Land Classification Using Remotely Sensed Data: Going Multi-Label

Konstantinos Karalas, Grigorios Tsagkatakis, Michalis Zervakis, and Panagiotis Tsakalides

***Abstract*—Obtaining an up-to-date high resolution description of land cover is a challenging task due to the high cost and labor intensive process of human annotation through field-studies. This work introduces a radically novel approach for achieving this goal by exploiting the proliferation of remote sensing satellite imagery, allowing for the up-to-date generation of global-scale land cover maps. We propose the application of multi-label classification, a powerful framework in machine learning, for inferring the complex relationships between acquired satellite images and the spectral profiles of different types of surface materials. Introducing a drastically different approach compared to unsupervised spectral unmixing, we employ contemporary ground-collected data from the European Environment Agency to generate the label set, and multispectral images from the MODIS sensor to generate the spectral features, under a supervised classification framework. To validate the merits of our approach, we present results using several state-of-the-art multi-label learning classifiers and evaluate their predictive performance with respect to the number of annotated training examples, as well as their capability to exploit examples from neighboring regions or different time instances. We also demonstrate the application of our method on hyperspectral data from the Hyperion sensor for the urban land cover estimation of New York city. Experimental results suggest that the proposed framework can achieve excellent prediction accuracy, even from a limited number of diverse training examples, surpassing state-of-the-art spectral unmixing methods.**

***Index Terms*—Remote sensing, pattern classification, satellite applications, land cover, unmixing, data processing, MODIS, time series, CORINE.**

I. Introduction

LAND cover analysis aims at monitoring and mapping the geobiophysical parameters of the Earth's surface, a process critical in environmental and urban sciences studying the ever-changing evolution of our planet [1]. The characterization of the natural resources and their dynamics is the singular most important way of providing sound answers to the greatest environmental concerns of humanity today, including climate change, biodiversity loss, as well as pollution of water, soil and air. These vital needs mandate an increased effort in creating accurate and timely high spatial resolution land cover maps. Despite the urgency, such endeavors are hindered by various constraints, the most prominent of which is the labor-intensive manual process of collecting ground-based data from field surveys. To that end, remote sensing systems represent a major resource for monitoring global-scale variations in land cover [2], where high resolution imaging sensors retrieving optical, radar, multispectral, and hyperspectral data, are being employed to achieve this demanding objective.

During the remote sensing mapping procedure, a classification technique has to be applied in order to annotate the acquired pixels with additional metadata. In typical satellite image classification [3], especially in situations where multiple spectral bands are acquired, each pixel is restricted in its characterization to a single class from a set of two or more mutually exclusive classes [4]. Unfortunately, this approach is guided by an assumption that is often violated in real-life scenarios where airborne and spaceborne imagery pixels are simultaneously characterized by multiple classes/labels. This is due to the mixing of multiple signals, a phenomenon attributed to the physical properties of light, the interactions of photons with matter and the atmosphere, and the characteristics of the acquisition process [5]. Consequently, single class assignment is unrealistic and can lead to map ambiguities.

State-of-the-art methods try to address this problem by a process known as spectral unmixing [6], which is able to distinguish different materials contributing to a pixel. Despite the importance of the unmixing methods, the majority of the proposed approaches rely on extremely limited and outdated hand-labeled datasets, such as the Cuprite mining district data. A consequence of the lack of real data is that typically one artificially applies a theorized forward mixing process and tests the capabilities of the proposed algorithm on performing the inverse process [7]. The utilization of simulated/synthetic data can provide some intuition regarding the merits of each approach, however, it makes generalization of the behavior of these algorithms very difficult when they are applied under real conditions [8].

In this work, we propose a novel approach for modeling the relations between spectral pixels and ground characteristics through the introduction of *multi-label learning* [9], a powerful supervised machine learning paradigm. Departing from traditional single-label classification, in multi-label learning each sample is associated with multiple labels simultaneously. More importantly, the labels are also ranked according to their relevance to the given sample [10], a premise that is appealing for remote sensing applications.

We claim that multi-label learning can provide valuable information on remotely sensed data, especially in the case of land cover estimation, where the heterogeneity of different

K. Karalas is with the School of Electronic and Computer Engineering, Technical University of Crete, Greece, and the Institute of Computer Science, FORTH, Greece (e-mail: kkaralas@isc.tuc.gr).

