}_j$)>;α_i and β_j are the Lagrange multipliers corresponding to the labeled sample (x_i) and the semilabeled sample ($\hat{x}_j$) respectively. When α_i and β_j are solved by(5), b is solved by

$$f(x)=\boldsymbol{w}^T\bullet\phi(x)+b=\left(\sum_{i=1}^{n}a_i y_i K(x,x_i)+\sum_{j=1}^{m}\beta_j\hat{y}_j K(x,\hat{x}_j)\right)+b \qquad (6)$$

Finally, the unlabeled sample x'(x'∈**U**) can be labeled by the following decision function:

$$y'=\operatorname{sgn}\{(\boldsymbol{w}^T\bullet x')+b\}=\operatorname{sgn}\left\{\left(\sum_{i=1}^{n}a_i y_i K(x',x_i)+\sum_{j=1}^{m}\beta_j\hat{y}_j K(x',\hat{x}_j)\right)+b\right\} \qquad (7)$$

It is worth noting that with the increase of iteration, the previous semilabeled samples are more convincing and the corresponding C_j would get bigger. The S^{3}VM can deal with the nonlinear problem using many different kernel functions, such as Gauss kernel function, radial kernel function or exponential kernel function.

The basic structure of the S^{3}VM is described above. The learning set **U** and the result **Y** classified by the S^{3}VM$_1$ are the inputs to Q that is described next.

2.3Query Strategy Q

Fig. 2 shows the structure of the query strategy Q consisting of two parts Q$_1$ and Q$_2$. Q$_1$is used to screen (**U**′, **Y**′) from (**U**, **Y**).Then, (**U**′, **Y**′) is refined by Q$_2$using the proposed rules f_1 and f_2, and the output is(**U**$_1$, **Y**$_1$).

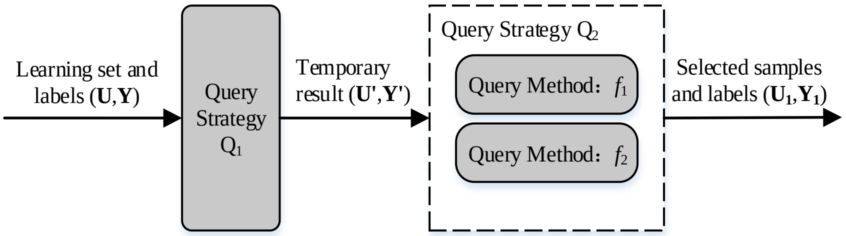


Fig. 2 Structure of Q

2.3.1 The implementation of Q$_1$ and Q$_2$

Q$_1$is designed on the basis of the MS. The MS is conducive to the convergence of the algorithm because the samples within the margin band are more informative than others. Therefore, **U**′is selected within the margin band, where the absolute value of the decision function is limited between the ρ and 1, defined as follows:

$$\mathbf{U}' = \left\{ x_i \mid x_i \in \mathbf{U}, \rho < abs(value_i) < 1, i = 1, \ldots, m \right\}$$
$$\text{Subject to}: \quad value_i = f(x_i) = \mathbf{w}^T \Box x_i + b \tag{8}$$

Q$_2$is designed by using the Euclidean distance between the samples and the screening samples from the **U**′by applying the rules *f*$_1$and *f*$_2$.The implementation is shown as follows:

I. The distance (**D**) between the sample of **U**′and the training set **L** is calculated by

$$\mathbf{D} = \{ d_{ij} \mid d_{ij} = \mathrm{norm}(x_i, x_j), x_i \in \mathbf{U'}, x_j \in \mathbf{L} \} \tag{9}$$

II. The minimum, maximum and their ratios between **U**′and **L** are calculated by

$$\mathbf{D}_{\min} = \left\{ d_i \mid d_i = \min{}_j(d_{ij}), d_{ij} \in \mathbf{D} \right\} \tag{10}$$

$$\mathbf{D}_{\max} = \left\{ d_i \mid d_i = \max{}_j(d_{ij}), d_{ij} \in \mathbf{D} \right\} \tag{11}$$

$$\Delta = \left\{ \delta_i \mid \delta_i = {}^{\mathbf{D}_{\mathbf{min}}(i)}\!\big/_{\mathbf{D}_{\mathbf{max}}(i)} \right\} \tag{12}$$

III. The rule*f*$_1$ is used to select x_1 from **U**′:

$$f_1: \quad x_1 = \mathbf{U}'(i)$$
$$\text{Subjected to:} \quad \mathbf{D}_{\mathbf{max}}(i) == \max(\mathbf{D}_{\mathbf{max}}), \quad i = 1, \ldots, k \tag{13}$$

IV. The rule*f*$_2$ is used toselect$x_2$from**U**′:

$$f_2: \quad x_2 = \mathbf{U}'(i)$$
$$\text{Subjected to:} \quad \Delta(i) = \max(\Delta), \quad i = 1, \ldots, k \tag{14}$$

V. Finally, update **U$_1$**:

$$\mathbf{U_1} = \mathbf{U_1} + x_1 + x_2 \tag{15}$$

Through the loop running I-V, **U$_1$** is updated continuously until the number of samples in **U$_1$** reaches the threshold.

The implementation of Q has been introduced. Next, the significance of*f*$_1$ and *f*$_2$is described to prove the rationality of Q.

2.3.2 Significance of *f*$_1$ and*f*$_2$

In order to facilitate the description, Fig.3 illustrates the distribution of the sample *x*($x \in \mathbf{U}'$) and the training set **L**.

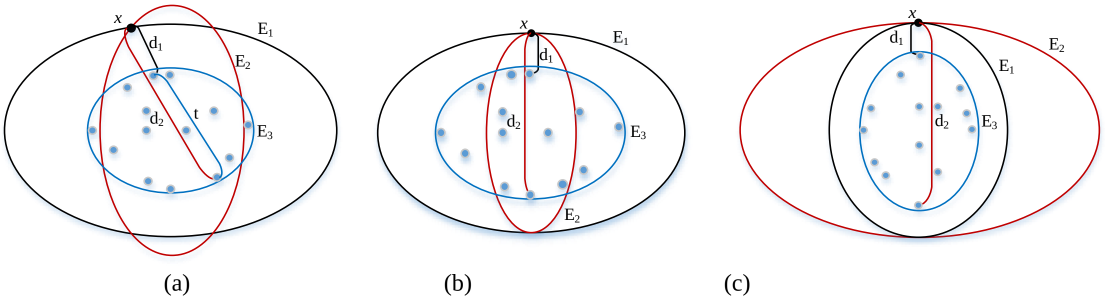


Fig. 3 Distribution of the sample *x* and training set **L**. (a) *x* distributed around E$_3$randomly. (b)*x* distributed in the direction of the short axis of E$_3$. (c)*x* distributed in the direction of the long axis of E$_3$.

