Mach Learn (2016) 103:343–375
DOI 10.1007/s10994-016-5559-7

Transductive hyperspectral image classification: toward integrating spectral and relational features via an iterative ensemble system

Annalisa Appice[1,2,3] · **Pietro Guccione**[4] · **Donato Malerba**[1,2,3]

Received: 13 December 2014 / Accepted: 23 February 2016 / Published online: 22 March 2016

Abstract Remotely sensed hyperspectral image classification is a very challenging task due to the spatial correlation of the spectral signature and the high cost of true sample labeling. In light of this, the collective inference paradigm allows us to manage the spatial correlation between spectral responses of neighboring pixels, as interacting pixels are labeled simultaneously. The transductive inference paradigm allows us to reduce the inference error for the given set of unlabeled data, as sparsely labeled pixels are learned by accounting for both labeled and unlabeled information. In this paper, both these paradigms contribute to the definition of a spectral-relational classification methodology for imagery data. We propose a novel algorithm to assign a class to each pixel of a sparsely labeled hyperspectral image. It integrates the spectral information and the spatial correlation through an ensemble system. For every pixel of a hyperspectral image, spatial neighborhoods are constructed and used to build application-specific relational features. Classification is performed with an ensemble comprising a classifier learned by considering the available spectral information (associated with the pixel) and the classifiers learned by considering the extracted spatio-relational information (associated with the spatial neighborhoods). The more reliable labels predicted by

Editors: Jesse Davis and Jan Ramon.

✉ Annalisa Appice
annalisa.appice@uniba.it

Pietro Guccione
guccione@poliba.it

Donato Malerba
donato.malerba@uniba.it

1 Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro", via Orabona 4, 70125 Bari, Italy

2 CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Rome, Italy

3 CILA - Centro Interdipartimentale di Logica e Applicazioni, Bari, Italy

4 Dipartimento di Ingegneria Elettrica ed Informazione, Politecnico di Bari, via Orabona, 4, 70125 Bari, Italy

the ensemble are fed back to the labeled part of the image. Experimental results highlight the importance of the spectral-relational strategy for the accurate transductive classification of hyperspectral images and they validate the proposed algorithm.

Keywords Relational classification · Iterative learning · Transduction · Collective inference · Ensemble learning

1 Introduction

Remote sensing focuses on collecting and interpreting information about a scene without having physical contact with the scene. Aircraft and satellites are the common platforms for remote sensing of the Earth and its natural resources (Goetz et al. 1985). They measure the electromagnetic radiation which is reflected from the Earth's surface materials, by producing measurements of energy in various parts of the electromagnetic spectrum. For applications in visible or Near Infrared, the spectrum, which can nominally range from 0.4 to 14 micrometers (μm) wavelength (20–750 THz frequency) is segmented into spectral regions (bands). A *spectral band* is a discrete interval (wavelength) of the electromagnetic spectrum in which a scanning instrument measures both reflectance and absorption of radiation at a specific geographic location. For example, visible light ranges in a band from 0.4 to 0.8 μm. A *spectral signature* is a range of contiguous wavelengths in the electromagnetic spectrum. A *hyperspectral image* (also called hyperspectral imagery dataset) is a collection of measurements taken on a topographic scene in a spectral signature with a large number of narrow, contiguous wavelength bands (see Fig. 1a). For example, the majority of hyperspectral data, collected through commonly deployed sensing systems (e.g. ROSIS, HySpex 1995; AVIRIS 2007), are measurements acquired simultaneously in 100–200 contiguous spectral bands at the nominal spectral resolution of 10 nanometers (Ablin and Sulochana 2013). They are collected over a few square Kms (from tens to hundreds of thousands of pixels), taken in high resolution (a few meters) and covering the wavelength region from 0.4 to 2.5 μm.

Different kinds of surfaces reflect radiating electromagnetic waves (e.m.) in different ways, due to the chemical composition, texture, color, roughness and moisture (Ablin and Sulochana 2013). This means that all the Earth's surface features, including minerals, vegetation, dry soil, water and snow, have unique spectral reflectance signatures. Hence, the spectral signature of different surfaces (e.g. soil, water, ice, vegetation) can contain informa-

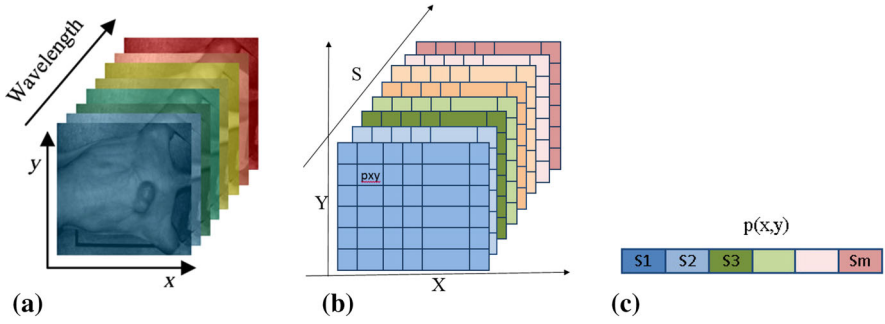

Fig. 1 Hyperspectral imagery dataset. **a** Hyperspectral image, **b** imagery matrix, **c** spectral signature of $p(x, y)$

tion to distinguish the different surface objects. Imaging spectroscopy (Green et al. 1998), also known as *hyperspectral imaging*, is concerned with the analysis and interpretation of spectral signatures of hyperspectral data acquired from a given scene. This kind of analysis can be used to detect slight changes in vegetation, soil, water and mineral reflectance (Plaza et al. 2009). Hyperspectral imaging is attracting growing interest in applications such as urban planning, agriculture, forestry and monitoring (Ablin and Sulochana 2013). In particular, *hyperspectral image classification* produces thematic maps from hyperspectral data. A thematic map represents the Earth's surface objects. Its construction implies that themes or categories, selected for the map, are distinguished in the remote sensed image. Classification assigns a known class (theme) to each pixel (imagery data example). Every pixel is expressed with a vector space model that represents the spectral signature as a vector of numeric features (namely *spectral features*) and is also associated with a specific position in a uniform grid, which describes the spatial arrangement of the scene. It is assigned a certain (possibly unknown) spectral response, i.e. class label.

Methodologically, the automatic classification of hyperspectral data is non-trivial due to factors such as the spatial correlation of the spectral features, the high cost of labeling the data and the large number of spectral bands (Plaza et al. 2009). In this paper, we propose a novel transductive collective classifier for dealing with all these factors in hyperspectral image classification. Collective inference exploits the spatial correlation between spectral responses (class labels) of neighboring pixels by simultaneously labeling interacting pixels. The transductive inference paradigm allows us to reduce the inference error for the given set of unlabeled data, as sparsely labeled pixels are learned by accounting for both labeled and unlabeled information.

The paper is organized as follows. The next section clarifies the motivation and the contribution of this paper. Section 3 reports the basics of the presented algorithm. Section 4 illustrates related work. Section 5 presents the proposed algorithm, while Sect. 6 reports the analysis of the learning complexity. Section 7 describes the datasets, the experimental setup and reports the results. Finally, in Sect. 8 some conclusions are drawn.

2 Motivation and contributions

The *spatial correlation* of the spectral signature refers to the relation (or dependence) between spectral signatures of pixels, due to their spatial proximity. Intuitively, spatial correlation means that the features for a specific pair of points are more (less) similar than would be expected for a random pair of points (Legendre 1993). In the case of hyperspectral images of geographical areas, spatial correlation exists in the positive form, as there is a *slowly progressive* spatial variation both in the spectral signature and in the spectral label (Miao et al. 2014). This means that by picturing the spatial variation of the observed features in a map, we may observe regions where the distribution of values is smoothly continuous, with some boundaries possibly marked by sharp discontinuities. The presence of spatial correlation violates the independence assumption (i.i.d.) made by most traditional classifications. This can lead to poor performance in statistical models (LeSage and Pace 2001) and machine learned models (Neville et al. 2004), although some models are robust to violations of this assumption (Dundar et al. 2007). In any case, an emerging trend in hyperspectral imaging is to accommodate spatial correlation into the hyperspectral classification process as this improves predictive performance (Plaza et al. 2009; Fauvel et al. 2013). This improvement was also observed by recent machine learning studies in other predictive tasks involving spatial correlation such as regression, interpolation and forecasting (Stojanova et al. 2013;

Appice and Malerba 2014; Appice et al. 2014). Further motivation is driven by the recently emerged perspective on the importance of a relational learning approach in spatial data mining (Malerba 2008).

Recent research in relational data mining has explored the use of the *collective inference* paradigm to exploit data correlation when learning predictive models. According to Jensen et al. (2004) and Getoor and Taskar (2007), collective inference refers to the combined classification of a set of correlated instances. In contrast to traditional algorithms which label data instances individually regardless of correlations among the instances, collective inference predicts the labels of instances simultaneously and exploits correlations among the instances. In hyperspectral imagery classification, collective inference offers a unique opportunity to explicitly account for the spatial variation of the spectral signature, by reducing the labeling uncertainty that may exist when only spectral information is used and helping to overcome the salt and pepper appearance of the classification (Fauvel et al. 2013; Chen et al. 2014; Khodadadzadeh et al. 2014b).

The high-dimensionality of spectral data and the small number of ground truth labels can cause problems such as a reduction in classification accuracy. This behavior is known as Hughes' phenomenon (Hughes 1968). In particular, the limited ground-truth samples are not always sufficient for a reliable estimate of the classifier's parameters. In fact, if the number of samples (training set) is too low compared to the number of variables, we risk overfitting the training data, i.e. we can learn a model that exactly fits the training data without accounting for a wider generalization (Chang 2007).

Both semi-supervised and transductive learning can help cope with limited labeled data in high dimensional problems (Shahshahani and Landgrebe 1994). They jointly exploit labeled and unlabeled data to reduce the impact of overfitting (Seeger 2001). Both settings have recently attracted an increasing amount of interest in remote sensing (Wang et al. 2014). The semi-supervised setting is a type of inductive learning, since the learned function is used to make predictions on any possible example. The *transductive setting* is less demanding - it is only interested in reducing the inference error of the given set of unlabeled data, without trying to improve the overall quality of the learned model. As pointed out by Vapnik (1995), the idea of transduction (labeling a test set) appears inherently easier than (semi-supervised) induction (learning a general rule).

This paper proposes spectral and spatio-relational transductive ensemble of classifiers (S^2TEC), that is, a novel hyperspectral imagery transduction classification algorithm to cope with limited number of labeled examples in the high dimensional spectral space. The proposed algorithm iteratively constructs various spatio-relational features over spatial neighborhoods via a *collective* iterative convergence algorithm. It uses an *ensemble system* of spectral and spatio-relational classifiers to determine labels of the imagery pixels and applies *transductive learning* to make accurate predictions. Spatio-relational features model the continuity of neighboring labels. They exploit the likely fact that two neighboring pixels may have the same label (label spatial correlation). This is somewhat the same principle as the application of the Markov random fields in hyperspectral imaging (Li et al. 2011, 2012, 2013a; Tarabalka et al. 2010b; Khodadadzadeh et al. 2014b).

Collective inference, transductive learning and ensemble learning have been explored already in the literature. However, to the best of our knowledge, this is the first study that combines these three strategies in a single learning framework. This framework, which represents one of the main contributions of this work, proves effective for the challenging problem of hyperspectral classification. Another contribution is the investigation of various application-specific relational operators to define a collective classification setting. We use both operators to describe the class frequency and operators to describe the class morphol-

ogy of a hyperspectral image. Although these operators have been already investigated in the hyperspectral classification literature (Guccione et al. 2015; Khodadadzadeh et al. 2014a; Pesaresi and Benediktsson 2001; Tan et al. 2014), they have been considered separately. In this study, inspired by the fact that they convey different kinds of information, we consider their combined use. Indeed, the frequency information is computed to quantitatively describe the label structure making a spatial average (a sort of "low-pass" filtering), while the morphology information is computed to qualitatively follow the borders separating land cover types (a sort of "high-pass" filtering).

We assess the efficacy of the algorithm in a real-world application, made complex by the presence of spatial information and by the scarceness of labeled information. In this setting, we claim that performing an iterative construction of application-specific relational features joined to transductive learning and ensemble classification can lead to accurate final classifications. This would happen even when starting from a reduced labeled set, since the combined framework reasonably converges towards a stable solution. In the paper, we justify this claim by showing empirically this point of view. By following the main stream of research in hyperspectral image classification, the effectiveness of our contributions is assessed via an empirical study on three benchmark hyperspectral datasets, corresponding to various contexts. These datasets are those used in the majority of hyperspectral imaging literature (e.g. Plaza et al. 2009; Tarabalka et al. 2010a; Khodadadzadeh et al. 2014b; Li et al. 2013a; Fauvel et al. 2012; Wang et al. 2014; Guccione et al. 2015). We evaluate the accuracy of both the proposed algorithm and several competitors by computing the metrics (overall accuracy, average accuracy, Cohen's kappa coefficient, F-1 score), which are usually considered by the hyperspectral image classification community and the relational learning community. The experimental results show that all components of our proposal contribute to its efficacy. The presented algorithm outperforms several supervised and transductive classifiers defined in the machine learning literature, as well as to specific competitors defined in the hyperspectral image analysis literature.

Hence, this work is relevant for the relational learning community as it contributes to assessing that combining collective classification and relational features can gain improvements over a propopositional/non collective setting. These improvements are shown to be relevant in an applicative context (remote sensing) that has recently gained importance. This work is also significant for the hyperspectral image classification community, as it describes an algorithm that deals with spectral and spatial information by gaining accuracy with respect to state-of-the-art algorithms.