G. Tsagkatakis is with the Institute of Computer Science, FORTH, Greece (e-mail: greg@ics.forth.gr).

M. Zervakis is with the School of Electronic and Computer Engineering, Technical University of Crete, Greece (e-mail: michalis@display.tuc.gr).

P. Tsakalides is with the Institute of Computer Science, FORTH and the Department of Computer Science, University of Crete, Greece (e-mail: tsakalid@ics.forth.gr).

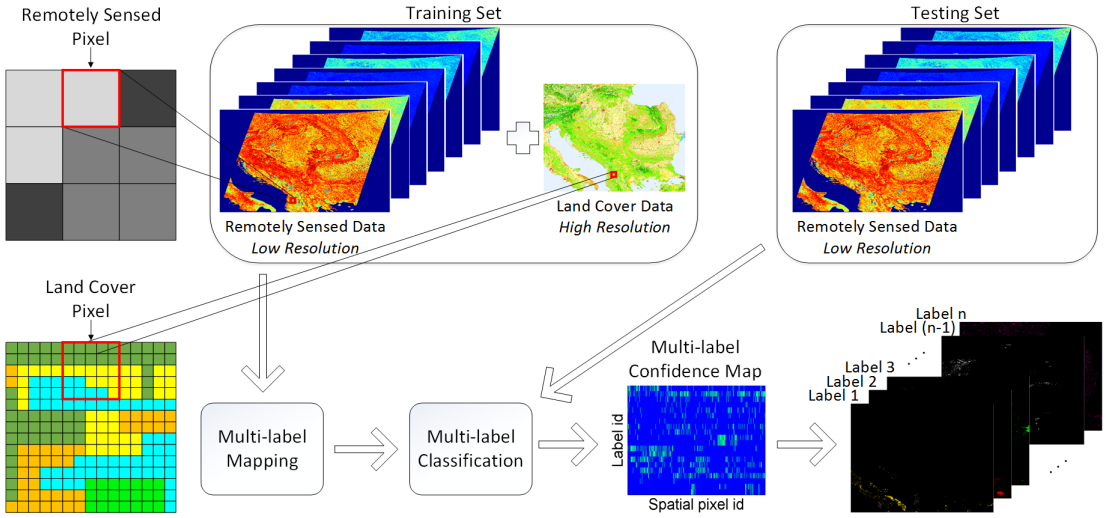


Fig. 1: Visual illustration of the multi-label classification process with remotely sensed data. A multi-label training set is generated by annotating multispectral satellite imagery with ground-sampled labels at higher spatial resolutions. Up-to-date land cover predictions are made through the use of multi-label classifiers that produce "multi-label confidence maps" encoding the presence of specific types of land cover.

regions introduces significant mixing of the corresponding signals. Compared to typical spectral unmixing, the proposed multi-label approach offers numerous advantages. In multi-label learning, one does not have to assume prior generating processes, in contrast to unmixing algorithms which rely on the expected type of mixing [11] and on the presence of pixels containing a single material, an assumption that is neither trivial nor universal. Moreover, multi-label classification employs a wide range of performance evaluation measures [10] which can provide a detailed characterization of the procedure from various viewpoints, whereas unmixing, being in principle an unsupervised procedure, has a limited number of well established performance quantification metrics and largely relies on the visual interpretation of abundance maps. Finally, the multi-label community has recognized that exploiting the dependencies between related labels and samples during the classification process is critical in order to improve prediction performance [12]. Thus, most of the state-of-the-art algorithms take into account such correlations, typically implicitly, in contrast to unmixing methods. All in all, multi-label classification can be thought as an alternative model to single-label classification for remotely sensed data, while in parallel it can provide supplementary solutions to the tasks of unmixing.

II. Reassessing Satellite Image Classification

Interestingly, there is a plethora of large unlabeled remote sensing datasets which remain unexploited, an issue which has fueled a lot of research around the identification of the optimal ways to utilize such data. One main obstacle is the problem of scale incompatibility. Whereas field-based measurements can be conducted at meter scales, distance to the ground and motion of the moving platforms are directly responsible for the considerably lower spatial resolution of remote sensing imagery. This spatial scale incompatibility between field-based and satellite-based sampling inevitably hinders the exploitation of the acquired measurements. On the other hand, despite their lower resolution, airborne and spaceborne platforms can provide imagery at significantly high temporal sampling rates, as more and more of these platforms are in flight and in orbit around the Earth.