For all the figures in Fig.3, **L** is shown as the blue dots and E$_3$is the boundary; *x* is shown as the black dots, $d_1 \in \mathbf{D_{min}}$ and $d_2 \in \mathbf{D_{max}}$(shown in the equation(10) and (11)); the points in E$_1$ have the same minimum distance with d_1; the points in E$_2$ have the same maximum distance with d_2.

For thef_1, Fig. 3(a) shows that the bigger $d_2$generates a longer distance between x and **L**, and a smaller probability of x belonging to **L**, so more information is contained in $x$which can decrease the amount of the semilabeled samples used by the SSL. Therefore, $f_1$canselectthe informative sample and reduce the semilabeled samples to the greatest extent.

For the f_2,Fig. 3(a) shows that the δ ($\delta \in \mathbf{\Delta}$,in (12))in x can be written as

$$\delta = d_1/d_2 \qquad (16)$$

In order to facilitate the description of f$_2$, the expression of δ needs to be changed.

Let: $$d_1 = d_2 - t \qquad (17)$$

Where t is the secant of x in E$_3$, then substitute (17) into (16):

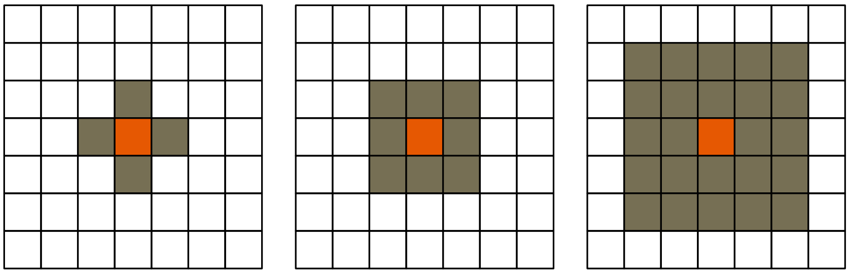
$$\delta = (d_2 - t)/d_2 \qquad (18)$$

Get: $$\delta = 1 - t/d_2 \qquad (19)$$

where it can be seen that: when d_2 is invariant, the bigger δ is, the smaller t is, which means that x would be closer to the short axis of E$_3$(shown in Fig. 3(b)) and can make distribution of **L** more uniform (compared with Fig. 3(c));when t is constant, the bigger δ is, the smaller$d_2$is, and the longer distance between x and **L**is, which means that the smaller of the probability of x belonging to the **L**, so the more information that x contains. Therefore, $f_2$cannot only reduce the semilabeled samples by selecting informative ones, but also make the distribution of **L** better.

2.4Search contextual patterns

This function makes the current pixel x($x \in \mathbf{U}_1$) as the center sample to find the context patterns $\mathbf{U}_2$ from the neighborhood. There are many choices for$\mathbf{U}_2$. Espinola et al. (2015)introduced the most common types shown in Fig.4, where Von Neumann neighborhood and Moore neighborhood are called the first-order system and the second-order system(Bruzzone and Persello 2009), and Von Neumann neighborhood is also called the four-directly neighbored pixels (Zhao et al. 2013).



(a) (b) (c)

Fig. 4 Examples of neighborhood systems. (a) Von Neumann neighborhood, (b) Moore neighborhood and (c) Extended Moore neighborhood

Context information is very significant. Firstly, the center sample and the context patterns are close in the space, which makes their characteristics exist great relevance that can be applied to calculate the image texture feature(D'Elia et al. 2014)or generate the classification map(Bruzzone and Persello 2009). More semilabeled samples are found using context information, which can improve the convergence speed for the SSL. Secondly, their labels can also provide very important information. In(Espinola et al. 2015), the cell automaton distinguishes the different pixels (such as the boundary, noise, or common) at each iteration by utilizing their labels. The robustness of the SSL algorithm can be increased by properly using the context information of labels. The next section describes the way to relabel samples by using the information of labels.

2.5 H: Relabeling samples

Q selects the samples between the maximum margin bands, where the selected samples are more likely to be contaminated and mislabeled manually. To increase the reliability of the selected samples, the central samples ($\mathbf{U}_1$,$\mathbf{Y}_1$) and the context patterns ($\mathbf{U}_2$,$\mathbf{Y}_2$)are relabeled by the function H with the following rules:

I. When $\mathbf{Y}_2$are consistent with $\mathbf{Y}_1$, they are left;

II. When $\mathbf{Y}_2$are inconsistent with $\mathbf{Y}_1$, $\mathbf{Y}_1$ is incorrect. To make **L** representative, the corresponding central sample is discarded and the context patterns are left.

III. When a part of $\mathbf{Y}_2$ is consistent with$\mathbf{Y}_1$, for$\mathbf{Y}_2$, the inconsistent context patterns are removed; for $\mathbf{Y}_1$, the proportion of consistent part to context patterns is calculated: if the proportion is greater than the threshold, rules A would be followed, else rules B would be chosen.

The semilabeled set ($\mathbf{U}_3$,$\mathbf{Y}_3$) is obtained after being relabeled by H. Then, those samples repeated with **L** are eliminated, which generates the semilabeled set $\mathbf{L}_1$= ($\mathbf{U}_4$,$\mathbf{Y}_4$)to add to **L.**

Section 2.1-2.5 describes the proposed SSL based method. In order to overcome the influence of the randomness of the learning set **L**, the classification results are fused in the next section.