3 Basics

Let $\mathcal{D}$ be a *hyperspectral imagery dataset*, that is, a set of pixels (examples). Every pixel represents a region of around a few square meters of the Earth's surface (i.e. it is a function of the sensor's spatial resolution). It is associated with the spatial coordinates XY, as well as with the m-dimensional vector of spectral features $\mathbf{S} = S_1, S_2, \ldots, S_m$ (see Fig. 1c). Every spectral feature S_i is a numeric feature that expresses how much radiation is reflected, on average, across the pixel region, at the ith band of the considered spectral profile. According to the general formulation of the transductive setting (Vapnik 1995), pixels of $\mathcal{D}$ are sparsely labeled according to an unknown target function, whose range is a finite set of classes $C = \{C_1, C_2, \ldots, C_k\}$. Every class C_i represents a distinct theme (i.e. type of Earth's surface). In general, pixels are equally distributed in space over a regular grid, so that a hyperspectral

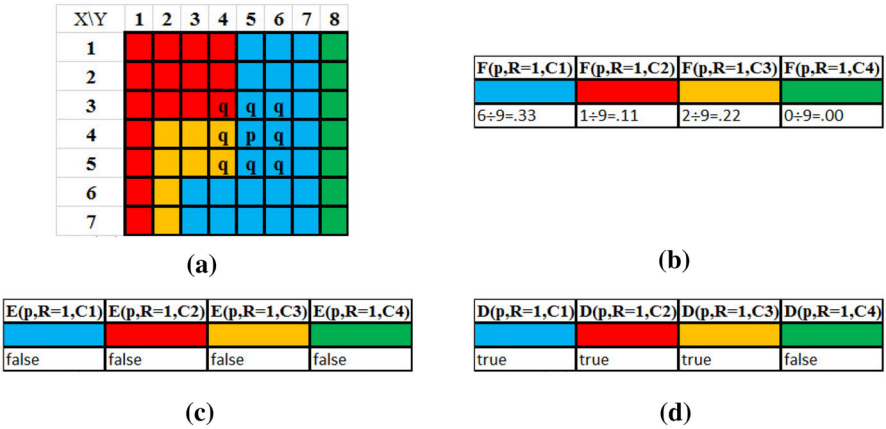

Fig. 2 Spatial neighborhoods and relational features: **a** the spatial neighborhood constructed for the pixel p with radius $R = 1$ and a square shape; **b** the relational features constructed for the pixel p over the spatial neighborhood $\mathcal{N}(p, R = 1)$ with the frequency operator; **c** the relational features constructed for the pixel p over the spatial neighborhood $\mathcal{N}(p, R = 1)$ with the erosion operator; **d** the relational features constructed for the pixel p over the spatial neighborhood $\mathcal{N}(p, R = 1)$ with the dilation operator

dataset can be represented as a matrix (see Fig. 1b). Thus, the spatial coordinate X is associated with the row index, while the spatial coordinate Y is associated with the column index of the matrix. Based on this premise, let $p(x, y)$ be a pixel located at the (x, y) row-column position of the imagery matrix. A *spatial neighborhood* is a set of pixels q (task-relevant pixels) surrounding p (target pixel) in the matrix.

In the imagery analysis literature, spatial neighborhoods frequently have a square shape (Plaza et al. 2009; Guccione et al. 2015), although alternative shapes like a circle or a cross can be also considered. Formally, let R be a positive, integer-valued radius, the *square-shaped* spatial neighborhood $\mathcal{N}(p, R)$ of pixel p (see Fig. 2a) is the set of imagery pixels $q(x + I, y + J)$, so that $-R \leq I, J \leq +R$. The construction of one or more spatial neighborhoods, coupled with every pixel of a hyperspectral imagery dataset, allows us to define the actual *(spatio-)relational structure* of the dataset. This definition of a relational data structure allows us to pass from a propositional representation of imagery data (spectral information) to a relational representation (spatial-aware information) of the same data.

In this study, spatio-relational features are constructed by resorting to a collective inference procedure, in order to express the label of a target pixel depending on the labels of all the related (task-relevant) neighbors of the target pixel. These features are formed through the application of the *frequency-based operator* (Guccione et al. 2015; Khodadadzadeh et al. 2014a) and/or the *morphology-based operators* (Pesaresi and Benediktsson 2001; Tan et al. 2014). Given a target imagery pixel p and its spatial neighborhood $\mathcal{N}(p, R)$ (with radius R), the frequency operator constructs k features, one feature for each class label C_i. The value of the feature is proportional to the pixels within $\mathcal{N}(p, R)$ that have label C_i (see Fig. 2b).

On the other hand, the morphological operators construct $4 \cdot k$ features, one feature for each morphological operator (erosion, dilation, opening and closing) and for each class label C_i. They use spatial neighborhoods as structuring elements. The erosion and dilation of a class label destroy (Fig. 3a–d) and enhance (Fig. 3e–h), respectively the structure and the density of borders separating land cover types present in the structuring element (Benediktsson et al. 2003; Soille 2003). Erosion $E(p, R, C_i)$ is true if all pixels within $\mathcal{N}(p, R)$ have label C_i

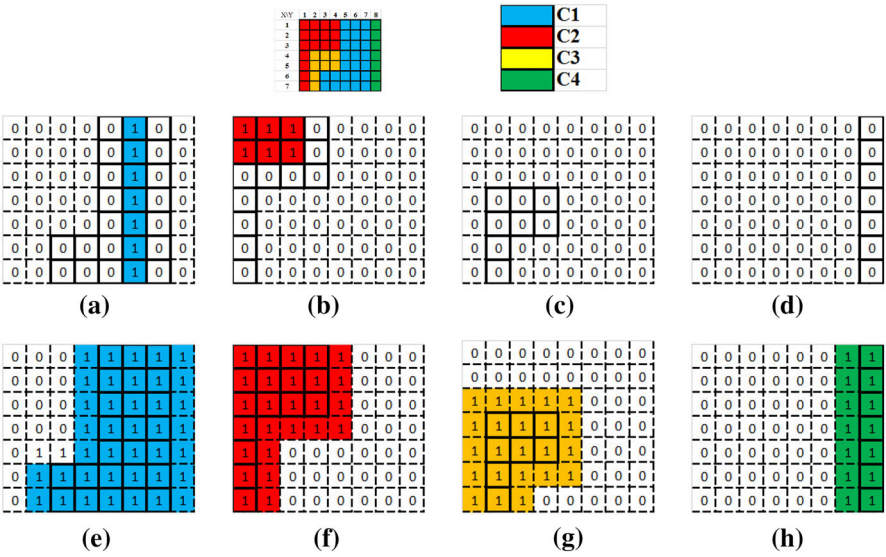

Fig. 3 Erosion (**a**–**d**) and dilation (**e**–**h**) of a hyperspectral image computed with square neighborhoods having radius $R = 1$

(see Fig. 2c), while dilation $D(p, R, C_i)$ is true if at least one pixel within $\mathcal{N}(p, R)$ has label C_i (see Fig. 2d). The opening and closing operators are combinations of erosion and dilation. Opening is erosion followed by dilation. It recovers most structures of the original image, i.e. structures that were not removed by the erosion and are bigger than the structuring element. Closing is dilation followed by erosion. With opening or closing we can obtain objects of the image which are larger or smaller than the structuring element (Fauvel et al. 2013).

4 Related work

This paper draws on methodological work in relational learning, collective classification and transductive learning as well as applied work on hyperspectral image classification.

4.1 Relational learning

Relational representations and relational learning algorithms have been investigated in the literature, in order to deal with spatial correlation in several real-world applications (see Malerba 2008 for a survey). Relational learning algorithms can be directly applied to various representations of spatial data, i.e. collections of geo-located entities. They account for spatial correlation that biases learning in spatial domain. Furthermore, discovered relational patterns reveal those spatial relationships, which correspond to spatial domains.

Closely related to the application context of this study are relational learning algorithms investigated to process document images, images from medical domains and images from one's daily life. Ceci et al. (2007) have proposed multi-relational data mining algorithms to account for spatial dimension of page layout when recognizing semantically relevant components in the layout extracted from a document image. Mizoguchi et al. (1997) have applied inductive logic programming to identify glaucomatous eyes from ocular fundus images,

while Sammut and Zrimec (1998) have applied inductive logic programming to construct concept descriptions from X-ray angiograms and classify types of blood vessels. Finally, both Chechetka et al. (2010) and Antanas et al. (2014) have investigated how spatial relational representations can be generated from images and defined using a logical background theory (i.e. a set of Prolog rules, as in relational learning). As an example, we can consider the relation "close aligned horizontally to the right" and the relation "part of". This relational knowledge is then used to recognize higher-level structures in images from daily life.

In contrast to relational approaches that use inductive logic programming, our work builds on the idea of constructing spatio-relational features of imagery data, which can be represented in a vector space model. This idea has recently received growing attention in hyperspectral imaging (Plaza et al. 2009; Benediktsson et al. 2003; Guccione et al. 2015; Khodadadzadeh et al. 2014a). The main motivation for it is that it allows us to easily inject spatial information into efficient, attribute-value algorithms (i.e. propositional learners), which are effective when learning spectral information (e.g. SVM Plaza et al. 2009). We note that hyperspectral imaging approaches, which resort to the construction of spatio-relational features, belong to the category of propositionalization approaches to relational data mining.

Propositionalization (Srinivasan and King 1999; Železný and Lavrac 2006; Ceci and Appice 2006; Krogel et al. 2003) can be seen as a transformation of relational learning problems into attribute-value representations amenable for propositional learners. Propositionalization algorithms are divided into two categories: logic-oriented algorithms and database-oriented algorithms (Krogel et al. 2003). Logic-oriented algorithms handle complex background knowledge and provide expressive first-order models, in order to generate attribute-value features. Database-oriented algorithms mainly explore foreign key relationships as a basis for a declarative bias during propositionalization and apply aggregation functions (e.g. mean, mode, SD, count), which are widely used in the database area, in order to generate attribute-value features. The spatio-relational feature construction in hyperspectral imaging can be considered as a kind of database-oriented propositionalization algorithm. It explores the spatial neighboring relationship between pixels, in order to aggregate information of sets of neighbor pixels (spatial neighborhoods). Contrary to existing propositionalization algorithms, application-specific operators (e.g. morphology-based operators Pesaresi and Benediktsson 2001; Tan et al. 2014 and/or frequency-based operators Guccione et al. 2015; Khodadadzadeh et al. 2014a) are considered to generate attribute-value features.

4.2 Collective inference

Collective inference is a fundamental approach to classification in relational domains (Jensen et al. 2004; Getoor and Taskar 2007). Collective classification algorithms predict labels of related instances simultaneously. These algorithms are grouped into global algorithms and local algorithms (Sen et al. 2008). Global algorithms train a classifier that seeks to optimize a global objective function. They are often based on a Markov random field and use loopy belief (LBP) propagation (Weiss 2001; Taskar et al. 2002; Neville and Jensen 2007; Sen et al. 2008) or mean-field (MF) relaxation labeling (Weiss 2001; Sen et al. 2008), in order to avoid the computational complexity of computing marginal probability distributions. Instead of working with the probability distribution associated with the MRF directly, they both use a simpler "trial" distribution. Local algorithms employ an iterative process whereby a local classifier predicts labels for each instance, by using features of the instances and relational features derived from the related instances. This type of approach involves an iterative process to update the labels and the relational features of the related instances, e.g.

iterative convergence-based approaches (ICA) (Neville and Jensen 2000; Getoor 2005) and gibbs sampling (GS) approaches (Jensen et al. 2004).

Global (MF and LBP) and local (ICA and GS) approaches have been empirically compared in Sen et al. (2008). This study has revealed that global approaches achieve, in several cases, a better accuracy than local ones. However, global approaches are also the most difficult to work with in learning and inference. In particular, choosing the initial weights, so that they will converge during training, is nontrivial. The most trouble is observed with MF, which may be unable to converge or, when it does, it may not converge to the global optimum. This analysis is consistent with previous work (Weiss 2001; Yanover and Weiss 2002) and, as reported by Sen et al. (2008), it should be taken into consideration when choosing to apply these algorithms. On the other hand, by focusing attention on local approaches, Sen et al. (2008) have also concluded that ICA and GS can produce very similar results, but ICA is much faster than GS. These considerations motivate our preference towards implementing collective inference through iterative convergence learning.

The iterative convergence approaches are investigated in many studies (Neville and Jensen 2000; Bilgic et al. 2007; McDowell et al. 2007; Fang et al. 2013). They account for the correlation of labels and compute the label of an instance depending on the labels of all its related neighbors. In particular, they express an instance by combining the instance features and the relational features constructed by using the labels of all the related neighbors of the instance. The relational features are computed by using an aggregation function over the neighbors, such as count, mode and proportion. Based on the descriptive features and the relational features, an algorithm trains a classifier and iteratively updates the predictions of all instances, by using the predictions for instances with known labels. This process continues until the algorithm converges. Saha et al. (2012) have recently described an iterative convergence algorithm to deal with multi-label classification problems. Finally, collective classification has been recently investigated in combination semi-supervised and transductive learning (Shi et al. 2011; McDowell and Aha 2012).