Our proposed scheme, based on a multi-label learning concept, manages to jointly model ground-based land cover data and multispectral satellite imagery of different spatial resolutions into composite useful representations. This way, we provide a genuine answer to the scale incompatibility problem which arises naturally through the sampling procedure. More specifically, we combine data from the CORINE Land Cover (CLC) maps of 2000 and 2006 compiled by European Environment Agency (EEA) [13] at 100m^2 spatial resolution, corresponding to the European environmental landscape annotated by experts, with satellite data products from the MODIS database [14] at 500m^2 spatial resolution. Due to this difference in scale, each multispectral pixel may be associated with multiple labels, leading to the case of multi-label annotation.

The multi-label learning framework has attracted considerable attention in the literature over the last decade due to its numerous real-world applications [15]. Traditionally, it is applied in text [16], audio [17] and image classification [18], where a document could belong to several topics, a music song could fit to different genres, and an image could be annotated by many tags, respectively. One of the main challenges in multi-label learning is how to effectively utilize the correlations among different labels during classification [12]. In order to conceptually understand the significance of label dependencies, one can think of two images with a blue background depicting a ship and an airplane. Distinguishing these two images based solely on the color features is a challenging task for a classifier, since both contain large regions with blue color. However, if the system is confident enough that the image should be annotated with the "airplane" label, then it is more likely that the blue regions of the image should be

annotated with "sky", rather than "sea".

According to the proposed learning approach, the output of our (software) system corresponds to the predicted labels of a testing area together with their ranking, combined in an informative visualization graph termed "the multi-label confidence map". A high level overview of our proposed learning model is depicted in Fig. 1. In a nutshell, the key contributions of the proposed system are the following:

- the formulation of an efficient approach for the combination of high spatial resolution land cover data with low spatial resolution satellite images.
- the development of an architecture capable of using up-to-date remote sensing data and produce land cover maps with minimal labor-intensive hand-operated labeling.
- the systematic evaluation of state-of-the-art multi-label classification approaches on a novel and highly complex dataset.
- potential use of alternative modalities for extending the scheme to various sources of data, in addition to multispectral and land cover examined mainly in this work.

To the best of our knowledge, this is the first work which applies a multi-label classification scheme in remote sensing data, an approach that can effectively address the issues that naturally arise due to the multiple scales of the data, without requiring the explicit and often unrealistic modeling of the underlying generative processes. A key benefit of our method is the generation of accurate and up-to-date high resolution land cover maps, obtained through a new labeled dataset composed of freely available real data which can leverage the abundance of satellite imagery. The complete dataset will be available online.

The rest of this paper is structured as follows. Section III provides an overview of the related state-of-the-art. Section IV presents the multi-label classification methods considered in this work, whereas Section V exposes the datasets that are employed together with the experimental setup and the evaluation metrics. Section VI reports the experimental results, while conclusions and extensions of this work are presented in Section VII.

III. State-of-the-art

A. Remote Sensing Mapping and Classification

Since the 90's, satellite data have been extensively used for land cover mapping and classification. Land cover datasets have gained considerable attention since they provide a critical input for ecological and socioeconomic models at local, regional, and global scale. The most established such datasets include the Global Land Cover 2000, the GlobCover, and the MODIS land cover product which provide a global mapping [19], whereas the CORINE project [20] encompasses data for the European continent. Each dataset is prepared using different sources, classification algorithms, methodologies or even spatial resolution, leading in many cases to areas of uncertainty [21]. All of the above datasets have been investigated with a plethora of typical [22] as well as more sophisticated classification methods [23]. Notable among them, Support Vector Machines (SVM) exhibit very good classification performance of airborne and satellite imagery with limited training dataset, especially by incorporating composite kernels [24].

Apart from the learning algorithms, considering the sensors that are employed in the process is also of crucial importance for the quality of the features and thus the construction of the land cover maps and classification. There are two main categories of optical remote sensing systems: *multispectral* imaging devices which typically acquire 5 to 20 spectral bands, and *hyperspectral* ones which can acquire hundreds of spectral bands. Nevertheless, it has to be underlined that except for spectral information, the spatial, temporal as well as radiometric resolution properties of the sensor determine dominantly the success of classification [25]. Note that a higher spatial resolution generally implies a smaller coverage area of the system [26].