2.6Fusion method

Data fusion is a very popular data processing technology to make up for the defects caused by the missing data or noise pollution. It can be used in many methods. The principal component analysis (PCA) extracts the principal components to fuse the training set, and the spectral information and spatial information are combined together to fuse images. Our fusion method is shown in Fig.5,where the part in the dashed rectangular box is the SSL based method proposed in Section 2.1:

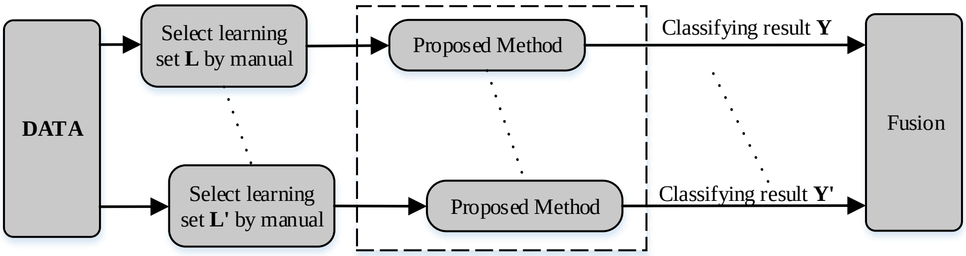

DATA
Select learning set **L** by manual
Select learning set **L'** by manual
Proposed Method
Proposed Method
Classifying result **Y**
Classifying result **Y'**
Fusion

Fig. 5 Fusion method

Firstly, because only a few labeled samples are needed to initialize the training set, the initial **L** (**L'** and etc.) are selected from the **DATA** manually and would not cost a lot. Then, **L** is processed by the proposed method to classify the learning set **U** for obtaining the classification result **Y**. The other initial sets are processed by the same way. Finally, the different classification results (**Y**,**Y'**and etc.) are treated as the different data models to be fused by the given different weights based on the detection accuracy.

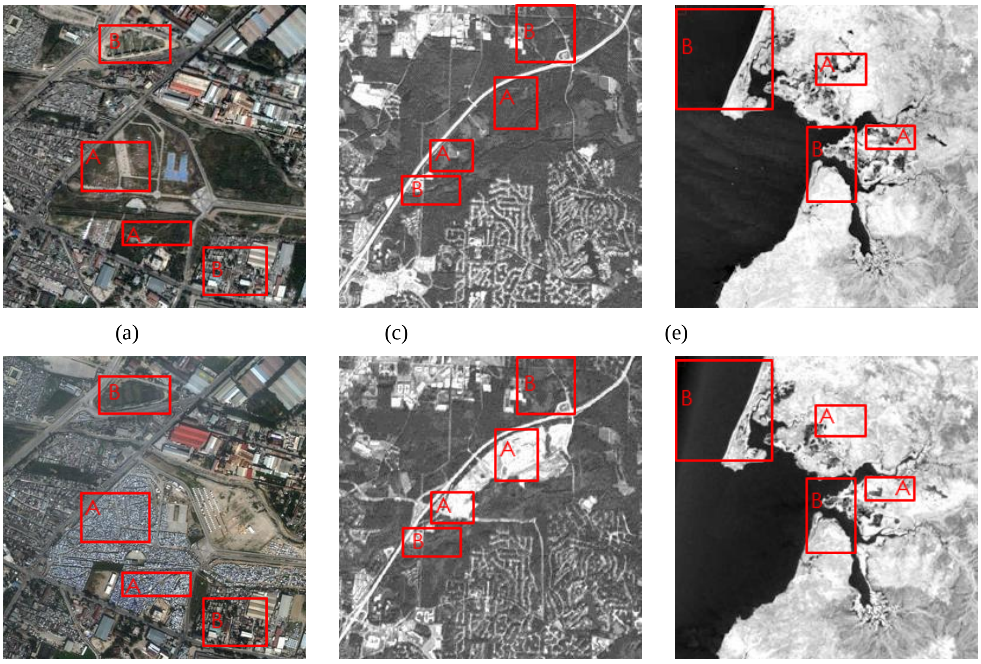
3 Experiments

In this section, the related experiments are designed to measure the performance of the proposed algorithm (Method Proposed, PM) in three aspects. Firstly, to compare PM with the supervised algorithm, the experiments on PM and SVM are carried out. Secondly, to compare PM with the semisupervised algorithm, the experiments on PM, CS^{4}VM and PS^{3}VM-D are carried out. Finally, to further improve the performance of PM, the different experimental results are fused.

3.1 Dataset description and parameters setting

3.1.1 Dataset Description

Our experiments on change detection are performed on real RSIs, as shown in Fig. 6, where the regions A and regions B indicate the changed regions and unchanged regions respectively. The aim of change detection is to distinguish the changed part and unchanged part of images (Habib et al. 2009; Celik 2010), and can be regarded as the binary classification problem. For the regions A: Fig. 6(a)-(b) are the city images taken before and after the earthquake and possess the rich texture features because of many artificial constructions. The change is due to the accumulation of people. Fig. 6(c)-(d)are the ground images taken at different periods and do not possess obvious texture features. The change is due to the appearance of many white patches. Fig. 6(e)-(f)are the coast images taken at different periods and also do not possess obvious texture features. The change is due to the appearance of a few small-scale gray patches. For the regions B of Fig.6, their geomorphic features are unchanged basically, only the gray changes a little.



(a) (c) (e)

(b) (d) (f)

Fig. 6 Remote sensing images ofthe city (a) before earthquake and (b) after earthquake, the ground (c) before change and (d) after change, and the coast (a) before change and (b) after change. A represents the changed areas, B represents the unchanged areas

After all the images are processed by the geometric registration and radiometric calibration, then the ratio map of grayscales is generated. To characterize the images, for Fig. 6(a)-(b) the gray level co-occurrence matrix is calculated as the **DATA₁**; for Fig. 6(c)-(d) and Fig. 6(e)-(f), the sample texture featuresare calculated as the **DATA₂** and **DATA₃** respectively. All the experimental data are normalized and divided into the training set and the learning set respectively.

3.1.2Parameters setting

The cost factor C of the PM is nonlinear, defined as follows:

$$C=\alpha \bullet (N\text{-}1)^2+C_1 \tag{20}$$

Where N is the number of iterations; C_1 is the initial cost factor; $\alpha$is the weight coefficient. α and $C_1$are set to 1, and the change range of C is [1,100]. With the increase of iterations, the reliability of the selected samples would increase, and the corresponding C would become larger than the previous one. For the PS^{3}VM-D, C is shown in (20), whereC_1 is set to 2 and$\alpha$is calculated by

$$\alpha=(C_{\max}\text{-}C_{\min})/(r-1)^2 \tag{21}$$

Where $C_{\max}$ and $C_{\min}$ are the maximum and minimum of C; r is designed artificially, with r = 10 herein. For the CS^{4}VM, C is defined as:

$$C=2\bullet(N-1)+C_1 \tag{22}$$

where N is also the number of iterations; The initial value of C_1 is 2. C is weighted by the K

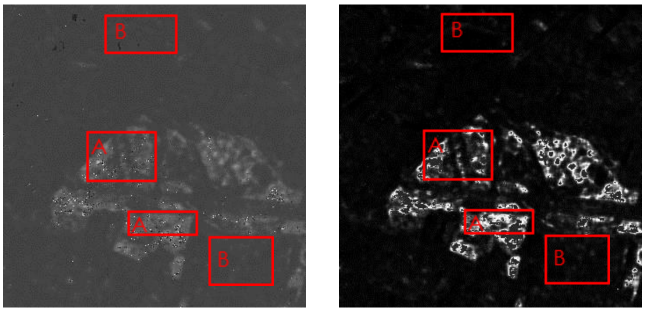
$$K=C/C'=2 \tag{23}$$

When the label of the contextual pattern is the same as the center sample, C is exploited, otherwise C' is explored.