4.3 Transductive learning

Transductive learning (Vapnik 1998) is a learning paradigm that exploits a large amount of unlabeled data when a small amount of labeled data is available. It assumes that the testing data are exactly the unlabeled data. Many transductive learning algorithms have been proposed in the literature. Joachims (1999) has formulated an optimization algorithm for learning transductive support vector mfachines (TSVMs). This algorithm exploits the structure in both training and testing data for better positioning the maximum margin hyperplane. Subsequently, Sindhwani and Keerthi (2006) have formulated a fast, multi-switch implementation of the TSVMs, called SVMLin, which is significantly more efficient and scalable than the previous algorithm. They have exploited data sparsity and linearity of the problem, in order to provide superior scalability. They have investigated a multiple switching heuristic that further improves TSVM training by an order of magnitude. In particular, according to the multi-switch modality, more than one pair of labels may be switched in each iteration. These speed enhancements turn TSVM into a feasible tool for large-scale applications. In addition, they adopted deterministic annealing techniques, in order to alleviate the problem of local minima in the TSVMs. Another family of transductive algorithms is investigated in graph mining. A graph is defined with the nodes representing both labeled and unlabeled instances, while the edges reflect the similarity of instances. Graph-based approaches usually assume label smoothness over the graph. One example is to exploit the structure of the entire data set in the search for min cuts (Blum and Chawla 2001) or for min average cuts (Joachims

2003) on the graph. Finally, recent advances include transductive algorithms for multi-label classifications, to effectively assign a set of multiple labels to each instance (Kong et al. 2013), as well as transductive relational probabilistic classifiers (Taskar et al. 2001; Malerba et al. 2009; Ceci et al. 2012), to apply transduction in probabilistic relational learning.

4.4 Hyperspectral image classification

Over the last two decades, several supervised machine learning algorithms have been applied to hyperspectral image classification. Spectral information is processed, in order to train a classifier with the labeled data samples. The quality of these classification algorithms was strongly related to the quality and number of training samples under the influence of Hughes' phenomenon. In this context, support vector machines (SVMs) have been widely used to deal with Hughes' phenomenon by addressing large feature spaces and producing solutions from sparsely labeled data (Melgani and Bruzzone 2004; Huang et al. 2002). Recently, multinomial logistic regression (Li et al. 2011, 2012) has been shown to provide an alternative approach to deal with ill-posed problems. Finally, multiple classifier systems and classifier ensembles have been proved successful in several hyperspectral image classification applications (Waske and Benediktsson 2007; Chan and Paelinckx 2008; Ceamanos et al. 2009).

In any case, a new learning trend has recently emerged in hyperspectral imagery analysis. It takes advantage of semi-supervised or transductive learning and also integrates spatial information to reduce the risk of overfitting possibly due to Hughes' phenomenon. In particular, transduction, possibly combined with spatial information synthesized through local neighborhoods, aims at iteratively labeling samples also from the test set. This fact increases the number of labeled samples, reducing the impact of the overfitting. For example, Bruzzone et al. (2006) have designed transductive SVMs for hyperspectral image classification. The algorithm is iterative and gradually searches the optimal discriminant hyperplane in the feature space. It uses a transductive process that incorporates unlabeled samples in the training phase. Ratle et al. (2010) have proposed a semi-supervised classification algorithm based on neural networks. The algorithm consists of adding a flexible embedding regularizer to the loss function used for training neural networks. Similarly, a plethora of spectral-spatial classifiers was defined in the hyperspectral imaging literature. Several algorithms, based on Markov random fields (MRFs), have been quite successful in hyperspectral imaging (Li et al. 2011, 2012, 2013a; Tarabalka et al. 2010b; Khodadadzadeh et al. 2014b). MRFs exploit general properties to efficiently describe dependencies between random variables. In this way, they can arrange the spatial dependency between pixels or regions of the image. In particular, MRF-based algorithms encourage segmentation and foster solutions in which adjacent pixels are likely to belong to the same class. In addition, MRFs are a generalization of an energy model (i.e. Ising model Geman and Geman 1984), so a stable solution (the correct classification) typically corresponds to the minimization of the image energy function.

On the other hand, Plaza et al. (2009), Bovolo et al. (2006) and Fauvel et al. (2012) have defined several spectral-spatial kernels, which model the inter-pixel relations as the mean of the pixel spectral signatures from a pixel's neighborhood system. This is based on the idea that the spectral signature of each pixel may be represented by some linear combinations of its neighboring pixels (spectral spatial correlation). Spatial information is directly included in the training process as a new constraint for the optimization problem. Tarabalka et al. (2010a) have investigated the use of a watershed transformation, in order to determine a segmentation map of the image. They defined a two-stepped classification process according to which the spectral-based SVM classification is followed by majority voting within the watershed regions. Huang and He (2012) have investigated the idea of learning SVMs from

the spectral profile, as well as from two types of spatial profiles of the imagery data. However, they have extracted spatial information based on the spectral information. Therefore, they construct spatial features, which do not change during the learning phase. Finally, there are studies which exploit spatial information in semi-supervised learning algorithms. Camps-Valls et al. (2007) have presented a semi-supervised graph-based method, designed to exploit both spectral and spatial information in the images through composite kernels. Wang et al. (2014) have recently proposed a spectral-spatial label propagation for the semi-supervised classification of hyperspectral imagery.

5 Spectral and spatio-relational classifier ensemble

This section is devoted to the description of the algorithm S^2TEC.

5.1 The transductive classification problem

Let $\mathcal{D}$ be a hyperspectral imagery dataset whose pixels are sparsely labeled according to an unknown target function C and are all described according to the spectral feature vector model $\mathbf{S}$ (details in Sect. 3). The transductive classification problem inputs both a labeled set $\mathcal{L} \subset \mathcal{D}$ and the projection of the unlabeled set $\mathcal{U} = \mathcal{D} - \mathcal{L}$ on the descriptive space $\mathbf{S}$, in order to output predictions of the class values of instances in the unlabeled set $\mathcal{U}$, which are as accurate as possible. The learner receives full information (including labels) on the instances in $\mathcal{L}$ and partial information (without labels) on the instances in $\mathcal{U}$ and is required to predict the class values only of the examples in $\mathcal{U}$.

5.2 Spectral and relational features

The vector of spectral features is input as part of the hyperspectral imagery dataset and used to populate the spectral data profile $\mathbf{S}$. The relational features are constructed by coupling imagery pixels with square-shaped spatial neighborhoods and synthesizing information on the spatial variation of labels over the imagery pixels of each neighborhood (see details in Sect. 3). Relational features are then used to populate the spatio-relational data profiles of the dataset. Two application-specific spatio-relational profiles are constructed (see details in Sect. 3), namely the frequency data profile $\mathbf{F}$ and the morphological data profile $\mathbf{M}$. The frequency data profile is the vector of the spatio-relational features which are constructed, for every pixel p, by applying the frequency operator with every class label and every spatial neighborhood coupled with p. The spatial morphological data profile is the vector of the spatio-relational features which are constructed, for every pixel p, by applying the morphological operators (dilation, erosion, opening and closing), with every class label and every spatial neighborhood coupled with p.

For every pixel p, we construct a set of neighborhoods $\mathcal{N}(p, R)$ with growing sides ($R \in RSet$), in order to capture the space-variant label distribution. In fact, the image labeling is usually modeled as non-stationary in the spatial domain (Isaaks and Srivastava 1990). This idea of using a range of sizes follows the point of view of Plaza et al. (2009) and Guccione et al. (2015), who showed how a range of different spatial neighborhoods must be used as structuring elements, in order to capture the shape or size of all elements present in an image. In addition, while spectral features do not change during learning, spatio-relational features can be updated every time predicted labels are changed under the influence of the transductive learning.

Algorithm 1 Iterative Convergence Learning

Require: $\mathcal{D}$: imagery data pixels as they are split in $\mathcal{L}$ (labeled set) and $\mathcal{U}$ (unlabeled set)
Require: XY: spatial coordinates of imagery pixels of $\mathcal{D}$
Require: $\mathbf{S}$: vector of spectral features
Require: $radiusSet$: set of radius values used to construct spatial neighborhoods
Require: C: map of ground-truth labels of $\mathcal{L}$
Ensure: C: map of labels of $\mathcal{D}$ (ground-truths for $\mathcal{L}$ and predicted labels for $\mathcal{U}$) {initialization phase}
1: $classifierS \leftarrow$ classifier($\mathcal{L}, \mathbf{S}, C$) {Supervised learning}
2: $C \leftarrow C \cup$ predictions($classifierS, \mathbf{S}, \mathcal{U}$) {Classifying unlabeled pixels}
3: $\mathcal{N} \leftarrow spatialNeighborhood(\mathcal{D}, XY, RSet)$ {Constructing the spatial neighborhoods of the imagery pixels of $\mathcal{D}$ with the radius values specified in $RSet$}
4: $\mathbf{F} \leftarrow$computeFrequencyFeatures($\mathcal{D}, C, \mathbf{F}$) {Constructing frequency features}
5: $\mathbf{M} \leftarrow$computeMorphologicalFeatures($\mathcal{D}, C, \mathbf{M}$) {Constructing morphological features} {loop phase}
6: **repeat**
7: $classifierF \leftarrow$classifier($\mathcal{L}, \mathbf{F}, C$)
8: $classifierM \leftarrow$classifier($\mathcal{L}, \mathbf{M}, C$)
9: $classifierS \leftarrow$classifier($\mathcal{L}, \mathbf{S}, C$)
10: $ensemble \leftarrow$ensemble($classifierF, classifierM, classifierS$){Defining the ensemble system of frequency-relational, morphological-relational and spectral classifiers}
11: **for** $p \in \mathcal{U}$ **do**
12: **if** consensus($ensemble, p$) **then**
13: assign the consensus class to p in C {Using the consensus pattern of the current ensemble system to select and label definitely pixels moved from $\mathcal{U}$ to $\mathcal{L}$}
14: $\mathcal{L} \leftarrow \mathcal{L} \cup \{p\}$
15: $\mathcal{U} \leftarrow \mathcal{U} - \{p\}$
16: **end if**
17: **end for**
18: $\mathbf{F} \leftarrow$computeFrequencyFeatures($\mathcal{D}, C, \mathbf{F}$) {Updating frequency features}
19: $\mathbf{M} \leftarrow$computeMorphologicalFeatures($\mathcal{D}, C, \mathbf{M}$) {Updating morphological features}
20: **until** $\mathcal{U} = \oslash$ OR number of pixel transferred from $\mathcal{U}$ to $\mathcal{L}$ is less than $MinTransfer$

5.3 Iterative convergence learning

A top-level description of the iterative convergence learning is given in Algorithm 1. The algorithm comprises an initialization phase and an iterative phase. Both phases concern an ensemble system that comprises three classifiers. These classifiers are learned iteratively from the spectral profile $\mathbf{S}$ ($classifierS$), the spatial (-frequency) profile $\mathbf{F}$ ($classifierF$) and the spatial (-morphological) profile $\mathbf{M}$ ($classifierM$) of the imagery data, respectively. The unlabeled part of the image is initially classified according to the classifier ($classifierS$) learned from the original labeled part $\mathcal{L}$ of the image with the features in the spectral profile $\mathbf{S}$. Both real labels and predicted labels are used to initialize the features of the two spatial profiles $\mathbf{F}$ and $\mathbf{M}$, respectively. At each iteration, pixels of the unlabeled part $\mathcal{U}$, which are identically predicted by the majority of classifiers in the present ensemble,[1] are definitively classified with the majority class of the ensemble and transferred to the labeled part $\mathcal{L}$ of the dataset. Spatio-relational features of the two spatial profiles are updated accordingly. A detailed description of the two phases is reported in the following.

The initialization phase (Algorithm 1, lines 1–5) consists of three steps:

1. The pixels of the unlabeled set $\mathcal{U}$ are initially labeled (Algorithm 1, lines 1–2), by using the spectral classifier learned from the labeled set $\mathcal{L}$, as it is originally described in the image in the space of spectral features $\mathbf{S}$.

[1] In this study, the ensemble comprises three classifiers, thus a pixel is considered a consensus pattern if it is labeled identically by at least two out of three classifiers.

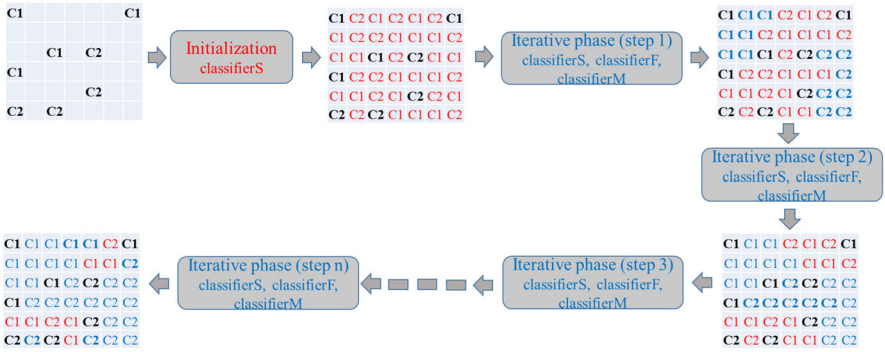

Fig. 4 Label assignment: starting from a sparsely labeled image, unknown labels are predicted by the spectral classifier ($classifierS$) during the initialization phase (*red colored labels*). During the iterative phase, labels that are reliable (*blue colored labels*) predicted by the iteratively constructed ensemble ($classifierS$, $classifierF$ and $classifierM$) definitively replace labels predicted in the initialization phase. At the end of the learning process, the image is fully labeled (Color figure online)

2. The spatial neighborhood structure of the imagery data $\mathcal{D}$ is constructed (Algorithm 1, line 3). For each pixel, a set of square-shaped spatial neighborhoods is built and associated to the pixel. Each neighborhood is constructed with a specified radius. The set of radius values ($radiusSet$) is a user-defined parameter.
3. The spatio-relational features are constructed to synthesize the information on the spatial variation of the class labels over spatial neighborhoods (Algorithm 1, lines 4–5). To initialize these features, the real labels are associated with the pixels of the labeled set $\mathcal{L}$, while the labels predicted by the initial spectral classifier (see step 1 of this initialization phase) are associated with the pixels of the unlabeled set $\mathcal{U}$ (see Fig. 4). The constructed features are used to populate the frequency profile (**F**) and the morphology profile (**M**) of both $\mathcal{L}$ and $\mathcal{U}$.

