ABSTRACT
This article addresses a new classification technique: Partially Supervised Classification (PSC), which is used to identify a specific land-cover class of interest from a remotely sensed image using unique training samples that belong to a specified class. This article also presents and discusses a newly proposed novel Support Vector Machine (SVM) algorithm for PSC. Accordingly, its training set includes labeled samples that belong to the class of interest and unlabeled samples of all classes randomly selected from a remotely sensed image. Moreover, all unlabeled samples are assumed to be training samples of other classes, and each of them is assigned a weight factor indicating the likelihood of this assumption; hence, the algorithm is called Weighted Unlabeled Sample SVM (WUS-SVM). Based on the WUS-SVM, a PSC method is proposed. Experimental results with both simulated and real datasets indicate that the proposed PSC method can achieve encouraging accuracy and is more robust than the 1-SVM and the Spectral Angle Mapping (SAM) method.

Keywords: partially supervised classification; support vector machine; unlabeled samples

INTRODUCTION
In spatial data mining, it is important to discover spatial distribution of objects based on remote sensing imagery. Classification of remote sensing imagery is one of the methods used to identify the spatial distribution of land-cover classes. Generally, there are two traditional classification techniques: unsupervised and supervised. Although training samples are unnecessary for nonsupervised classification, the classes derived by this kind of technique are unknown. In supervised classification,
each class is defined by training samples that are selected by users. However, sufficient and exhaustive training samples are required for supervised classification. In many cases, a classifier that can recognize only a specific land-cover class of interest in a remotely sensed image is sufficient. Since obtaining training samples is always expensive and time-consuming, it would be very useful if a classifier were designed only to require training samples that belong to the specific class of interest ($C_{int}$). This technique has been termed Partially Supervised Classification (PSC) (Jeon & Landgrebe, 1999).

There are two main classification schemes: relative and absolute (Jeon & Landgrebe, 1990). The former, such as the Maximum-Likelihood Classifier (Richards, 1993) and Support Vector Machine (SVM) (Vapnik, 1995), usually provides acceptable classification accuracy but is not suitable for PSC, because it requires all the training samples of different classes to be present in the corresponding remotely sensed image under the analysis process. The latter, such as the parallelepiped classifier (Richards, 1993) and I-SVM (Koppel & Schler, 2004; Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001; Tax, 2001), allows the classification task to be performed solely using training samples that belong to $C_{int}$, thus, it is suitable for PSC. In spite of this advantage, the accuracy of this type of approach is always limited or heavily dependent on the selection of certain thresholds or parameters.

Many techniques have been developed to achieve a satisfactory accuracy of PSC. Most of them try to acquire information of the classes other than $C_{int}$ using unlabeled samples of both $C_{int}$ and other classes ($C_{others}$) and deal with PSC by a relative classifier. Liu, Lee, Yu, and Li (2002) developed a technique called S-EM based on the EM algorithm (Dempster, Laird, & Rubin, 1977). However, S-EM is not accurate because of its weak classifier. Jeon and Landgrebe (1990, 1999) proposed two methods based on probability density estimation. Fernández-Prieto (2002) proposed an improved method in which Radial Basis Function network and Markov Random Fields are used. Although they can improve the accuracy of PSC, all the methods are based on the assumption that the probability density of $C_{int}$ is known or can be estimated correctly by training samples. However, in many cases, the probability density of $C_{int}$ is unknown, and it is difficult to be estimated from training samples, especially when only limited training samples are available.

Instead of taking density estimation as an intermediate step, SVM, which forms a decision function directly, can achieve a satisfied accuracy even with a small training sample set (Vapnik, 1995). Because of this important fact, several SVM-based algorithms have been proposed, such as PEBL (Yu, Han, & Chang, 2002), Roc-SVM (Li & Liu, 2003) and biased SVM (Liu, Yang, Li, Lee, & Yu, 2003) with a common feature of transforming PSC into a binary classification to distinguish $C_{int}$ from $C_{others}$. Both PEBL and Roc-SVM are based on a two-step strategy. In the first step, a set of reliable samples belonging to $C_{others}$ are identified, whereas in the second step, an SVM classifier is trained with training samples of both classes, $C_{int}$ and $C_{others}$. Together, these two steps can be seen as an iterative method of increasing the number of unlabeled samples that are classified as $C_{others}$ while maintaining the correctly classified samples of $C_{int}$. Unlike these two-step algorithms, biased SVM is to assume that all unlabeled samples belong to $C_{others}$ and to try to minimize the number of those unlabeled samples classified as $C_{int}$ while constraining labeled samples of $C_{int}$ in order to be correctly classified. It shows that biased SVM is superior to PEBL and Roc-SVM (Liu et al., 2003).

There are some related studies, including semi-supervised learning (Bruzzone, Chi, & Mattia, 2005; Dundar, 2003; Kristin & Bennett, 1998; Shahshahani & Landgrebe, 1994), transductive learning (Joachims, 1999), and co-training (Blum & Mitchell, 1998), which also take advantage of unlabeled samples in order to improve classification performance. However, they are different from our study, since we use no labeled sample of $C_{others}$.

This article presents and discusses a newly proposed novel SVM algorithm for PSC (Liu,
Robust Statistical Methods for Rapid Data Labelling
Jamie Godwin and Peter Matthews (2014). *Data Mining and Analysis in the Engineering Field* (pp. 107-141).
www.igi-global.com/chapter/robust-statistical-methods-for-rapid-data-labelling/109979?camid=4v1a