All the methods are performed using the Gauss kernel function, with the radial width set 0.6. The Von Neumann neighborhood is applied by the PM and CS^{4}VM.

3.2 Comparison with SVM

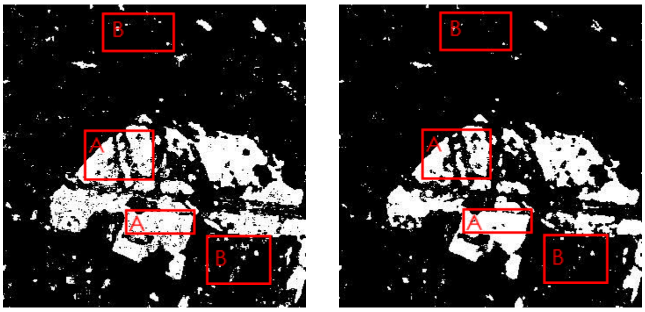
Considering the **DATA₁**, the performance of the SVM and PM is compared. Fig. 7 shows the detection grayscales of the two algorithms.



(a) (b)

Fig. 7 Detection grayscales achieved on **DATA₁**in terms of (a) SVM and (b) PM.A represents the changed areas, B represents the unchanged areas

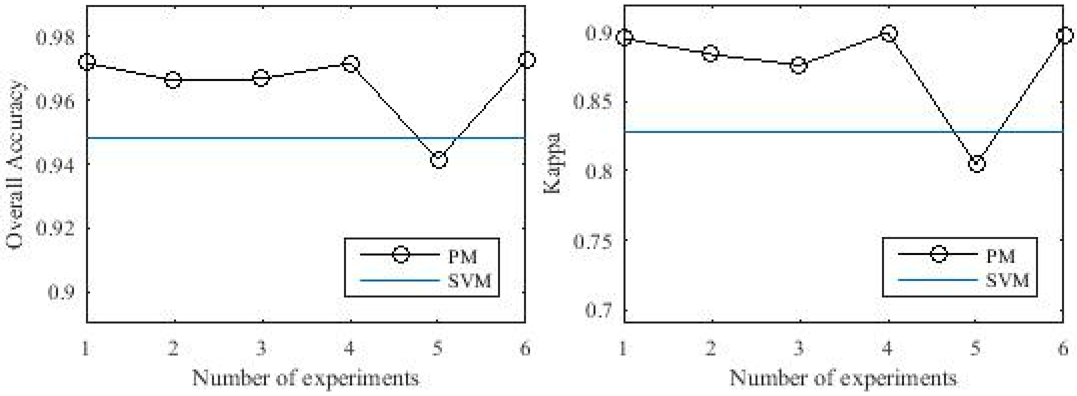
Through the threshold processing, the detection maps are obtained, as shown in Fig. 8 where the white area is detected with the changes and the black area is detected as unchanged; A represents the changed areas and the B represents the unchanged areas.



(a) (b)

Fig. 8 Detection maps achieved on the **DATA$_1$** in terms of (a) SVM and (b) PM, where the white area is detected with the changes and the black area is detected as unchanged; A represents the changed areas and the B represents the unchanged areas

There is much omission in A and false alarm in B in Fig. 8(a), while the detection error is much less in Fig. 8(b). In order to compare the performance of the PM in detail, 6 experiments of PM were implemented to obtain the results in terms of the overall accuracy (OA) and Kappa, as shown in Fig. 9.



(a) (b)

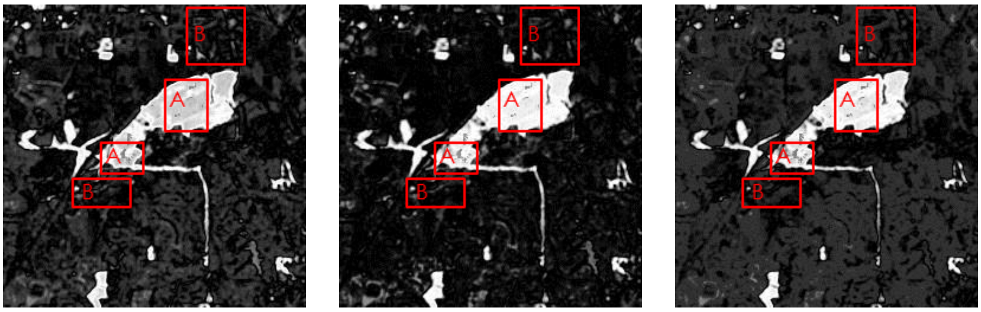
Fig. 9 Results achieved on **DATA$_1$** in terms of (a) OA and (b) Kappa

From Fig.9, except the results of the fifth experiment in which the PM are slightly lower than the SVM, others are significantly higher than the SVM in OA and Kappa. Therefore, in comparison with the SVM, the detection accuracy and performance of the PM have been improved greatly.

The training set of the SVM is obtained by the threshold segmentation. The threshold must be large enough to ensure the accuracy of the training set, which means that the samples with a high diversity are probably not within the threshold range. Moreover, the SVM does not exploit the contextual information. Although only a few initial training samples were used, the PM effectively absorbs the abundant and informative unlabeled samples to add to the training set by the AL, which makes it outperform the detection performance compared with the SVM.

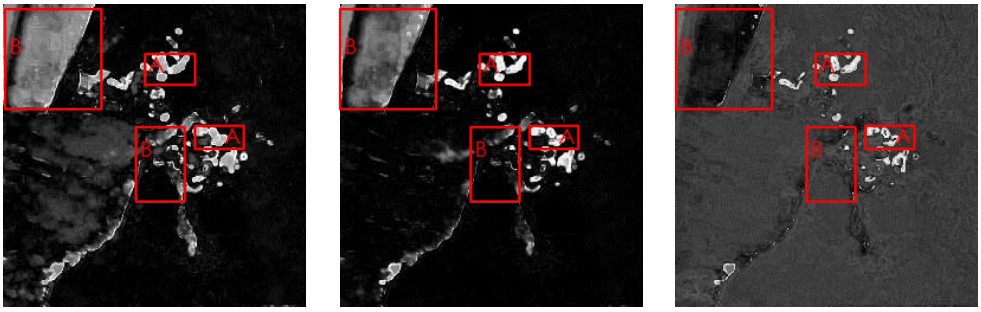
3.3 Compared with the PS^{3}VM-D and CS^{4}VM

Considering the **DATA$_2$**and **DATA$_3$**, the performance of the PS^{3}VM-D, CS^{4}VM and PM is compared. Fig. 10 and Fig. 11 show the detection grayscales of the three algorithms.