The iterative phase is produced by the main loop (Algorithm 1, lines 6–20) and consists of three steps:

1. The ensemble of the multiple classifiers (Algorithm 1, line 10) is learned from the currently labeled set $\mathcal{L}$. This ensemble is composed of: (1) $classifierF$ (Algorithm 1, line 7), learned from $\mathcal{L}$ as it is spanned on the vector of frequency-defined relational features **F**; (2) $classifierM$ (Algorithm 1, line 8), learned from $\mathcal{L}$ as it is spanned on the vector of morphological-defined relational features **F**; $classifierS$ (Algorithm 1, line 9), learned from $\mathcal{L}$ as it is spanned on the vector of spectral features **S**. The ensemble is used to predict labels of pixels of the currently unlabeled set $\mathcal{U}$.
2. For each pixel in the currently unlabeled set $\mathcal{U}$, each classifier in the ensemble is used to predict its label. We consider consensus patterns, pixels which are identically labeled by the majority of classifiers of the ensemble. A consensus pixel is finally labeled with the consensus label (Algorithm 1, line 13) determined by the ensemble and definitely moved from $\mathcal{U}$ to $\mathcal{L}$ (Algorithm 1, lines 14–15, Fig. 4). On the other hand, pixels, which are left in $\mathcal{U}$, stay still associated with the labels predicted by the spectral classifier learned during the initialization phase (see Fig. 4).
3. The spatio-relational features of both the frequency profile and the morphological profile are updated according to the new consensus labels, which have been finally updated in C (Algorithm 1, lines 18–19).

This iterative inference stops when the unlabeled set is empty or the number of pixels definitely transferred from unlabeled set $\mathcal{U}$ to labeled set $\mathcal{L}$ is less than a threshold denoted as $MinTransfer$. By default, $MinTransfer = 10$. The iterative inference procedure is guaranteed to converge as eventually one of the stopping criteria will be satisfied. If each iteration transfers more than $MinTransfer$ pixels from $\mathcal{U}$ to $\mathcal{L}$ (Algorithm 1, lines 14–15), then $\mathcal{U} = \emptyset$ and the first condition is satisfied. Otherwise, if the number of pixels, transferred from $\mathcal{U}$ to $\mathcal{L}$, at the present iteration, is less than $MinTransfer$, the second stopping condition is satisfied. In both cases, the imagery data are all fully labeled during the learning process (see Fig. 4). In fact, if there are still pixels, which have never been transferred during the iterative phase, they stay assigned to the classes decided by the spectral classifier learned during the initialization phase.

6 Learning complexity

For this analysis, we assume that $N^{\mathcal{L}}$ denotes the number of labeled pixels ($N^{\mathcal{L}} = |\mathcal{L}|$), while $N^{\mathcal{U}}$ denotes the number of unlabeled pixels ($N^{\mathcal{U}} = |\mathcal{U}|$) of the imagery data $\mathcal{D}$, so that $N = N^{\mathcal{L}} + N^{\mathcal{U}}$. As the pixels are transferred from the labeled set to the unlabeled set during the iterative convergence learning process, $N^{\mathcal{L}(i)}$ ($N^{\mathcal{U}(i)}$) is used to denote the number of pixels in the labeled set (unlabeled set) at the ith iteration of the iterative process. k is the number of distinct classes ($k = |C|$). r is the number of square-shaped neighborhood objects constructed for each pixel of $\mathcal{D}$ ($r = |RSet|$). R_{max} is the radius of the largest spatial neighborhoods constructed per pixel ($R_{max} = \max_{R \in RSet} R$), so that $((2R_{max} + 1)^2 = 4R_{max}^2 + 4R_{max} + 1$ is the maximum number of pixels grouped per square neighborhood. $nIter$ is the number of iterations performed with the iterative convergence learning algorithm. $\Lambda(|Data|, |FeatureSpace|)$ denotes the cost of learning a supervised classifier[2] from a training set $Data$ as it is spanned on a feature space $FeatureSpace$.

The computational complexity of S^{2}TEC depends on the cost of (1) classifying unlabeled pixels according to the supervised classifier learned from $\mathcal{L}$ on $\mathbf{S}$; (2) constructing the neighborhood structure with the radius values collected in $RSet$; (3) constructing the relational features according to both the frequency profile and the morphological profile; (4) constructing the ensemble of classifiers; (5) identifying pixels of $\mathcal{U}$ which are identically labeled by the majority of classifiers in the ensemble and moving these pixels from $\mathcal{U}$ to $\mathcal{L}$ with their consensus labels; (6) updating relational features according to the new consensus labels. Steps (1), (2) and (3) are part of the initialization phase, while steps (4), (5) and (6) are part of the iterative convergence learning phase, thus they occur $nIter$ times. The time cost of learning the spectral classifier from the initial labeled set $\mathcal{L}$ is $O(\Lambda(N^{\mathcal{L}}, m))$. The time cost for constructing the neighborhood structure of the imagery data is $N \cdot (4R_{max}^2 + 4R_{max} + 1)$, that is, $O(NR_{max}^2)$. The time cost of constructing the relational features by using both the frequency operator and the morphological operators is $5 \cdot k \cdot r \cdot (4R_{max}^2 + 4R_{max} + 1) \cdot N$, that is, $O(krR_{max}^2 N)$. Therefore, the time complexity of the initialization phase is $O(\Lambda(N^{\mathcal{L}}, m) + NR_{max}^2 + krR_{max}^2 N)$, that is, $O(\Lambda(N^{\mathcal{L}}, m) + (kr + 1)R_{max}^2 N)$. At the iteration i of the iterative convergence learning algorithm, the time cost of constructing the ensemble of classifiers is $O(\Lambda(N^{\mathcal{L}(i)}, \underbrace{kr}_{|\mathbf{F}|}) + \Lambda(N^{\mathcal{L}(i)}, \underbrace{4kr}_{|\mathbf{M}|}) + \Lambda(N^{\mathcal{L}(i)}, \underbrace{m}_{|\mathbf{S}|}))$, that is, $O(\Lambda(N^{\mathcal{L}(i)}, F))$, with $F = \max\{kr, 4kr, m\}$. The time cost of determining and transferring consensus pixels from $\mathcal{U}(i)$ to $\mathcal{L}(i)$ is $O(N^{\mathcal{U}(i)})$, while the time cost of updating the relational features is

[2] This cost depends on the algorithm selected as the base classifier.

$O(kr R^2_{max} N)$). Therefore, the time complexity of the iterative convergence learning phase is $\sum_{i=1}^{nIter} (\Lambda(N^{\mathcal{L}(i)}, F) + N^{\mathcal{U}(i)} + kr R^2_{max} N)$, that is, $O(\sum_{i=1}^{nIter} (\Lambda(N^{\mathcal{L}(i)}, F) + kr R^2_{max} N))$ as $N^{\mathcal{U}(i)} \leq N$.

7 Experimental evaluation and discussion

S^{2}TEC, whose implementation is publicly available,[3] is written in Java. It integrates the inductive Support Vector Machine (SVM) [4] (Cortes and Vapnik 1995) as a base classifier of the transductive ensemble system. This choice is motivated by several studies reported in the literature (e.g. Plaza et al. 2009; Fauvel et al. 2013; Chen et al. 2014), which show that inductive SVMs are applied to hyperspectral image classification with great success, outperforming several other inductive classifiers. As the hyperspectral classification problem is a multi-class problem, we learn multi-class SVMs with the "one-against-all" strategy. SVMs are learned with the Gaussian kernel rule, while parameters are optimally selected according to a grid-search method and a three-fold cross validation of the labeled set. S^{2}TEC is evaluated on three benchmark hyperspectral images, in order to seek answers to the following questions:

1. Is the defined transductive schema more accurate than the base inductive learner and the traditional transductive approaches that do not use collective inference (see Sect. 7.2)?
2. How does the performance (accuracy, learning time, memory usage) of the classification change by varying the number of performed iterations (see Sect. 7.3)?
3. Is the classification robust to change in the size of the initial labeled set and the size of the spatial neighborhoods (see Sect. 7.3)?
4. How do the individual components of the transductive schema affect its overall accuracy (see Sect. 7.4)?
5. How does the schema's accuracy compare to the state-of-the-art hyperspectral imaging classifiers (see Sect. 7.5)?

The experiments are run on a Xeon 2.4 Ghz 2 core processor.

7.1 Hyperspectral image datasets

Three real data sets, namely Indian Pines, Pavia University and Salinas Valley (http://www.grss-ieee.org/community/technical-committees/data-fusion/data-sets/), are used in this experimental study. In detail, *AVIRIS Indian Pines* was obtained by the airborne visible infrared imaging spectrometer (AVIRIS) sensor over the Indian Pines region in Northwestern Indiana in 1992. The image contains 220 spectral bands, but 20 spectral bands have been removed due to the noise and water absorption phenomena. The spatial resolution is of 20 m and the spatial size is 145×145 pixels, which are classified into 16 mutually exclusive classes (see Fig. 5a). This data set represents a very challenging land-cover classification scenario, in which the primary crops of the area (mainly corn and soybeans) were very early in their growth cycle, with only about 5 % canopy cover (Plaza et al. 2009). Discriminating among the major crops under these circumstances can be a very difficult task. This scenario is also made more complex by the imbalanced number of available labeled pixels per class. *ROSIS Pavia University* was obtained by the reflective optics system imaging spectrometer (ROSIS)

[3] http://www.di.uniba.it/~appice/software/S2TEC/.

[4] We use the Java implementation of SVM included in the WEKA toolkit (Witten and Frank 2005).

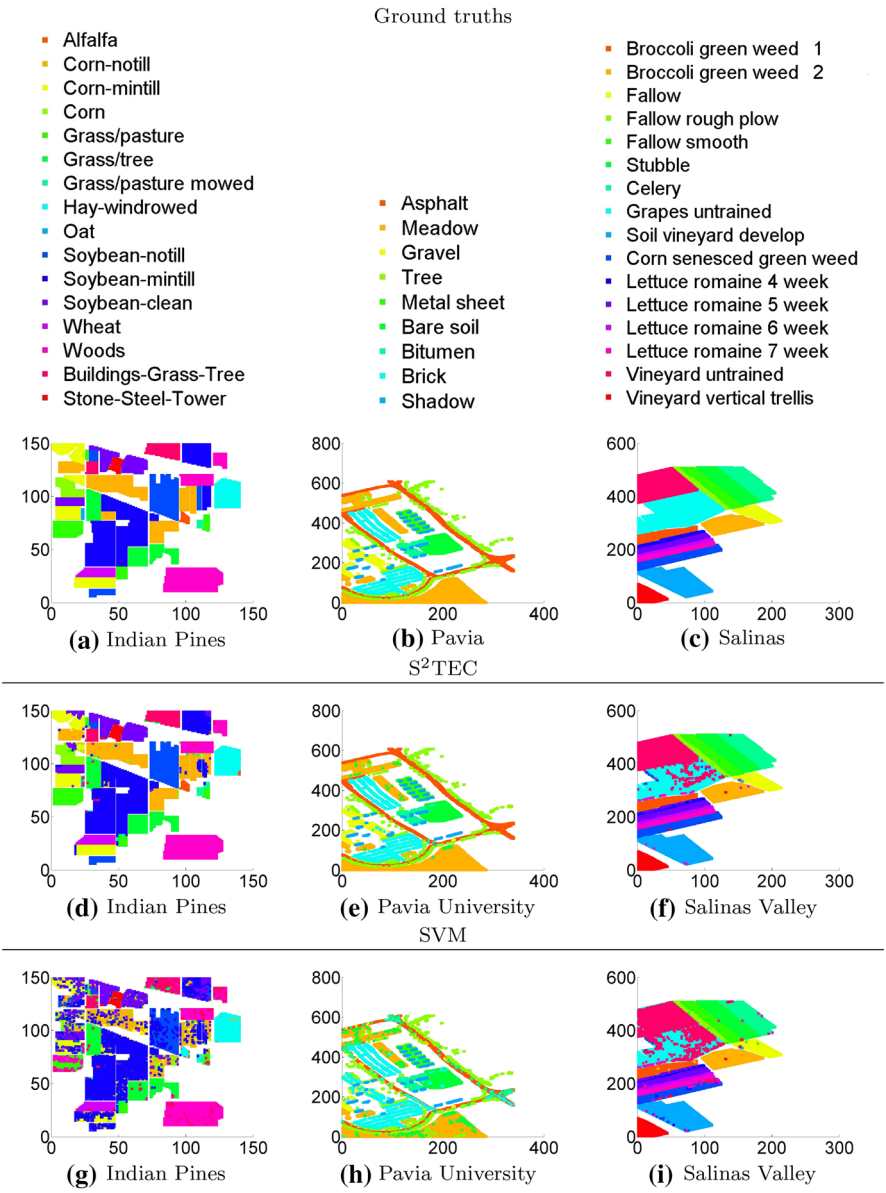

Fig. 5 The AVIRIS data of Indian Pines and Salinas Valley, the ROSIS data of Pavia University: ground truths (**a**–**c**), as well as classification maps generated by both S^2TEC (**d**–**f**) and SVMs (**g**–**i**)

sensor during a flight campaign over the Engineering School at the University of Pavia, in 2003. Water absorption bands were removed, and the original 115 bands were reduced to 103 bands. It has a spatial resolution of 1.3 m. The image has a spatial size of 610 × 340 pixels, which are classified into 9 classes (see Fig. 5b). Finally, *AVIRIS Salinas Valley* was collected by AVIRIS over Salinas Valley, Southern California, in 1998. It has a spatial resolution of

3.7 m. The area contains a spatial size of 512 × 217 pixels and 206 spectral bands. The 20 water absorption bands are discarded. Pixels are classified into 16 classes (see Fig. 5c).