One of the first sensors that provided multispectral satellite data at a large scale was the NOAA's AVHRR instrument, which triggered many studies on land cover discrimination [27]. More recent and broadly used medium resolution remote sensing systems include PROBA-V on SPOT, TM and ETM+ on Landsat, and MODIS onboard Terra and Aqua satellites [14]. In order to compensate for the coarse resolution provided by these multispectral instruments, the use of time evolution of surface reflectance (time series) has proven to be valuable and thus it is adopted in most relevant studies [28]. On the opposite side, the most explored hyperspectral remote sensing scenes which are appropriate for supervised classification (containing ground-truth tables) were gathered by the AVIRIS (*e.g.*, Indian Pines, Salinas Valley, Kennedy Space Center) and the ROSIS (*e.g.*, Pavia Center, Pavia University) airborne sensors, which generate 224 and 115 contiguous spectral bands, respectively [29].

B. Spectral Unmixing

Under normal operating conditions, in remote sensing imaging systems each pixel (spectral vector) captures and encodes a multitude of signals. More precisely, on one hand nonlinear mixing of signals occurs when the light scattered by multiple materials in the scene is reflected of additional objects, as well as when two surrounding materials are homogeneously mixed. On the other hand, even in the ideal case where the incident light interacts with a single material, linear mixing occurs due to the instrumentation and various sources of noise [11].

Given the mixing of signals, there is a compelling need for a process that can separate the pixel spectra into a collection of pure materials, called endmembers. Spectral unmixing [6] aims at calculating the number of the endmembers, distinguishing their spectral signatures, and estimating their fractional abundances (*i.e.*, the proportion of each endmember's presence) in each pixel [11]. Typical spectral unmixing methods introduce certain assumptions regarding the mixing process, where the Linear Mixing Model (LMM), despite its simplicity, has been very successful in this context. Furthermore, due to the physical aspects of the data acquisition process, the unknown fractional abundance vector for a given

pixel is assumed to adhere to the Abundance Non-negativity Constraint and the Abundance Sum-to-one Constraint.

In order to decompose a mixed pixel spectrum, there exist two major classes of endmember extraction algorithms: the *geometrical* and the *statistical*. The geometrical approaches exploit the fact that mixed pixels lie inside a simplex. They are further divided into two subcategories: the *pure pixel based*, which assume that there is at least one pure pixel per endemember in the training data, *e.g.*, N-FINDR [30] and Vertex Component Analysis (VCA) [31], and the *minimum volume based*, which do not introduce such a prerequisite but seek to minimize the volume of the simplex, *e.g.* Simplex Identification via Split Augmented Lagrangian (SISAL) [32]. In the statistical methods, the abundance fractions are modeled as random variables and the spectral unmixing is formulated as a statistical inference problem. These include the Independent Component Analysis, which has been criticized due to the fact that the abundance fractions associated to each pixel are not statistically independent [33], and Bayesian approaches [34], which have a high computational complexity.

The abundance estimation part comprises the last step of the unmixing process. It can be solved via classical convex optimization methods, such as the Constrained Least Squares, which in this context minimizes the total squared error under the abundance non-negativity constraint, as well as the Fully Constrained Least Squares, which adds the abundance sum-to-one constraint to the constrained least squares problem. Meanwhile, sparse regression approaches have become popular, such as the the Sparse Unmixing by variable Splitting and Augmented Lagrangian (SUnSAL) [35], where sparse linear mixtures of spectra are investigated in a fashion similar to that of Compressed Sensing [36]. More recently, effort has been given to study nonlinear mixing models in order to handle specific kinds of nonlinearities, such as the Polynomial Post-Nonlinear Mixing Model (PPNMM) and its associated unmixing algorithm based on the subgradient method proposed in [37].

IV. Multi-label Classification

Intense research by the machine learning community has produced a large number of multi-label classification approaches, *e.g.*, [9], [10]. Existing approaches can be broadly divided into three categories: *problem transformation*, *algorithm adaptation*, and *ensemble methods* [38].