(a) (b) (c)

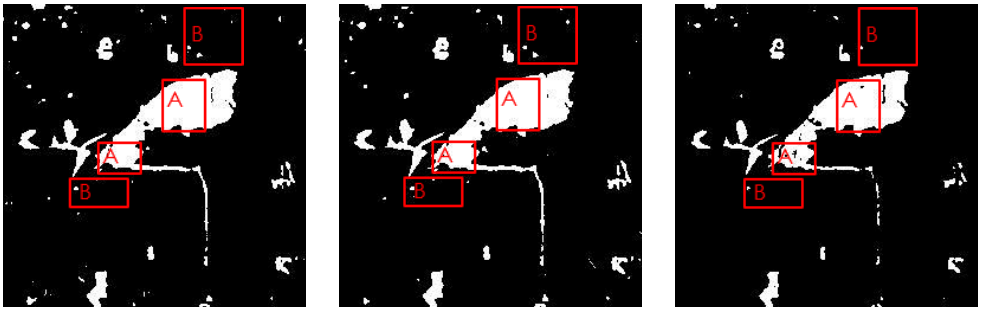
Fig. 10 Detection grayscales achieved on **DATA**$_2$ in terms of (a) CS^{4}VM, (b) PS^{3}VM-D and (c) PM.A represents the changed areas, B represents the unchanged areas



(a) (b) (c)

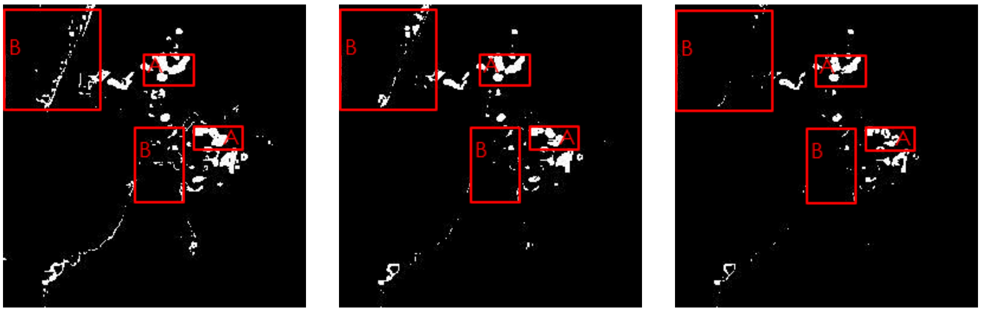
Fig. 11 Detection grayscales achieved on **DATA**$_3$ in terms of (a) CS^{4}VM, (b) PS^{3}VM-D and (c) PM.A represents the changed areas, B represents the unchanged areas

Through the threshold processing, the detection maps of **DATA**$_2$ and **DATA**$_3$are obtained, as shown in Fig. 12 and Fig. 13 respectively, where the white area is detected with the changes and the black area is detected as unchanged; A represents the changed areas and B represents the unchanged areas.



(a) (b) (c)

Fig. 12 Detection maps achieved on **DATA**$_2$ in terms of (a) CS^{4}VM, (b) PS^{3}VM-D and (c) PM， where the white area is detected with the changes and the black area is detected as unchanged; A represents the changed areas and the B represents the unchanged areas



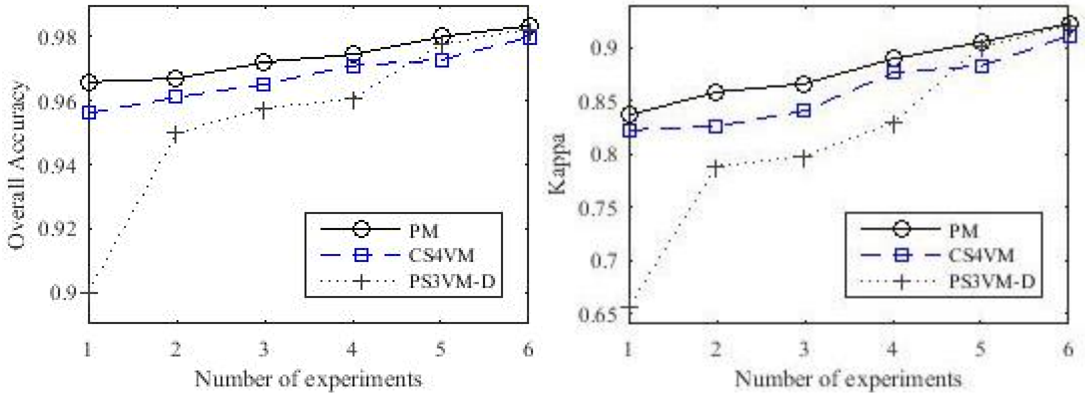
(a) (b) (c)

Fig. 13 Detection maps achieved on **DATA**$_3$ in terms of (a) CS^{4}VM, (b) PS^{3}VM-D and (c) PM， where the white area is detected

with the changes and the black area is detected as unchanged; A represents the changed areas and the B represents the unchanged areas