These data sets are selected for the following reasons: (1) They have a very high spatial resolution. (2) They contain rich spectral information (100–200 bands) and a high number of classes (9–16 classes). (3) They correspond to different scenarios. (4) Ground truths are available for these data.[5] Additionally, they are considered in the majority of recent, relevant works on hyperspectral image classification (e.g Melgani and Bruzzone 2004; Plaza et al. 2009; Li et al. 2011, 2012, 2013a; Tarabalka et al. 2010b; Guccione et al. 2015). In fact, although the most advanced sensor, namely the AISA system, is currently able to capture up to 488 bands in the interval 400–970 nm, for an image of 512 or 1024 pixels (details at http://www.spectralcameras.com/aisa), the use of 100–200 bands in the optical-near infrared electro magnetic interval with a resolution of a few meters, still represents the state-of-the-art in the sensing literature.

7.2 Comparative analysis

For this study, we consider all the datasets described above.

7.2.1 Experimental set-up

We run S^{2}TEC by setting the percentage of pixels labeled in the image equal to 5 %,[6] and the size of spatial neighborhoods equal to 5–10, 15 and 20, respectively. We compare S^{2}TEC to the inductive SVM, to the Fast Linear tranductive SVM (SVMLin) (Sindhwani and Keerthi 2006) and to the spectral graph transducer (SGT) (Joachims 2003) (see a description in Sect. 4.3). SVMLin is run with the optimal configuration of parameters identified by Sindhwani and Keerthi (2006). SGT is run with the optimal configuration of parameters identified by Joachims (2003) and with the number of neighbors k ranging between 25, 50 and 100. The inductive SVM, as well as the transductive SVMLin and SGT are all defined for binary classification problems. We use the "one-against-all" strategy, already adopted in S^{2}TEC, in order to adapt these binary transductive classifiers to the multi-class problem.

We evaluate the accuracy of the algorithms in terms of overall accuracy (OA), average accuracy (AA) and Cohen's kappa coefficient (κ) (Richards 1993). In addition, we analyze the F-1 score of predictions performed for each class. For each dataset, the labeled pixels are randomly selected from the available ground truth of the image by using the stratified random sampling without replacement;[7] the remaining pixels are used as the unlabeled part of the learning process. Five partitioning trials between labeled and unlabeled sets are generated; metrics are averaged on these trials.

We use the non-parametric Wilcoxon two-sample paired signed rank test (Orkin and Drogin 1990), in order to compare the accuracy of the considered algorithms. To perform the test, we assume that the experimental results of the two algorithms compared are independent pairs. We test the null hypothesis H_0: "no difference in distributions" against the two-sided

[5] Although, the data acquisition can be a relatively easy process, the generation of a reliable ground truth is a very expensive process.

[6] This setting is usual in the hyperspectral image classification (Plaza et al. 2009).

[7] This is slightly different from the traditional k-fold cross validation, where the data are "partitioned" into k equally sized folds; the learner is trained on a $k-1$ data folds and tested on the hold-out data fold (or vice-versa). The experimental methodology based on the stratified random sampling is commonly used in the hyperspectral image classification literature, where it is used to simulate the case of sparsely labeled datasets.

Table 1 The accuracy performance: OA, AA and κ of S^2TEC (size = 5, 10, 15, 20), SVM, SVMLin, SGT (k = 25), SGT (k = 50) and SGT (k = 100)

Algorithm	Indian Pines			Pavia University			Salinas Valley		
	OA	AA	κ	OA	AA	κ	OA	AA	κ
S^2TEC	**.945**	**.876**	**.937**	**.992**	**.988**	**.989**	**.980**	**.988**	**.977**
SVM	.739	.626	.700	.931	.912	.909	.925	.957	.916
SVMLin	.605	.517	.553	.804	.747	.737	.891	.943	.878
SGT (25)	.596	.515	.531	.805	.749	.731	.859	.923	.843
SGT (50)	.598	.533	.533	.804	.745	.729	.855	.917	.839
SGT (100)	.604	.537	.541	.800	.739	.723	.853	.914	.836

Results are collected on five trials with 5 % of labeled pixels. The highest accuracy is in bold

alternative H_1: "there is a difference in distributions". In all experiments reported in this empirical study, the significance level used in the test is set at 0.05.

7.2.2 Results and discussion

Table 1 shows the average accuracy metrics (OA, AA and κ) of the compared algorithms. They show that S^2TEC is more accurate than the inductive SVM learner and the transductive competitors in all datasets. In addition, according to a pairwise Wilcoxon signed rank test, all differences between S^2TEC and other algorithms are statistically significant (with $p \leq .05$). Inductive SVMs, generally, perform much better than transductive SVMLins and SGTs. To interpret these results, let us consider that all these competitors are learned by using only the spectral information, without accounting for the spatial information. Additionally, they are actually all defined as binary classifiers, so we use the "one-against-all" strategy, in order to adapt binary inductive/transductive classifiers to the multi-class problem formulation. However, S^2TEC applies the "one-against-all" strategy every time a multi-class classifier has to be learned as a set of binary classifiers during the transduction. On the contrary, SVMLins and SGTs complete a separate transductive learning process with every binary classifier and apply the "one-against-all" strategy to the binary classifiers finally constructed via the transduction. Based on these premises, we can conclude that the transductive process performed with the binary classifiers without accounting for the spatial information (which S^2TEC does, however), may even lead to less accuracy. To support this conclusion, let us consider that Sect. 7.4 contains the results of two transductive SVMs, namely T + SVM and selfSVM (see Table 5) learned for the Indian Pines data. They perform transductive inference of the spectral data by applying the "one-against-all" strategy to every set of binary classifiers learned during the transduction. We note that, also in this case, transductive learning, performed without benefiting from collective inference and ensemble learning, does not outperform the inductive learner, although the observed accuracy gap is smaller.

Table 2 shows the per-class F-1 score for each classifier. These results show that S^2TEC exhibits high F-1 score ($\geq$ 0.9 per Indian Pines) per class, except for the classes Alfalfa and Oat in Indian Pines. Both these classes are minority classes (see column 2 of Table 2) in this image. Competitors also produce predictions with poor accuracy for the same classes. On the other hand, this analysis per class confirms that both the inductive base learner and the transductive competitor learners are outperformed by our transductive one in the detection of

almost all classes. The only exceptions are the classes "Broccoli green weed 1" and "Fallow" of Salinas Valley where the F-1 score of SVMLin is slightly higher than the F-1 score of S^2TEC (0.996 vs 0.995 for "Broccoli green weed 1" and 0.994 vs 0.991 for "Fallow"), as well as the class "Fallow rough plow" for Salinas Valley where the F-1 score of SGT (k=100) is slightly higher than the F-1 score of S^2TEC (.994 vs .991).

Table 2 Class by class analysis: percentage of pixels per class (column 2) and F-1 score of S^2TEC (column 3, size=5,10,15,20) against SVM (column 4), SVMLin (column 5), SGT (k=25, column 6), SGT (k=50, column 7) and SGT (k=100, column 8)

Class	%	S^2TEC	SVM	SVMLin	SGT (25)	SGT (50)	SGT (100)
Indian Pines							
Alfalfa	0.45	**.612**	.254	.198	.120	.256	.259
Corn-N	13.93	**.913**	.692	.603	.427	.408	.430
Corn-M	8.10	**.912**	.602	.383	.351	.377	.396
Corn	2.31	**.954**	.370	.202	.228	.217	.183
Grass/pasture	4.71	**.912**	.799	.784	.722	.743	.710
Grass/tree	7.12	**.972**	.879	.860	780	.786	.795
Grass/pasture M	0.27	**.626**	.419	.163	.681	.629	.616
Hay-W	4.66	**.983**	.920	.837	.921	.917	.919
Oat	0.20	**.156**	.092	.040	.008	.064	.049
Soybean-N	9.48	**.922**	.669	.493	.428	.448	.493
Soybeans-M	23.95	**.956**	.744	.616	.612	.608	.611
Soybeans-C	5.79	**.940**	.599	.403	.197	.186	.216
Wheat	2.0	**.987**	.912	.709	.822	.821	.826
Woods	12.34	**.996**	.899	.880	.901	.907	.892
Buildings-GT	3.77	**.967**	.521	.443	.209	.228	.203
Stone-ST	0.91	**.987**	.875	.213	.918	.917	.959
Pavia University							
Asphalt	15.5	**.994**	.923	.766	.827	.825	.822
Meadow	43.6	**.996**	.867	.906	.886	.886	.883
Gravel	4.91	**.973**	.795	.607	.605	.598	.598
Tree	7.16	**.984**	.947	.843	.845	.846	.838
Metal sheet	3.14	**.999**	.991	.992	.989	.988	.985
Bare soil	11.76	**.993**	.895	.653	.362	.356	.337
Bitumen	3.11	**.992**	.842	.484	.716	.712	.703
Brick	8.61	**.978**	.865	.609	.701	.698	.696
Shadow	2.21	**.994**	.998	.959	.983	.978	.975
Salinas Valley							
Broccoli w1	3.71	.995	.994	**.996**	.995	.994	.991
Broccoli w2	6.88	**.998**	.995	.997	.992	.990	.987
Fallow	3.65	.991	.980	**.994**	.940	.935	.936
Fallow RP	2.58	.991	.989	.988	.992	.993	**.994**
Fallow S	4.95	**.993**	.987	.988	.952	.950	.952
Stubble	7.31	**.999**	.997	.998	.994	.994	.979

Table 2 continued

Class	%	S^2TEC	SVM	SVMLin	SGT (25)	SGT (50)	SGT (100)
Celery	6.61	**.998**	.996	.991	.988	.988	.988
Grapes U	20.82	**.959**	.845	.761	.720	.719	.720
Soil V	11.46	**.997**	.992	.993	.978	.973	.972
Corn SW	6.06	**.988**	.956	.933	.866	.849	.837
Lettuce 4w	1.97	**.990**	.959	.971	.942	.915	.907
Lettuce 5w	3.56	**.995**	.985	.981	.976	.969	.966
Lettuce 6w	1.69	**.993**	.966	.986	.953	.956	.957
Lettuce 7 week	1.98	**.976**	.939	.919	.944	.945	.946
Vineyard U	3.34	**.940**	.947	.634	.529	.527	.523
Vineyard VT	13.43	**.992**	.986	.988	.973	.969	.963

Results are collected on five trials with 5 % of labeled pixels. The highest accuracy is in bold

In general, the analysis of accuracies highlights that, although transductive inference has been specifically defined to deal with scarce labels (Vapnik 1998), it can be unsatisfactory for gaining accuracy in the data imagery scenario. On the contrary, coupling transductive inference with collective inference and ensemble learning can really improve accuracy in the identification of the pixels belonging to the different classes. The classification of pixels belonging to minority classes remains a problem, that requires further investigation.

Finally, we illustrate some considerations concerning the spatial distribution of misclassified pixels. We compare the classification maps built by both S^2TEC (Fig. 5d–f) and its inductive SVM counterpart (Fig. 5g–i). These maps show that S^2TEC takes advantage of the presented spectral-relational methodology. It gains accuracy when discriminating objects of interest on the map, by reducing visibly the salt-and-pepper distribution of pixels misclassified by the base SVM learner. For example, we can note that S^2TEC diminishes visibly the number of pixels of the class "vineyard untrained" that the base SVM learner wrongly detects as part of the object "grapes untrained" in the Salinas Valley (see Fig. 5f–i). This result is in agreement with the F-1 scores reported in Table 2, which show that both the inductive SVM learner and the transductive competitor learners (SVMLin and SGT) perform poorly when labeling pixels of the classes "grapes untrained" and "vineyard untrained". On the other hand, the pixels misclassified by S^2TEC generally fall into the margin bound between homogeneously, well-classified zones, that is, where enhancing the classification accuracy with the spatial separability among classes is a more difficult task.

7.3 Sensitivity analysis

For this analysis, we consider the Indian Pines dataset that, according to considerations formulated by Plaza et al. (2009), is a very challenging classification problem (see details in Sect. 7.1).

7.3.1 Experimental set-up

We perform a sensitivity analysis of the performance of S^2TEC along the number of iterations, the size of the initial labeled set and the size of the spatial neighborhoods. Firstly, we consider the labeled sets sampled with the percentage 5 % and we monitor both the performance of

S^2TEC along the dimension of the number of performed iterations. For this analysis, S^2TEC constructs spatial neighborhoods with sizes growing from 5 to 10, 15 and 20. Secondly, we vary the percentage of pixels which are labeled in the image among 3, 5 and 10 %, while we run S^2TEC by constructing spatial neighborhoods with sizes growing from 5, 10, 15 to 20. Finally, we construct spatial neighborhoods with sizes: 5–10–15, 5–10–15–20 and 5–10–15–20–25, while we run S^2TEC with the initial labeled set sampled with the labeling percentage equal to 5 %.

We evaluate the performance of the compared algorithms in terms of overall accuracy, average accuracy and Cohen's kappa coefficient. We also analyze the learning time (in seconds), the maximum number of iterations performed to complete the task and the peak of memory usage (in MegaBytes) during learning. As five partitioning trials between labeled and unlabeled sets are generated, the results are always averaged across these trials.