The intuition underlying problem transformation methods, is to decompose the original multi-label learning problem into a set of smaller and easier-to-learn binary classification problems to obtain a solution through well-established learning architectures. On the other hand, algorithm adaptation approaches adjust their internal structure in order to directly tackle multi-label data by employing a type of problem transformation. Representative techniques which have been adapted for the multi-label case include SVM [39], Boosting [16], Decision Trees (DT) [40], k-Nearest Neighbors (kNN) [41], and Artificial Neural Networks [42]. Ensemble methods have appeared more recently and are deployed on top of problem transformation or algorithm adaptation methods as wrappers, improving their generalization ability by gathering knowledge from multiple components [43]. According to this paradigm, multiple base learners are combined during the training phase to construct an ensemble, while a new instance is classified by integrating the outputs of single-label classifiers.

In our formulation, we assume a multi-label training set $\mathcal{D} = \{(\boldsymbol{x}_i, Y_i) \mid i = 1, \dots, n\}$, where Y_i is the actual labelset of the i-th example, and $\mathcal{L} = \{\lambda_j \mid j = 1, \dots, m\}$ is the set of all labels. For each unseen instance $\boldsymbol{x}$, we define $Z_{\boldsymbol{x}}$ as its predicted set of labels, and $r_{\boldsymbol{x}}(\lambda)$ as the associated ordered listing (rank) for label λ. The objective of multi-label classification is to estimate a set of decision rules $\mathcal{H}$ that maximize the probability of $\mathcal{H}(\boldsymbol{x}) = Z_{\boldsymbol{x}}$ for each example $\boldsymbol{x}$. Based on this notation, in the following section we discuss key representative examples from each category.

A. Problem Transformation Methods

Binary Relevance (BR) [44] is one of the earliest approaches in multi-label classification [9], where a single-label binary classifier is trained independently for each label, regardless of the rest of the labels (one-versus-all strategy). The method produces the union of the labels predicted by the binary classifiers, with the capability of ranking based on the classifier output scores. More specifically, in the BR approach, one trains a set of m classifiers such that:

$$\mathcal{H}_{BR} = \{h_j \mid h_j(\boldsymbol{x}) \rightarrow \lambda_j \in \{0, 1\}, j = 1, \dots, m\} \ . \quad (1)$$

BR is a straightforward approach for handling multi-label problems and is thus typically employed as a baseline method. The theoretical motivation and intuitive nature of BR are enhanced by additional attractive characteristics, such as moderate computational complexity (polynomial w.r.t. the number of labels), the ability to optimize several loss functions, and the potential of parallel execution [45]. An inherent drawback of the BR approach is the lack of consideration for label correlations which can lead to under or over estimation of the active labels, or the identification of multiple labels that never co-occur [46].

Another fundamental yet less extensively used transformation method is Label Powerset (LP) [44], where each existing combination of labels in the training set is considered as a possible label for the newly transformed multi-class classification problem. This way, the number of distinct mutually exclusive classes is upper bounded by $f = \min(n, 2^m)$, however, in practice it is much smaller [47]. For the classification of a new instance, the single-label classifier of LP outputs the most probable class, which can be now translated to a set of labels:

$$\mathcal{H}_{LP} = \{h_j \mid h_j(\boldsymbol{x}) \rightarrow \lambda_j \in \{0, 1\}, j = 1, \dots, f\} \ . \quad (2)$$

In contrast to BR, LP methods can capture inter-relationships among labels, at the cost of significantly higher computational complexity, which scales exponentially with the number of labels. Therefore LP is challenged in domains with large values of n and m. Furthermore, although this method is good at exact matches, it is prone to overfitting since it can only model labelsets which have been previously observed in the training set [48].

One can see that this type of transformations are universally applicable, since any traditional single-label classifier (DT, SVM, Naive Bayes, etc.) can be employed in order to obtain multi-label predictions. The overall complexity of classification is heavily dependent on the underlying single-label classification algorithm and the number of distinct label collections. Due to these properties, problem transformation methods are very attractive in terms of both scalability and flexibility, while they remain competitive with state-of-the-art methods [46].