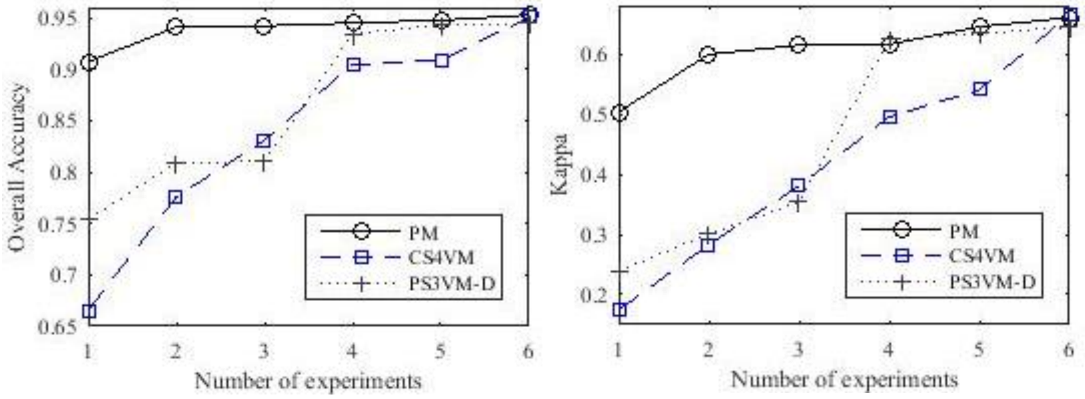
In Fig. 12, the omission of PS^{3}VM-D is the least and that of the PM is the greatest in A; the false alarm of PM is the least and that of the CS^{4}VMisthegreatest in B. It can be seen that the performance of the three algorithms is ideal and that of PM is more accurate than others for **DATA$_2$**. In Fig.13, the false alarm of PM is significantly less than that of the other algorithms in B, followed by thePS^{3}VM-D. The omission of three algorithms is almost the same in A. Therefore, the detection result of PM is the best for **DATA$_3$**.

In order to know the stability of the three algorithms, 6 different experiments were implemented for the CS^{4}VM, PS^{3}VM-D and PM, respectively with the same initial training sets to obtain the results in terms of OA and Kappa. The line charts of different results are drawn from small to large for **DATA$_2$**and **DATA$_3$**, which areshown in Fig.14 and Fig.15.



(a) (b)

Fig. 14 Line charts of CS^{4}VM, PS^{3}VM-D and PM achieved on **DATA$_2$** in terms of (a) OA and (b) Kappa



(a) (b)

Fig. 15 Line charts of CS^{4}VM, PS^{3}VM-D and PM achieved on **DATA$_3$** in terms of (a) OA and (b) Kappa

For **DATA$_2$**and **DATA$_3$**, except that the maximum of the OA and Kappa are basically the same for the three algorithms, the rest of PM are significantly higher than those of thePS^{3}VM-D and CS^{4}VM. For different initial sets, the change range of OA and Kappa for PS^{3}VM-D and CS^{4}VM is great and that for PM is small. Therefore, the detection accuracy and stability of the PMare higher than other algorithms.

DATA$_3$is selected to illustrate the convergence of the CS^{4}VM, PS^{3}VM-D and PM. The curve of OA is plotted in Fig. 16, where the X axis represents the number of iterations and the Y axis represents OA.

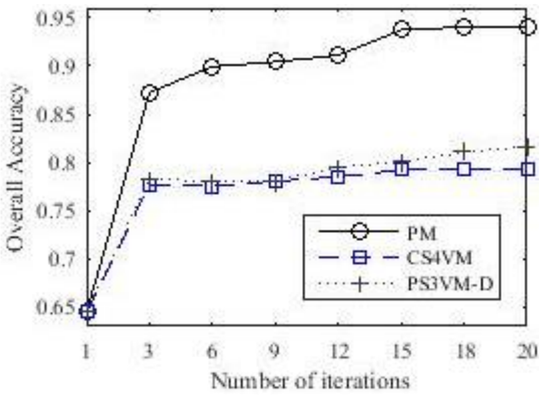


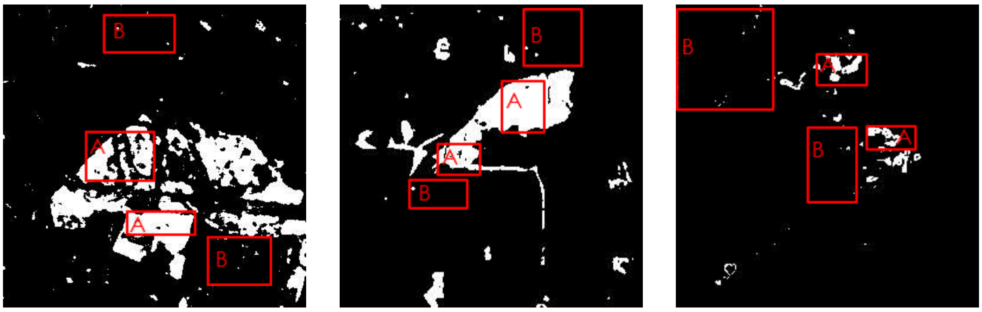
Fig. 16 Line charts of PM achieved on **DATA$_3$**

With the increase of iterations, the OAs of the three algorithms all change, where the OA of the CS^{4}VM increases from 0.64 to 0.79, the OA of thePS^{3}VM-D increases from 0.64 to 0.82 and the OA of PM increases from 0.64 to 0.94. The curve of the PM rises fastest, which suggests that the PM can get the faster convergence speed than other algorithms.

The CS^{4}VM can increase the relevance of the training samples by searching the contextual samples, but it does not consider whether those samples are beneficial to improve the convergence rate. The PS^{3}VM-D guarantees the information of the selected samples by searching them within the margin band, but the probability that these samples are polluted is also great. Compared to the CS^{4}VM and the PS^{3}VM-D, the PM uses the AL to select the informative samples, then use the context information to increase the samples correlation, which improves the detection accuracy and ensures the reliability of the selected samples. We use a new query function Q to make the distribution of the training samples more uniform for improving the detection efficiency and accuracy. Moreover, the convergence of the PM is further improved by applying the semilabeled samples.

3.4Fusion for PM

To obtain the more accurate results and verify the effectiveness of the fusion for the SSL methods, the experimental results of the PM are fused. The OA and Kappa are listed in Table 1, and the fusion detection maps are shown in Fig. 17, where the white area is detected with the changes and the black area is detected as unchanged. A represents the changed areas and B represents the unchanged areas.



(a) (b) (c)

Fig. 17 Fusion detection maps achieved on (a) **DATA$_1$**, (b) **DATA2**and (c) **DATA$_3$**for PM, where the white area is detected with the changes and the black area is detected as unchanged; A represents the changed areas and the B represents the unchanged areas

Table 1OA and Kappa of different experiments for **DATA$_1$**, **DATA$_2$** and **DATA$_3$**

	1		2		3		4		5		6		Fusion	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
DATA1	0.9718	0.8955	0.9662	0.8839	0.9668	0.8760	0.9717	0.8997	0.9413	0.8045	0.9724	0.8977	**0.9744**	**0.9060**
DATA2	0.9659	0.8582	0.9721	0.8661	0.9835	0.9225	0.9800	0.9053	0.9670	0.8365	0.9744	0.8894	**0.9837**	**0.9250**
DATA3	0.9475	0.6155	0.9420	0.6139	0.9409	0.6447	0.9451	0.5991	0.9072	0.5031	0.9528	0.6599	**0.9534**	**0.6638**

In TABEL 1, both the OA and Kappa get higher after fusion. By comparing Fig. 17(a), Fig. 17(b) and Fig. 17(c) with Fig. 8(b), Fig. 12(c) and Fig. 13(c) respectively, it is found that the false alarm in B and the omission in A reduces a lot after the fusion for **DATA$_1$** and **DATA$_3$**. The false alarm in B region and the omission in A decreases slightly for **DATA$_2$**. Obviously, the data fusion could effectively compensate the detection error, reduce the risk caused by the different initial training sets and improve the detection accuracy.