7.3.2 Results and discussion

Number of iterations We start by studying the performance of S^2TEC along the dimension of the number of performed iterations. This analysis is performed by considering the same samples of the comparative study presented in Sect. 7.2. The accuracy metrics (OA, AA and κ), the computation time (in s) and the memory usage (in MB) are the plots in Fig. 6a–c. We observe that accuracy is gained as new iterations are performed. This confirms the effectiveness of the iterative learning approach. We can also observe that the highest accuracy gain is obtained in the initial iterations of the learning process, which are also those showing the highest increment in the usage of the time-memory resources consumed by the process.

Size of the initial labeled set and size of the spatial neighborhoods Then we proceed by studying the performance of S^2TEC as a function of the size of the initial labeled set, as well as a function of the number and size of the spatial neighborhoods. The average and the SD of the accuracy metrics, the memory usage peak (MB) and the number of performed iterations are reported in Table 3. We observe that, as expected, the classifier gains accuracy by augmenting the number of pixels in the originally labeled set (rows 2–4, Table 3). In addition, the SD of the accuracy metrics decreases as the number of the initially labeled pixels increases in the experiment. This result is in agreement with the literature (Li et al. 2013b). On the other hand, the classifier gains accuracy by enlarging the size of the spatial neighborhoods (rows

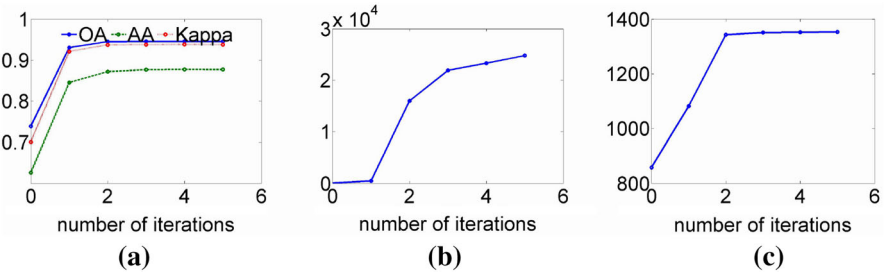

Fig. 6 Sensitivity study (Indian Pines): the accuracy (*Y* axis, **a**), the computation time (*Y* axis, **b**), and the memory usage peak (*Y* axis, **c**) are plotted along the dimension of the number of performed iterations (*X* axis). S^2TEC is run by considering the labeled sets generated by sampling 5 % of pixels and by constructing the relational features over the spatial neighbourhoods with size growing from 5, 10, 15 to 20. Results generated on five trials are averaged

Table 3 Sensitivity study (Indian Pines): the accuracy metrics, the memory usage peak and the number of performed iterations are collected on five trials, the average (SD) of the measures is computed on these trials

Labeled (%)	OA	AA	κ	Memory (MB)	nIterations
3	.928 (.010)	.804 (.039)	.918 (.001)	1352.2 (.894)	5.2 (0.399)
5*	.945 (.010)	.876 (.036)	.937 (.011)	1352.4 (1.788)	5.4 (0.799)
10	.975 (.006)	.927 (.014)	.971 (.007)	1352.8 (1.673)	4.4 (.489)
Neighborhood sizes	OA	AA	κ	Memory (MB)	nIterations
5, 10, 15	.936 (.011)	.874 (.031)	.927 (.012)	931.8 (1.673)	5.2 (0.399)
5, 10, 15, 20*	.945 (.036)	.876 (.046)	.937 (.011)	1352.4 (1.788)	5.4 (0.799)
5, 10, 15, 20, 25	.953 (.014)	.884 (.049)	.946 (.016)	1959.6 (17.469)	5.2 (0.399)

S^2TEC is run by setting the neighborhood sizes equal to 5–10–15–20 when varying the labeled set percentage between 3, 5 and 10 % and by setting the labeled set percentage equal to 5 % when varying the neighborhood sizes between 5–10–15, 5–10–15–20 and 5–10–15–20–25. "*" is used to mark the parameter setup of the specific configuration used in the comparative study

6–8, Table 3), as in this way we increase the chances of building relational features that better match the spatial variation of classes, even when classes vary over space with different density and texture. Additionally, the SD of these accuracy metrics, generally, assumes low values. The learning process is completed in five iterations on average regardless of both the number and the size of the spatial neighborhoods used to construct the spatio-relational features, as well as of the size of the initial labeled set. Finally, the memory usage is mainly influenced by the size and the number of neighborhoods constructed. The higher the number of neighbors processed through collective inference, the greater the amount of memory consumed by the learning process. The memory usage is approximately stable with respect to the size of the initial training set.

7.4 Learning component analysis

For this analysis, we consider the Indian Pines dataset.

7.4.1 Experimental set-up

We investigate how the classification accuracy can be influenced by the several learning components (i.e. SVM kernels, feature profiles, transductive learning, ensemble learning and iterative collective inference) that contribute to the definition of S^2TEC. By combining these components differently, we define several learning frameworks, whose characteristics are summarized in Table 4.

S^2TEC-linear is equivalent to algorithm S^2TEC with the SVMs learned by considering the linear kernel in place of the Gaussian kernel.

S^2TEC(S + F) is equivalent to algorithm S^2TEC with the spectral profile and only the frequency profile considered for populating the ensemble. S^2TEC(S + M) is equivalent to S^2TEC with the spectral profile and only the morphology profile considered for populating the ensemble. In both frameworks, a pixel is a consensus pixel for the ensemble, if both classifiers of the ensemble assign the same label to the pixel.

S-SVM adopts a two-level learning schema. In the first level, a classifier is induced from the labeled set with the spectral features. It is used to predict labels of the unlabeled set. In

Table 4 Compared learning frameworks

Learning components	Framework name
SVM (Gaussian kernel), spectral profile, spatial (-frequency) profile and spatial (-morphology) profile, transductive learning, ensemble learning and iterative collective inference	S^2TEC
SVM (Linear kernel), spectral profile, spatial (-frequency) profile and spatial (-morphology) profile, transductive classifier, ensemble learning and iterative transductive collective inference	S^2TEC-linear
SVM (Gaussian kernel), spectral profile and spatial (-frequency) profile, transductive classifier, ensemble learning and iterative transductive collective inference	S^2TEC(S + F)
SVM (Gaussian kernel), spectral profile and spatial (-morphology) profile, transductive classifier, ensemble learning and iterative transductive collective inference	S^2TEC(S + M)
SVM (Gaussian kernel), spectral profile, spatial (-frequency) profile and spatial (-morphology) profile, inductive classifier and transductive collective inference	S-SVM
SVM (Gaussian kernel), spectral profile, spatial (-frequency) profile and spatial (-morphology) profile, inductive classifier and inductive collective inference	I + C
SVM (Gaussian kernel), spectral profile, spatial (-frequency) profile and spatial (-morphology) profile, inductive classifier and iterative transductive collective inference	I + C + I
SVM (Gaussian kernel), spectral profile, transductive classifier and iterative inference	T + SVM
SVM (Gaussian kernel), spectral profile, transductive classifier and self training	selfSVM

the second level, both frequency features and morphological features are constructed for the entire data set. A new classifier is induced from the labeled set with all these spatial features. It is used to finally classify the unlabeled part of the image.

I + C is a fully inductive variant of S^2TEC. The learning phase is performed by using the labeled part of the image only. Each classifier is induced from the labeled part of the image; both the frequency features and the morphological features of the labeled examples are constructed through collective inference by considering examples of the labeled set only. Three classifiers are induced with the features of the three considered profiles (spectral, spatial-frequency and spatial-morphology). The ensemble of these classifiers is used to classify the unlabeled set with the majority rule. As unlabeled data are considered neither to learn the classifiers nor to construct the relational features, the iterative schema is left out of this case.

IvC + I is a version of S^2TEC, which learns classifiers in the inductive setting, but performs collective inference with the iterative schema in the transductive setting. Similarly to I + C, three classifiers are learned, in the inductive setting. Features of the spatial profiles (frequency and morphology) are constructed iteratively by using the entire dataset. Thus, differently from I + C, new classifiers can be iteratively learned from the spatial data profiles, even staying in

Table 5 Analysis of learning components (Indian Pines): the accuracy metrics are collected on five trials, the average (SD) of the measures is computed on these trials

The neighborhood sizes are equal to 5, 10, 15, 20, while the labeled set percentage is equal to 5 %. The compared learning frameworks are those described in Table 4

Learning framework	OA	AA	κ
S^2TEC	**.945** (.036)	**.876** (.046)	**.927** (.011)
S^2TEC-linear	.930 (.013)	.818 (.027)	.921 (.014)
S^2TEC(S + F)	.887 (.013)	.811 (.032)	.871 (.015)
S^2TEC (S + M)	.868 (.010)	.776 (.020)	.849 (.012)
S-SVM	.898 (.028)	.769 (.028)	.885 (.032)
I + C	.876 (.010)	.733 (.013)	.858 (.011)
I + C + I	.940 (.012)	.865 (.028)	.911 (.013)
T + SVM	.723 (.018)	.630 (.014)	.682 (.021)
selfSVM	.739 (.011)	.626 (.011)	.701 (.013)

the inductive setting, as the spatio-relational features can be updated according to the new labels assigned by the ensemble to the originally unlabeled pixels. The algorithm stops when the number of classifications changed by the ensemble is less than a threshold (10 in this study).

T + SVM considers only spectral information. It performs iterative learning in the transductive setting. At each iteration, the SVM is learned from the labeled part and used to predict the unlabeled part of the image. An estimate of the probability of a label is here predicted by fitting a logistic regression classifier to the SVM classifier (Platt 1999). Unlabeled examples are sorted in descending order according to the probability of the predicted label. The top-k unlabeled examples are moved from the unlabeled set to the labeled one for the next iteration. In this study $k = 500$. The algorithm stops when all data have been moved from the unlabeled part to the labeled one.

Finally, selfSVM considers only spectral information. It performs iterative learning with the self training approach described by Li et al. (2008). Initially, the SVM is induced from the labeled part with the spectral features. This classifier is used to label all the pixels of the unlabeled set. Subsequently, the training set is the "entire" dataset with real labels associated with examples of the labeled part and predicted labels associated with examples of the unlabeled part. Iteratively, the SVM is learned from this training set and used to re-predict labels of the unlabeled part. The algorithm stops when the number of classifications changed by the SVM is less than a threshold (10 in this study).

We compare the overall accuracy, average accuracy and Cohen's kappa coefficient of these learning frameworks by constructing spatial neighborhoods with size 5, 10, 15, 20 and considering the initial labeled set sampled with the labeling percentage equal to 5 %.

7.4.2 Results and discussion

Table 5 reports the accuracy metrics collected for the alternative learning frameworks described in Table 4. These results deserve several considerations.

Firstly, the comparison between S^2TEC and S^2TEC-linear allows us to perform the analysis of the presented algorithm as a function of the kernel considered when learning SVMs. The results (rows 1–2, Table 5) show that the Gaussian kernels yield more accurate classifications than the linear kernels. This confirms the general trend described in Bruzzone et al. (2006), Tarabalka et al. (2010b), Bovolo et al. (2006), Melgani and Bruzzone (2004), Huang and He (2012) and Li et al. (2012), which usually resorts to SVMs learned with Gaussian

kernels, in order to address the problem of hyperspectral image classification. The higher accuracy is mainly due to the fact that a Gaussian kernels can capture the intimate nonlinear nature of the problem better, while a linear kernel leads to the construction of a linear classifier in high-dimensional spaces of features which can be nonlinearly related to the input space.

Secondly, the comparison among S^2TEC, S^2TEC(S+F), S^2TEC(M+F), T+SVM and selfSVM allows us to perform the analysis of the presented algorithm as a function of the processed feature profiles. In particular, the results (rows 1, 3–4, 8–9, Table 5) show that the learning frameworks using both spectral and spatial profiles (S^2TEC, S^2TEC(S+F), S^2TEC(S+M)) yield more accurate classifications than the learning frameworks using the spectral profile only (T+SVM and selfSVM). This confirms the considerations already reported in Li et al. (2011, 2012), Tarabalka et al. (2010b), Khodadadzadeh et al. (2014b), Li et al. (2013a), Plaza et al. (2009), Bovolo et al. (2006), Fauvel et al. (2012), Tarabalka et al. (2010a), Camps-Valls et al. (2007) and Wang et al. (2014), which inspire the emerging, recent trend of considering spatial information, in addition to spectral information in imagery data. At the same time, by focusing this analysis on the learning frameworks using the spatial information, the results (rows 1, 3–4, Table 5) show that the classification accuracy produced with one spectral profile and "two" spatial profiles (S^2TEC) is higher than the classification accuracies produced with one spectral profile and one spatial profile, i.e. frequency (S^2TEC(S+F)) or morphology (S^2TEC(S+M)). This confirms the considerations already reported in Huang and He (2012), which inspire our idea of considering "various" (and possibly independent) spatial profiles of imagery data.

Thirdly, the comparison between S^2TEC, S-SVM and I+C allows us to evaluate the contribution of iterative learning in combination with collective inference. In the compared frameworks, collective inference is performed by learning the relational features constructed by using the labels of the related neighbors of the instance. Both S-SVM and I+C do not perform iterative learning, while S^2TEC resorts to iterative learning for collective inference. The results (rows 1, 5–6, Table 5) show that the use of iterative learning really improves the classification accuracy. This confirms the results of previous studies (Neville and Jensen 2000; Getoor 2005; Bilgic et al. 2007; McDowell et al. 2007; Fang et al. 2013) in collective inference, which have assessed the effectiveness of iterative learning, in order to account for the correlation of labels.