B. Algorithm Adaptation Methods

The Multi-Label k-Nearest Neighbors (ML-kNN) method [41] constitutes an adaptation of the kNN algorithm for multi-label data following a Bayesian approach. It is a lazy learning algorithm which is based on retrieving the k nearest neighbors in the training set and then counting the number of neighbors belonging to each class (*i.e.*, a random variable W) [49]. Based on prior and posterior probabilities for the frequency of each label within these neighboring instances, it utilizes the Maximum a Posteriori (MAP) principle in order to determine the labelset for the unseen sample $\boldsymbol{x}$. The posterior probability of label $\lambda_j \in \mathcal{L}$ is thus given by:

$$P\left(\lambda_j \in Z_{\boldsymbol{x}} \mid W = w\right) = \frac{P\left(W=w \mid \lambda_j \in Z_{\boldsymbol{x}}\right) P\left(\lambda_j \in Z_{\boldsymbol{x}}\right)}{P(W=w)} \ . \quad (3)$$

Then, for each λ_j, ML-kNN builds a probabilistic classifier $h_j(\cdot)$ applying the rule: $h_j(\boldsymbol{x}) =$

$$\begin{cases} 1 \ \ P\left(\lambda_j \in Z_{\boldsymbol{x}} \mid W = w\right) > P\left(\lambda_j \notin Z_{\boldsymbol{x}} \mid W = w\right) \\ 0 \ \ \text{otherwise} \ . \end{cases} \quad (4)$$

A classifier's output of 1 indicates that λ_j is active for $\boldsymbol{x}$, while 0 indicates the opposite. Despite the fact that ML-kNN inherits merits from both lazy learning and Bayesian reasoning (*e.g.*, adaptive decision boundary due to the varying neighbors identified for each test instance), it is ignorant of the possible correlations between labels. Thus it is essentially a BR method which learns a single classifier $h_j(\cdot)$ for each label, independently from the others [10].

The Instance-Based Logistic Regression (IBLR) method [50] is also derived from the family of kNN inspired algorithms. The core idea, is to consider the label information in the neighborhood of a query as "extra features" of that query, and then to treat instance-based learning as a logistic regression problem. For each label λ_j, the algorithm builds a logistic regression classifier $h_j(\cdot)$ according to the model:

$$\log\left(\frac{\pi^{(j)}_{\boldsymbol{x'}}}{1-\pi^{(j)}_{\boldsymbol{x'}}}\right) = \omega^{(j)}_{\boldsymbol{x'}} + \sum_{l=1}^{m} \alpha^{(j)}_l \cdot \omega^{(j)}_{+l}(\boldsymbol{x}), \quad (5)$$

where $\pi^{(j)}_{\boldsymbol{x'}}$ denotes the (posterior) probability that $\lambda_j \in \mathcal{L}$ is relevant for $\boldsymbol{x'}$, $\omega^{(j)}_{x'}$ is a bias term, $\alpha^{(j)}_l$ denotes a coefficient indicating to what extent the relevance of λ_j is influenced by the relevance of λ_l, and $\omega^{(j)}_{+l}(\boldsymbol{x})$ is a summary of the presence of label λ_l in the neighborhood of $\boldsymbol{x'}$, $\mathcal{N}_k(\boldsymbol{x'})$, defined by:

$$\omega^{(j)}_{+l}(\boldsymbol{x'}) = \sum_{\boldsymbol{x} \in \mathcal{N}_k(\boldsymbol{x'})} h_l\left(\boldsymbol{x}\right) \ . \quad (6)$$

Here, $h_l(\boldsymbol{x})$ is 1 if and only if λ_l is associated with $\boldsymbol{x}$, and 0 otherwise. The main advantage of IBLR over ML-kNN is that the former attempts to take into account label inter-dependencies arising by the estimation of regression coefficients.

C. Ensemble Methods

Ensemble of Classifier Chains (ECC) [46] has established itself as a powerful learning technique with modest computational complexity. It is based on the successful Classifier Chains (CC) model [46], which involves the training of m binary classifiers, similar to BR methods. However, unlike the naive BR scheme, in CC binary classifiers are linked along a "chain" so that each classifier is built upon the preceding ones. In particular, during the training phase, CC enhances the feature space of each link in the chain with binary features from ground-truth labeling. Since true labels are not known during testing, CC augments the feature vector by all prior BR predictions. Formally, the classification process begins with h_1 which determines $P(\lambda_1 \mid \boldsymbol{x})$, and propagates along the chain for every following classifier $h_2, \ldots, h_j$ predicting:

$$P(\lambda_j \mid \boldsymbol{x}, \lambda_1, \ldots, \lambda_{j-1}) \rightarrow \lambda_j \in \{0,1\}, j = 2, \ldots, m \ . \quad (7)$$