4. Conclusion

In this paper, a novel method is proposed for the change detection of RS images. It fully incorporates the methods and concepts in the SSL, and adopts them to fit the situation where the labeled samples are insufficient.

The novelty of this paper lies in: a) By considering the advantages of the AL and the context information, a novel semisupervised method is designed; b) by analyzing the distribution of the samples, a new query function is designed to select the semilabeled samples using the Euclidean distance; c) Based on the idea of data fusion, the discrete results of the PM are effectively fused. In the experiments of change detection for actual RSIs, the PM has made a significant improvement in the detection accuracy, convergence rate, and the stability in comparison with other existing methods. It can be further improved by using other effective fusion methods.

References

Anand, S., Mittal, S., Tuzel, O., & Meer, P. (2014). Semi-Supervised Kernel Mean Shift Clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 36*(6), 1201-1215, doi:10.1109/TPAMI.2013.190.

Bazi, Y., Melgani, F., & Al-Sharari, H. D. (2010). Unsupervised change detection in multispectral remotely sensed imagery with level set methods. *Geoscience and Remote Sensing, IEEE Transactions on, 48*(8), 3178-3187.

Bovolo, F., Bruzzone, L., & Marconcini, M. (2008). A Novel Approach to Unsupervised Change Detection Based on a Semisupervised SVM and a Similarity Measure. *Geoscience and Remote Sensing, IEEE Transactions on, 46*(7), 2070-2082, doi:10.1109/TGRS.2008.916643.

Bruzzone, L., & Persello, C. (2009). A Novel Context-Sensitive Semisupervised SVM Classifier Robust to Mislabeled Training Samples. *Geoscience and Remote Sensing, IEEE Transactions on, 47*(7), 2142-2154, doi:10.1109/TGRS.2008.2011983.

Celik, T. (2010). Change Detection in Satellite Images Using a Genetic Algorithm Approach. *Geoscience and Remote Sensing Letters, IEEE, 7*(2), 386-390, doi:10.1109/LGRS.2009.2037024.

Chapelle, O., Schölkopf, B., & Zien, A. (2006). Semi-supervised learning.

D'Elia, C., Ruscino, S., Abbate, M., Aiazzi, B., Baronti, S., & Alparone, L. (2014). SAR Image Classification Through Information-Theoretic Textural Features, MRF Segmentation, and Object-Oriented Learning Vector Quantization. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 7*(4), 1116-1126, doi:10.1109/JSTARS.2014.2304700.

Demir, B., Persello, C., & Bruzzone, L. (2011). Batch-Mode Active-Learning Methods for the Interactive Classification of Remote Sensing Images. *Geoscience and Remote Sensing, IEEE Transactions on, 49*(3), 1014-1031, doi:10.1109/TGRS.2010.2072929.

Didaci, L., Fumera, G., & Roli, F. (2012). Analysis of Co-training Algorithm with Very Small Training Sets. In G. Gimel'farb, E. Hancock, A. Imiya, A. Kuijper, M. Kudo, S. Omachi, et al. (Eds.), *Structural, Syntactic, and Statistical Pattern Recognition* (Vol. 7626, pp. 719-726, Lecture Notes in Computer Science): Springer Berlin Heidelberg.

Espinola, M., Piedra-Fernandez, J. A., Ayala, R., Iribarne, L., & Wang, J. Z. (2015). Contextual and Hierarchical Classification of Satellite Images Based on Cellular Automata. *Geoscience and Remote Sensing, IEEE Transactions on, 53*(2), 795-809, doi:10.1109/TGRS.2014.2328634.

Geiß, C., & Taubenböck, H. (2013). Remote sensing contributing to assess earthquake risk: from a literature review towards a roadmap. *Natural Hazards, 68*(1), 7-48, doi:10.1007/s11069-012-0322-2.

Gomez-Chova, L., Bruzzone, L., Camps-Valls, G., & Calpe-Maravilla, J. Semi-Supervised Remote Sensing Image Classification based on Clustering and the Mean Map Kernel. In *Geoscience and Remote Sensing Symposium, 2008. IGARSS 2008. IEEE International, 7-11 July 2008 2008* (Vol. 4, pp. IV - 391-IV - 394). doi:10.1109/IGARSS.2008.4779740.

Gong, C., Junwei, H., Lei, G., Zhenbao, L., Shuhui, B., & Jinchang, R. (2015). Effective and Efficient Midlevel Visual Elements-Oriented Land-Use Classification Using VHR Remote Sensing Images. *Geoscience and Remote Sensing, IEEE Transactions on, 53*(8), 4238-4249, doi:10.1109/TGRS.2015.2393857.

Habib, T., Inglada, J., Mercier, G., & Chanussot, J. (2009). Support Vector Reduction in SVM Algorithm for Abrupt Change Detection in Remote Sensing. *Geoscience and Remote Sensing Letters, IEEE, 6*(3), 606-610, doi:10.1109/LGRS.2009.2020306.

Hansen, M. C., & Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment, 122*, 66-74, doi:http://dx.doi.org/10.1016/j.rse.2011.08.024.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE, 13*(4), 18-28, doi:10.1109/5254.708428.

Izquierdo-Verdiguier, E., Laparra, V., Gomez-Chova, L., & Camps-Valls, G. (2013). Encoding Invariances in Remote Sensing Image Classification With SVM. *Geoscience and Remote Sensing Letters, IEEE, 10*(5), 981-985, doi:10.1109/LGRS.2012.2227297.

Junwei, H., Dingwen, Z., Gong, C., Lei, G., & Jinchang, R. (2015a). Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning. *Geoscience and Remote Sensing, IEEE Transactions on, 53*(6), 3325-3337, doi:10.1109/TGRS.2014.2374218.