Fourthly, the comparison between S^2TEC, I+C and I+C+I allows us to perform the analysis of the presented algorithm as a function of the learning setting adopted (inductive learning vs transductive learning). We compare classification accuracies produced with classifiers learned in the inductive setting and collective inference performed in the inductive setting (I+C), to classification accuracies produced with classifiers learned in the inductive setting and collective inference performed with iterative learning in the transductive setting (I+C+I). We also compare them to classification accuracies produced with classifiers learned in the transductive setting and collective inference performed in the transductive setting (S^2TEC). The results (rows 2, 6–7, Table 5) show that the highest accuracy is achieved when the learning process is completed in a purely transductive setting (S^2TEC), while the lowest accuracy is achieved when the learning process is completed in a purely inductive setting (I+C). These results confirm the point of view of Vapnik (1998) that learning classifiers by accounting for both labeled and unlabeled data contribute to improving accuracy.

Finally, we observe that S^2TEC can produce the highest accuracy in Table 5 only by learning SVMs with Gaussian kernels, using both spectral and multiple spatial profiles, as

well as performing transductive learning, ensemble learning and iterative collective inference in the defined learning framework.

7.5 Hyperspectral image processing perspective

Several algorithms, which have been designed in the hyperspectral image classification literature, have been evaluated by considering Indian Pines, Pavia University and/or Salinas Valley datasets. In this study, we consider the most recent (and competitive) results produced by investigating both transductive SVMs and spectro-spatial classifiers in these data scenarios.

Hyperspectral Trasnsductive SVMs Melgani and Bruzzone (2004) have proposed a transductive SVM specially designed for hyperspectral image classification, while Plaza et al. (2009) have evaluated the performance of this classifier by using Indian Pines imagery data and a semi-supervised experimental setting. In their experiment, from the 16 different land-cover classes (see Fig. 5a), 7 have been discarded since the authors have judged that an insufficient number of training samples was available. The remaining 9 classes (corn notill, corn mintill, grass/pasture, grass/tree, hay-windrowed, soybean notill, soybean mintill, soybean cleantill and woods) have been used to generate 4757 training samples and 4588 validation samples. 5 % of training samples have been sampled to feed the labeled set, while the remaining 95 % of training samples have been used to feed the unlabeled set. The transductive SVM has been built from both labeled and unlabeled data of the training set and then evaluated on the validation set. In this study, we perform some experiments by simulating this experimental setting. We divide data into a training (4757 pixels) and a validation set (4588 pixels). We sparsely label 5 % of training pixels, using the ground truths, for the iterative convergence learning ensemble. We consider SVMs built from the spectral space at the last iteration of the ensemble, in order to label the validation test and collect the accuracy metrics. However, we do not use exactly the same data samples adopted by Plaza et al. (2009), we consider the same sampling sizes and run S^2TEC on several sampling trials. In particular, we perform five random partitions between the training set and the validation set and, for each training partition, we generate five random trials to sample labeled data, for a total of 25 random trials. Results are averaged on these trials. The results show that, by using this "semi-supervised" setting , S^2TEC achieves average OA equal to 0.854 (with SD equal to 0.022) and κ equal to 0.828 (with SD equal to 0.026). Both measures greatly outperform OA = 0.762 and $\kappa = 0.710$ performed by TSVM in Plaza et al. (2009). This confirms, once again, the ability of our methodology to outperform existing transductive classifiers.

Spectral-spatial classifiers The classification accuracy performed by several recently defined algorithms, integrating spectral and spatial information, is analyzed in (Li et al. 2013a, 2011, 2012; Tarabalka et al. 2010b, a; Guccione et al. 2015). Evaluated classifiers include: LORSAL-MLL that resorts to a multilevel logistic prior encoding the spatial information and using active learning (Li et al. 2011); MPM-LBP that considers spectral and spatial information contained in the original hyperspectral data by using loopy belief propagation and active learning (Li et al. 2013a); MLRsubMLL that integrates spectral and spatial information in a Multinomial Logistic Regression (MLR) algorithm and uses Multilevel Logistic Markov-Gibbs with Markov random field prior, in order to synthesize the spatial information (Li et al. 2012); SVMMRF that firstly applies a probabilistic support vector machine spectral-based classification and then refines the classification results by using spatial contextual information through a Markov random field regularization (Tarabalka et al. 2010b); a spatial-aware SVM

that learns SVMs after extending the spectral feature space with a spatial-aware morphological profile (Plaza et al. 2009; Li et al. 2013a);[8] Watershed that uses watershed segmentation, in order to define information on spatial structures and perform spectral-based SVM classification, followed by majority voting within the watershed regions (Tarabalka et al. 2010a; Li et al. 2013a); as well as IRMC that implements two MLR classifiers, which are fed with spectral features and spatial features, respectively, and work iteratively so that every classifier exploits the decision of the other one (Guccione et al. 2015).

For Indian Pines, results have been produced in the literature with 5 % (IRMC), 6 % (SVMMRF) and 10 % (LORSAL-MLL, MPM-LBP, MLRsubMLL and SVMMRC) of the pixels labeled according to the available ground truths. For Pavia University, results have been produced in the literature with 5 % (IRMC) and 9 % (spatial SVM, LORSAL-MLL, MPM-LBP, MLRsubMLL, SVMMRC and Watershed) of the pixels labeled according to the available ground truths. For Salinas Valley, results have been produced in the literature with 5 % (IRMC) of the pixels labeled according to the available ground truths. In all these datasets, the remaining pixels have been unlabeled according to the proper transductive setting. Finally, all the original classes of Indian Pines, Pavia University and Salinas Valley have been used in the literature to evaluate the competitors considered in this section. The accuracy reported in Li et al. (2013a, 2011, 2012); Tarabalka et al. (2010b) is achieved by starting from labeled samples which are different from those considered for this study. Thus, the comparison in these cases is not properly safe. However, the low SD of the metrics computed for S^2TEC on several trial supports the theory that it should perform equally well if run with the labeled sets used in Li et al. (2013a, 2011, 2012); Tarabalka et al. (2010b).

Accuracy results are reported in Table 6. The accuracy metrics of all the competitors are collected with the percentage of the pixels labeled for the learning phase greater than or equal to 5 %. In any case, S^2TEC almost always outperforms its competitors with 5 % of the pixels labeled. The only exceptions are observed when MPM-LBP, MLRsubMLL and SVMMRF are used to classify Indian Pines imagery data. We note that MPM-LB, that uses loopy belief propagation and active learning, has been evaluated in Li et al. (2013a) with 10 % of the pixels labeled in Indian Pines. Hence, it is reasonable to compare S^2TEC and MPM-LB, when both algorithms are run with 10 % of the pixels labeled in Indian Pines. In this case, we observe that S^2TEC achieves higher OA and κ than MPM-LBP, while MPM-LB achieves higher AA than S^2TEC.

Let us now analyze the accuracy of MLRsubMLL and SVMMRF in Indian Pines, both of which use Markov random fields. Also in this case, the best accuracy of these competitors is achieved only in association to metric AA. In particular, OA and κ achieved by both MLRsubMLL (with 10 % of the pixels labeled) and SVMMRF (with the 6 % of the pixels labeled) are lower than OA and κ achieved by S^2TEC (with both 5 and 10 % of the pixels labeled). In contrast, AA achieved by S^2TEC is lower than AA achieved by both MLRsubMLL and SVMMRF, although differences in AA are smaller when S^2TEC is run with 10 % of the pixels labeled. In order to interpret this result, we look at the class-by-class results reported in Li et al. (2013a) for MPM-LBP, in Li et al. (2012) for MLRsubMLL and in Tarabalka et al. (2010b) for SVMMRF. We observe that the gain in average accuracy is due to the ability of these competitors to achieve higher accuracy than S^2TEC, when classifying the pixels belonging to the minority classes "Oat" and "AlfaAlfa". However, these competitors are outperformed by S^2TEC when considering all the remaining classes.

[8] This spatial-aware morphological profile includes opening and closing features, as well as additional spatial information like size, orientation and local contrast.

Table 6 Accuracy metrics (± SD) of state-of-art, spatio-spectral, hyperspectral classifiers

Algorithm	References	Labeled (%)	OA	AA	κ
Indian Pines					
S^2TEC	Table 1	5	.945±.010	.876±.036	.937±.011
S^2TEC	Table 3	10	.975±.006	.927±.014	.971±.007
LORSAL-MLL	Li et al. (2011, 2013a)	10	.927	.951	.916
MPM-LBP	Li et al. (2013a)	10	.947	.962	.939
MLRsubMLL	Li et al. (2012)	10	.936	.939	.926
SVMMRF	Tarabalka et al. (2010b)	6	.920	.958	.909
IRMC	Guccione et al. (2015)	5	.873±.031	.821±.075	.856±.035
Pavia University					
S^2TEC	Table 1	5	.992±.001	.988±.003	.989±.001
Spatial SVM	Plaza et al. (2009) Li et al. (2013a)	9	.852	.907	.808
LORSAL-MLL	Li et al. (2011, 2013a)	9	.855	.925	.818
MPM-LBP	Li et al. (2013a)	9	.857	.922	.820
MLRsubMLL	Li et al. (2012)	9	.941	.935	.922
SVMMRF	Tarabalka et al. (2010b)	9	.976	.945	.959
Watershed	Tarabalka et al. (2010a) Li et al. (2013a)	9	.854	.913	.813
IRMC	Guccione et al. (2015)	5	.867±.008	.837±.007	.814±.010
Salinas Valley					
S^2TEC	Table 1	5	.980±.006	.988±.003	.977±.006
IRMC	Guccione et al. (2015)	5	.955±.018	.952±.016	.950±.020

8 Conclusion

In this paper, we propose a new transductive hyperspectral image classification algorithm, that integrates spectral and spatio-relational features into an ensemble system developed with an iterative convergence algorithm. According to our knowledge, this represents an innovative contribution in the related field, showing that collective inference can be used with transductive learning and ensemble learning, in order to enhance the accuracy of traditional classifiers. In this context, collective inference is used to account for spatial autocorrelation of labels, transductive learning is considered to deal with sparsely labeled data, while ensemble learning is aopted to deal with spectral and spatial information. The empirical study proves that the presented methodology outperforms traditional classifiers, transductive classifiers and several spatio-spectral algorithms proposed in the hyperspectral image classification literature. However, it still fails to correctly classify some pixels along the margin bound between homogeneously, well classified zones.

Some directions for further work are still to be explored. New learning systems, such as co-training, can be investigated as an alternative to ensemble. At the same time, active learning solutions can be explored, in order to "intelligently" augment the labeled set along the iterative process. In addition, we can account for the conclusions suggested by the analysis

of hyperspectral competitors and try to improve classification accuracy of minority classes by resorting to Markov random fields. On the other hand, we can also explore the possibility of performing collective inference through Gibbs sampling (as an alternative to iterative convergence learning) also in combination with Markov random fields and/or active learning. Additionally, it would be interesting to study new ways to adaptively determine both the size and the shape of spatial neighborhoods. Finally, the algorithm combines principles of both collective inference and transduction, in order to address a classification task in the relational setting. Although it is specifically designed for hyperspectral imagery classification, the algorithm can be considered general-purpose once the schema to determine relational profiles of data are established. A future piece of work will focus on the definition of a general schema to determine relational profiles of data with correlation. In this way, we will be able to investigate the effectiveness of the proposed algorithm in new application environments. In any case, this is out of the scope of the presented manuscript.

Acknowledgments We would like to acknowledge the support of the European Commission through the project MAESTRA—Learning from Massive, Incompletely annotated, and Structured Data (Grant Number ICT-2013-612944). The authors wish to thank Donato Aquilino and Roberto Stanziale for their support in developing the component of the software for computing the relational features, Thorsten Joachims for his support in using SGT, Vikas Sindhwani for his support in using SVMLin, Lynn Rudd for her help in reading the manuscript and Luigi Mascolo for his useful discussions on studies investigating SVMs.

References

Ablin, R., & Sulochana, C. (2013). A survey of hyperspectral image classification in remote sensing. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(8), 2986–3000.

Antanas, L., van Otterlo, M., Mogrovejo, J. O., Tuytelaars, T., & Raedt, L. D. (2014). There are plenty of places like home: Using relational representations in hierarchies for distance-based image understanding. *Neurocomputing*, *123*, 75–85.

Appice, A., & Malerba, D. (2014). Leveraging the power of local spatial autocorrelation in geophysical interpolative clustering. *Data Mining and Knowledge Discovery*, *28*(5–6), 1266–1313.

Appice, A., Guccione, P., Malerba, D., & Ciampi, A. (2014). Dealing with temporal and spatial correlations to classify outliers in geophysical data streams. *Information Sciences*, *285*, 162–180.

AVIRIS. 2007. http://aviris.jpl.nasa.gov/

Benediktsson, J., Pesaresi, M., & Amason, K. (2003). Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(9), 1940–1949.

Bilgic, M., Namata, G. M., & Getoor, L. (2007). Combining collective classification and link prediction. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, ICDMW 2007* (pp. 381–386). IEEE Computer Society.

Blum, A., & Chawla, S. (2001). Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001* (pp. 19–26). Morgan Kaufmann Publishers Inc.

Bovolo, F., Bruzzone, L., & Marconcini, M. (2006). A novel context-sensitive SVM for classification of remote sensing images. In *IEEE International Conference on Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006* (pp. 2498–2501).

Bruzzone, L., Chi, M., & Marconcini, M. (2006). A novel transductive SVM for semisupervised classification of remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(11), 3363–3373.

Camps-Valls, G., Bandos Marsheva, T., & Zhou, D. (2007). Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(10), 3044–3054.