The binary feature vector $(\lambda_1, \ldots, \lambda_m)$ represents the predicted label set of $\boldsymbol{x}$, $Z_{\boldsymbol{x}}$. Despite the incorporation of label information, the prediction accuracy is heavily dependent on the ordering of the labels, since only one direction of dependency between two labels is captured. To overcome this limitation, ECC extends this approach by constructing multiple CC classifiers with random permutations over the label space. Hence, each CC model is likely to be unique and able to give different multi-label predictions, while a good label order is not mandatory. More specifically, to obtain the output of ECC, a generic voting scheme is applied, where the sum of the predictions is calculated per label, and then a threshold t_s is applied to select the relevant labels, such that $\lambda_j \geq t_s$.

Another effective ensemble-based architecture for solving multi-label classification tasks is the Random k-Labelsets (RAkEL) [47], which embodies LP classifiers as base members. The RAkEL system tries to estimate correlations between the labels by training each LP classifier of the ensemble with a small randomly selected (without replacement) k-labelset (*i.e.*, a size-k subset of the set of labels). This randomness is of primary importance in order to guarantee computational efficiency. For a classification of a new instance, each model provides binary predictions for each label λ_j in the corresponding k-labelset. Let E_j be the mean of these predictions for each label $\lambda_j \in \mathcal{L}$. Then, the output is positive for a given label, if the average decision is greater than a 0.5 threshold:

$$Z_{\boldsymbol{x}} = \left\{\lambda_j \mid E_j > 0.5, 1 \leq j \leq m\right\} \ . \quad (8)$$

In other words, when the actual number of votes exceeds half of the maximum number of votes that λ_j receives from the ensemble, then it is regarded to be relevant (majority voting rule). Although RAkEL models label correlations effectively and overcomes the aforementioned disadvantages of the LP transformation, the random selection of subsets is likely to

negatively affect the ensemble's performance, since the chosen subsets may not cover all labels or inter-label correlations [43].

V. Data and Evaluation Description

In this Section, we present the specific sources of data, both satellite- and ground-based, that are used in our analysis. One of the key contributions of this work lies in employing real imaging data provided by the MODIS sensor and real ground-truth land cover data from the EEA. Furthermore, we provide a detailed discussion of the various evaluation metrics that are adopted for the performance quantification of multi-label classification algorithms.

A. MODIS data - Obtaining features

NASA's MODIS Earth Observation System is considered one of the most valuable sources of remote sensing data, aimed at monitoring and predicting environmental dynamics. The MODIS sensor can achieve global coverage with high temporal resolution, scanning the entire Earth's surface (aboard the Terra and Aqua satellites) in $1-2$ days, from an altitude of 705km. MODIS acquires data in 36 spectral bands ranging from $400-14400$nm, where the first two bands have a spatial resolution (pixel size at nadir) of 250m^2, bands 3 to 7 of 500m^2, and the rest bands at 1km^2 approximately. The sensor provides 12 bits radiometric sensitivity and achieves a swath of 2330km (across track) by 10km (along track at nadir). MODIS data are open-access and continuously updated since 2000.

The MODIS land native product files distributed by the Land Processes Distributed Active Archive Center[1] come in the Hierarchical Data Format (HDF) and in Sinusoidal (SIN) projection. As a result, MODIS data are grouped in 460 equal non-overlapping spatial tiles starting at (0,0) in the upper left corner and proceeding to the right (horizontal) and downward (vertical) until the lower right corner at (35,17). Each one of them captures approximately 1200×1200km of real land. Nevertheless, SIN projection is not widely used and thus a common geographic projection is needed for our study. For this reason, we utilized the MODIS Reprojection Tool[2] (MRT), which provides a basic set of routines for transformation of MODIS imagery into standard geographic projections. This way, we re-sampled the original data and changed the projection to Universal Transverse Mercator (UTM) to become compatible with the global coordinate system (WGS 84 datum), which is also adopted by the Global Positioning System (GPS). The area of our interest, shown in Fig. 2, comprises of a central portion of the European continent, namely h19v04 (without portions of Ukraine and Moldova) and h18v04 image tiles.