Junwei, H., Dingwen, Z., Xintao, H., Lei, G., Jinchang, R., & Feng, W. (2015b). Background Prior-Based Salient Object Detection via Deep Reconstruction Residual. *Circuits and Systems for Video Technology, IEEE Transactions on, 25*(8), 1309-1321, doi:10.1109/TCSVT.2014.2381471.

Kawakita, M., & Kanamori, T. (2013). Semi-supervised learning with density-ratio estimation. *Machine Learning, 91*(2), 189-209, doi:10.1007/s10994-013-5329-8.

Li, J., Bioucas-Dias, J. M., & Plaza, A. (2010). Semisupervised Hyperspectral Image Segmentation Using Multinomial Logistic Regression With Active Learning. *Geoscience and Remote Sensing, IEEE Transactions on, 48*(11), 4085-4098, doi:10.1109/TGRS.2010.2060550.

Maulik, U., & Chakraborty, D. (2012). A novel semisupervised SVM for pixel classification of remote sensing imagery. *International Journal of Machine Learning and Cybernetics, 3*(3), 247-258, doi:10.1007/s13042-011-0059-3.

Maulik, U., & Chakraborty, D. (2014). Fuzzy Preference Based Feature Selection and Semisupervised SVM for Cancer Classification. *NanoBioscience, IEEE Transactions on, 13*(2), 152-160, doi:10.1109/TNB.2014.2312132.

Munoz-Mari, J., Tuia, D., & Camps-Valls, G. (2012). Semisupervised Classification of Remote Sensing Images With Active Queries. *Geoscience and Remote Sensing, IEEE Transactions on, 50*(10), 3751-3763, doi:10.1109/TGRS.2012.2185504.

Pasolli, E., Melgani, F., Tuia, D., Pacifici, F., & Emery, W. J. (2014). SVM Active Learning Approach for Image Classification Using Spatial Information. *Geoscience and Remote Sensing, IEEE Transactions on, 52*(4), 2217-2233, doi:10.1109/TGRS.2013.2258676.

Persello, C., & Bruzzone, L. (2014). Active and Semisupervised Learning for the Classification of Remote Sensing Images. *Geoscience and Remote Sensing, IEEE Transactions on, 52*(11), 6937-6956, doi:10.1109/TGRS.2014.2305805.

Schohn, G., & Cohn, D. Less is more: Active learning with support vector machines. In *ICML, 2000* (pp. 839-846): Citeseer

Scholkopf, B., Kah-Kay, S., Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T., et al. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *Signal Processing, IEEE Transactions on, 45*(11), 2758-2765, doi:10.1109/78.650102.

Shahshahani, B. M., & Landgrebe, D. A. (1994). The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. *Geoscience and Remote Sensing, IEEE Transactions on, 32*(5), 1087-1095, doi:10.1109/36.312897.

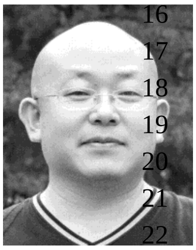
Tuia, D., Volpi, M., Copa, L., Kanevski, M., & Munoz-Mari, J. (2011). A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification. *Selected Topics in Signal Processing, IEEE Journal of, 5*(3), 606-617, doi:10.1109/JSTSP.2011.2139193.

Yi, Y., Wu, J., & Xu, W. (2011). Incremental SVM based on reserved set for network intrusion detection. *Expert Systems with Applications, 38*(6), 7698-7707.

Zhao, C., Li, X., Ren, J., & Marshall, S. (2013). Improved sparse representation using adaptive spatial support for effective target detection in hyperspectral imagery. *International Journal of Remote Sensing, 34*(24), 8669-8684, doi:10.1080/01431161.2013.845924.

Zhu, X. (2010). Semi-Supervised Learning. In C. Sammut, & G. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 892-897): Springer US.

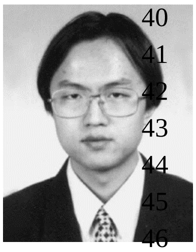
Author Biographies



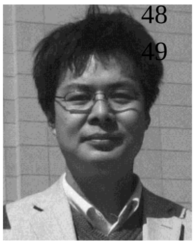
FeiGao received the B.S. and M.S. degree from the Xi’an Petroleum Institute, Xi’an, China, in 1996 and 1999, respectively, and the Ph.D. degrees from Beijing University of Aeronautics and Astronautics (BUAA), Beijing, China, in 2005. He is currently an Associate Professor with the School of Electronic and Information Engineering, BUAA. He is interested in radar signal processing, moving target detection and image processing.

Wenchao Lv received the B.S. degree in electronic and information engineering fromBeijing University of Posts and Telecommunications, Beijing, China, in 2014. He is currently pursuing the M.E. degree in electronic and communication engineering at Beijing University of Aeronautics and Astronautics, Beijing, China. His current research activity is in machine learning and synthetic aperture radar image classification.

Yaotian Zhang received the B.S. degree in Electronic Engineering from BeihangUniversityin 2003. He was awarded Ph.D degree in Signal Processing from Beihang University in 2009. Since2010, he has worked in School of Electronic and Information Engineering of Beihang University as an assistant professor. His research interest covers image understanding, target detection and Micro-Doppler signal analysis.



Jinping Sun received the M.Sc. and Ph.D. degrees from the Beijing University of Aeronautics and Astronautics (BUAA), Beijing, China, in 1998 and 2001, respectively. He is currently a Professor with the School of Electronic and Information Engineering, BUAA. His research interests include high-resolution radar signal processing, image understanding, and robust beamforming.



Jun Wang received the B.S. degree from the Northwestern Polytechnical University, Xi’an, China, in 1995 and the M.S. and Ph.D. degrees from the Beijing University of Aeronautics and Astronautics

(BUAA), Beijing, China, in 1998 and 2001, respectively. He is currently a Professor with the School of Electronic and Information Engineering, BUAA. He is interested in signal processing, DSP/FPGA real-time architecture, target recognition and tracking, and so on. His research has resulted in 38 papers in journals, books, and conference proceedings.

Erfu Yang is a Lecturer in the Department of Design Manufacture and Engineering Management (DMEM) at the University of Strathclyde, Glasgow, UK. His main research interests include robotics, autonomous systems, mechatronics, manufacturing automation, computer vision, image/signal processing, nonlinear control, process modelling and simulation, condition monitoring, fault diagnosis, multi-objective optimizations, and applications of machine learning and artificial intelligence including multi-agent reinforcement learning, fuzzy logic, neural networks, bio-inspired algorithms, and cognitive computation, etc. He has over 60 publications in these areas, including more than 30 journal papers and 5 book chapters.