Ceamanos, X., Waske, B., Benediktsson, J., Chanussot, J., & Sveinsson, J. (2009). Ensemble strategies for classifying hyperspectral remote sensing data. In J. Benediktsson, J. Kittler, & F. Roli (Eds.), *Multiple Classifier Systems, Lecture Notes in Computer Science* (Vol. 5519, pp. 62–71). Berlin: Springer.

Ceci, M., & Appice, A. (2006). Spatial associative classification: Propositional vs structural approach. *Journal of Intelligent Information Systems*, *27*(3), 191–213.

Ceci, M., Berardi, M., & Malerba, D. (2007). Relational data mining and ILP for document image understanding. *Applied Artificial Intelligence*, *21*(4&5), 317–342.

Ceci, M., Appice, A., Viktor, H. L., Malerba, D., Paquet, E., & Guo, H. (2012). Transductive relational classification in the co-training paradigm. In P. Perner (Ed.), *Proceedings of the 8th International Conference Machine Learning and Data Mining in Pattern Recognition, MLDM 2012, Lecture Notes in Computer Science* (Vol. 7376, pp. 11–25). Springer.

Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999–3011.

Chang, C. I. (2007). *Hyperspectral data exploitation: Theory and applications*. New York: Wiley.

Chechetka, A., Dash, D., & Philipose, M. (2010). Relational learning for collective classification of entities in images. In *Statistical Relational Artificial Intelligence, Papers from the 2010 AAAI Workshop, AAAI, AAAI Workshops* (Vol. WS-10-06).

Chen, C., Li, W., Su, H., & Liu, K. (2014). Spectral-spatial classification of hyperspectral image based on kernel extreme learning machine. *Remote Sensing*, *6*(6), 5795–5814.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dundar, M., Krishnapuram, B., Bi, J., & Rao, R. B. (2007). Learning classifiers when the training data is not IID. In M. M. Veloso (Ed.), *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI 2007* (pp. 756–761).

Fang, M., Yin, J., & Zhu, X. (2013). Transfer learning across networks for collective classification. In *Proceedings of the 13th International Conference on on Data Mining, ICDM 2013* (pp. 161–170). IEEE Computer Society.

Fauvel, M., Chanussot, J., & Benediktsson, J. (2012). A spatial-spectral kernel-based approach for the classification of remote-sensing images. *Pattern Recognition*, *45*(1), 381–392.

Fauvel, M., Tarabalka, Y., Benediktsson, J., Chanussot, J., & Tilton, J. (2013). Advances in spectral-spatial classification of hyperspectral images. *Proceedings of the IEEE*, *101*(3), 652–675.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI–6*(6), 721–741.

Getoor, L. (2005). Link-based classification. In *Advanced Methods for Knowledge Discovery from Complex Data, Advanced Information and Knowledge Processing* (pp. 189–207). London: Springer.

Getoor, L., & Taskar, B. (2007). *Introduction to statistical relational learning (adaptive computation and machine learning)*. Cambridge, MA, London: The MIT Press.

Goetz, A., Vane, G., Solomon, J., & Rock, B. (1985). Imaging spectrometry for earth remote sensing. *Science*, *228*(4704), 1147–1153.

Green, R. O., Eastwood, M. L., Sarture, C. M., Chrien, T. G., Aronsson, M., Chippendale, B. J., et al. (1998). Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, *65*(3), 227–248.

Guccione, P., Mascolo, L., & Appice, A. (2015). Iterative hyperspectral image classification using spectral-spatial relational features. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(7), 3615–3627.

Huang, C., Davis, L., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, *23*, 725–749.

Huang, R., & He, W. (2012). Using tri-training to exploit spectral and spatial information for hyperspectral data classification. In *2012 International Conference on Computer Vision in Remote Sensing, CVRS 20012* (pp. 30–33).

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, *14*(1), 55–63.

Isaaks, E. H., & Srivastava, M. R. (1990). *An introduction to applied geostatistics*. USA: Oxford University Press.

Jensen, D., Neville, J., & Gallagher, B. (2004) Why collective inference improves relational classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 2004* (pp. 593–598). ACM.

Joachims, T. (1999). Transductive inference for text classification using Support Vector Machines. In I. Bratko & S. Dzeroski (Eds.), *Proceedings of the 16th International Conference on Machine Learning, (ICML 1999)* (pp. 200–209). Morgan Kaufmann.

Joachims, T. (2003). Transductive learning via spectral graph partitioning. In T. Fawcett & N. Mishra (Eds.), *Proceedings of the 20th International Conference on Machine Learning, ICML 2003* (pp. 290–297). AAAI Press.

Khodadadzadeh, M., Li, J., Plaza, A., Gamba, P., Atli Benediktsson, J., & Bioucas-Dias, J. (2014a). A new framework for hyperspectral image classification using multiple spectral and spatial features. In *2014*

IEEE International Conference on Geoscience and Remote Sensing Symposium (IGARSS) (pp. 4628–4631).

Khodadadzadeh, M., Li, J., Plaza, A., Ghassemian, H., Bioucas-Dias, J. M., & Li, X. (2014b). Spectral-spatial classification of hyperspectral data using local and global probabilities for mixed pixel characterization. *IEEE Transactions of Geoscience and Remote Sensing*, *52*(10), 6298–6314.

Kong, X., Ng, M., & Zhou, Z. H. (2013). Transductive multilabel learning via label set propagation. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 704–719.

Krogel, M. A., Rawles, S., Zelezny, F., Flach, P. A., Lavrac, N., & Wrobel, S. (2003). Comparative evaluation of approaches to propositionalization. In T. Horvarth & A. Yamamoto (Eds.), *Inductive Logic Programming, Lecture Notes in Computer Science* (Vol. 2835, pp. 197–214). Berlin: Springer.

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, *74*(6), 1659–1673.

LeSage, J. H., & Pace, K. (2001). Spatial dependence in data mining. In R. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, & R. Namburu (Eds.), *Data mining for scientific and engineering applications* (pp. 439–460). Dordrecht: Kluwer Academic.

Li, J., Bioucas-Dias, J., & Plaza, A. (2011). Hyperspectral image segmentation using a new bayesian approach with active learning. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(10), 3947–3960.

Li, J., Bioucas-Dias, J., & Plaza, A. (2012). Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(3), 809–823.

Li, J., Bioucas-Dias, J., & Plaza, A. (2013a). Spectral-spatial classification of hyperspectral data using loopy belief propagation and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(2), 844–856.

Li, J., Reddy Marpu, P., Plaza, A., Bioucas-Dias, J., & Atli Benediktsson, J. (2013b). Generalized composite kernel framework for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(9), 4816–4829.

Li, Y., Guan, C., Li, H., & Chin, Z. (2008). A self-training semi-supervised SVM algorithm and its application in an eeg-based brain computer interface speller system. *Pattern Recognition Letters*, *29*(9), 1285–1294.

Malerba, D. (2008). A relational perspective on spatial data mining. *International Journal of Data Mining, Modelling and Management*, *1*(1), 103–118.

Malerba, D., Ceci, M., & Appice, A. (2009). A relational approach to probabilistic classification in a transductive setting. *Engineering Applications of Artificial Intelligence*, *22*(1), 109–116.

McDowell, L., & Aha, D. W. (2012). Semi-supervised collective classification via hybrid label regularization. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*. Omnipress.

McDowell, L., Gupta, K. M., & Aha, D. W. (2007). Case-based collective classification. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 20th International Florida Artificial Intelligence Research Society Conference* (pp. 399–404). AAAI Press.

McDowell, L., Gupta, K. M., & Aha, D. W. (2009). Cautious collective classification. *Journal of Machine Learning Research*, *10*, 2777–2836.

Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, *42*(8), 1778–1790.

Miao, L., Shuying, Z., Zhang, B., Shanshan, L., & Changshan, W. (2014). A review of remote sensing image classification techniques: The role of spatio-contextual information. *European Journal of Remote Sensing*, *47*, 389–411.

Mizoguchi, F., Ohwada, H., Daidoji, M., & Shirato, S. (1997). Using inductive logic programming to learn rules that identify glaucomatous eyes. In N. Lavrac, E. Keravnou, & B. Zupan (Eds.), *Intelligent data analysis in medicine and pharmacology, the Springer international series in engineering and computer science* (Vol. 414, pp. 227–242). US: Springer.

Munoz-Mari, J., Bovolo, F., Gomez-Chova, L., Bruzzone, L., & Camp-Valls, G. (2010). Semisupervised one-class support vector machines for classification of remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(8), 3188–3197.

Neville, J., & Jensen, D. (2000). Iterative classification in relational data. In *Proceedings of 17th International Joint Conference on Artificial Intelligence*. AAAI Press.

Neville, J., & Jensen, D. (2007). Relational dependency networks. *Journal of Machine Learning Research*, *8*, 653–692.

Neville, J., Simsek, O., & Jensen, D. (2004). Autocorrelation and relational learning: Challenges and opportunities. In *Proceedings of Workshop on Statistical Relational Learning* (pp. 290–299). AAAI Press.

Orkin, M., & Drogin, R. (1990). *Vital statistics*. New York: McGraw Hill.

Pesaresi, M., & Benediktsson, J. (2001). A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *39*(2), 309–320.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, B. Scholkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* (pp. 61–74). MIT Press.

Plaza, A., Benediktsson, J. A., Boardman, J. W., Brazile, J., Bruzzone, L., Camps-Valls, G., et al. (2009). Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, *113*(Supplement 1), S110–S122.

Ratle, F., Camps-Valls, G., & Weston, J. (2010). Semisupervised neural networks for efficient hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(5), 2271–2282.

Richards, J. A. (1993). *Remote sensing digital image analysis: An introduction* (2nd ed.). Secaucus, NJ: Springer.

ROSIS & HySpex. (1995). http://messtec.dlr.de/en/technology/dlr-remote-sensing-technology-institute/hyperspectral-systems-airborne-rosis-hyspex/index.php

Saha, T., Rangwala, H., & Domeniconi, C. (2012). Multi-label collective classification using adaptive neighborhoods. In *Proceedings of the 11th International Conference on Machine Learning and Applications, ICMLA 2012* (Vol. 1, pp. 427–432).

Sammut, C., & Zrimec, T. (1998). Learning to classify x-ray images using relational learning. In C. Nedellec & C. Rouveirol (Eds.), *Proceedings of the European Conference of Machine Learning, ECML 1998, Lecture Notes in Computer Science* (Vol. 1398, pp. 55–60). Berlin: Springer.

Seeger, M. (2001). *Learning with labeled and unlabeled data*. Technical report.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, *29*(3), 93–106.

Shahshahani, B., & Landgrebe, D. (1994). The effect of unlabeled samples in reducing the small sample size problem and mitigating the hughes phenomenon. *IEEE Transactions on Geoscience and Remote Sensing*, *32*(5), 1087–1095.

Shi, X., Li, Y., & Yu, P. (2011). Collective prediction with latent graphs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011* (pp. 1127–1136). ACM.

Sindhwani, V., & Keerthi, S. S. (2006). Large scale semi-supervised linear SVMs. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, & K. Järvelin (Eds.), *Proceedings of the 29th Annual International Conference on Research and Development in Information Retrieval, SIGIR 2006* (pp. 477–484). : ACM.

Soille, P. (2003). *Morphological image analysis: Principles and applications* (2nd ed.). Springer Berlin Heidelberg.

Srinivasan, A., & King, R. D. (1999). Feature construction with inductive logic programming: A study of quantitative predictions of biological activity aided by structural attributes. *Data Mining and Knowledge Discovery*, *3*(1), 37–57.

Stojanova, D., Ceci, M., Appice, A., Malerba, D., & Dzeroski, S. (2013). Dealing with spatial autocorrelation when learning predictive clustering trees. *Ecological Informatics*, *13*, 22–39.

Sun, S. (2013). A survey of multi-view machine learning. *Neural Computing and Applications*, *23*(7–8), 2031–2038. doi:10.1007/s00521-013-1362-6.

Tan, K., Li, E., Du, Q., & Du, P. (2014). Hyperspectral image classification using band selection and morphological profiles. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(1), 40–48.

Tarabalka, Y., Chanussot, J., & Benediktsson, J. (2010a). Segmentation and classification of hyperspectral images using watershed transformation. *Pattern Recognition*, *43*(7), 2367–2379.

Tarabalka, Y., Fauvel, M., Chanussot, J., & Benediktsson, J. (2010b). SVM- and MRF-based method for accurate classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, *7*(4), 736–740.

Taskar, B., Segal, E., & Koller, D. (2001). Probabilistic classification and clustering in relational data. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence, IJCAI 2001* (Vol. 2, pp. 870–876). Morgan Kaufmann Publishers Inc.

Taskar, B., Abbeel, P., & Koller, D. (2002). Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, UAI 2002* (pp. 485–492). Morgan Kaufmann Publishers Inc.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY: Springer.

Villa, A., Benediktsson, J., Chanussot, J., & Jutten, C. (2011). Hyperspectral image classification with independent component discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(12), 4865–4876.

Wang, L., Hao, S., Wang, Q., & Wang, Y. (2014). Semi-supervised classification for hyperspectral imagery based on spatial-spectral label propagation. *ISPRS Journal of Photogrammetry and Remote Sensing*, *97*, 123–137.

Waske, B., & Benediktsson, J. (2007). Fusion of support vector machines for classification of multisensor data. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(12), 3858–3866.

Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in MRFs. In M. Opper & D. Saad (Eds.), *Advanced Mean Field Methods* (pp. 229–243). Cambridge, MA, London: MIT Press.

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Yanover, C., & Weiss, Y. (2002). Approximate inference and protein folding. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (pp. 84–86). MIT Press.

Zelezný, F., & Lavrac, N. (2006). Propositionalization-based relational subgroup discovery with RSD. *Machine Learning*, *62*(1–2), 33–63.