In order to benefit from the high temporal resolution observations of MODIS, while simultaneously mitigating the low spatial resolution, we consider annual time series to monitor the best possible density and intensity of green vegetation growth. Our model takes into account a well known monitoring indicator for vegetation health and dynamics, namely

[1]https://lpdaac.usgs.gov/

[2]https://lpdaac.usgs.gov/tools/modis_reprojection_tool

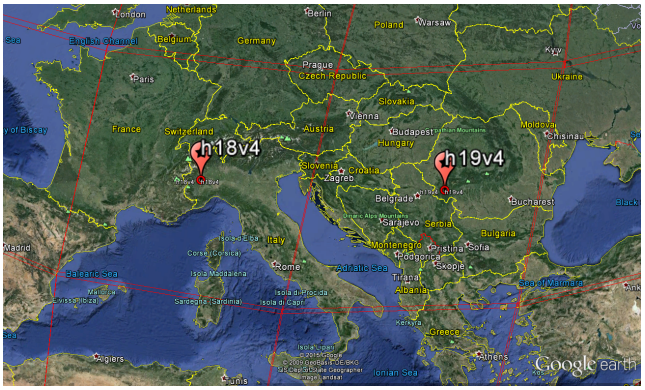


Fig. 2: Geographic distribution of MODIS h18v04 and h19v04 tiles. The h18v04 region captures South-Central Europe, while h19v04 a large part of the Balkans, capturing a diverse set of land cover types.

the Normalized Difference Vegetation Index (NDVI) [51] from the Level-3 product MOD13A1, collection 5 (500m^2 spatial resolution, 16 days temporal granularity). It is empirically related to the reflectance measurements in the red and Near InfraRed (NIR) portion of the spectrum through:

$$\text{NDVI} = \frac{\rho_{\text{NIR}} - \rho_{\text{red}}}{\rho_{\text{NIR}} + \rho_{\text{red}}} \ . \tag{9}$$

Due to the high discriminating capabilities of NIR versus visible wavelength, NDVI is more sensitive than a single wavelength and able to separate very well the living from stressed or dead plantation. Therefore, NDVI carries valuable information regarding surface properties and can effectively quantify the "floral" content of an area, *i.e.*, the chlorophyll concentrations. Furthermore, as a ratio, it has the advantage of minimizing different types of noise (variations in irradiance, clouds, view angles, atmospheric attenuation and even calibration), but it also leads to insensitivities with respect to vegetation variations over certain land cover conditions [52]. NDVI is designed to standardize the vegetation indices values between -1 and $+1$, where higher values indicate more photosynthetically active land cover types. We collected approximately 2 measurements for a ten-month period (March until December) leading to 19 NDVI values/features. For the final data calibration, we refer to the quality assurance metadata [53] supplied with the MOD13A1 product in order to assemble only reliable pixels (i.e., exclude unprocessed data).

Land Surface Temperature (LST) has been proved to play a significant role in detecting several climatic, hydrological, ecological, and biogeochemical changes [14], which are crucial parameters for land cover estimation. LST observations are retrieved from the Thermal InfraRed (TIR) bands and are able to combine the results of all surface–atmosphere interactions and corresponding energy fluxes, measuring the additive compositions of TIR from background soils and overlying vegetation canopy. This way, whereas NDVI measurements estimate efficiently the vegetation cover, LST is more applicable for targets that are not chlorophyll sensitive [54]. The LST data are included in the Level-3 product MOD11A2, which stores the average values during an 8 day period on a 1km^2 SIN grid, leading to 38 values for the period March-December. In

order to obtain the same spatial resolution with MOD13A1, we perform an oversampling to 500m^2 spatial resolution. As a result, we enhance the previously selected 19 features by adding measurements (feature level fusion) related to the LST daytime, extending the number of features to 57.

B. CORINE Land Cover data - Obtaining Labels

The CLC inventory was initiated in 1990 and has been updated in 2000 and 2006, while the latest version of the 2012 update is still under production. CLC consists of 44 classes, including artificial surfaces, agricultural and forest areas, wetlands, and water bodies overall. In this work, we utilize data from 2000[3] and 2006 at 100m^2 resolution (Version 17). The QGIS[4] software is employed in order to transform these raster-based Geographic Information Systems (GIS) measurements to WGS 84 datum in order to become compatible with MODIS data, and subsequently extract the regions corresponding to the h19v04 and the h18v04 tiles.

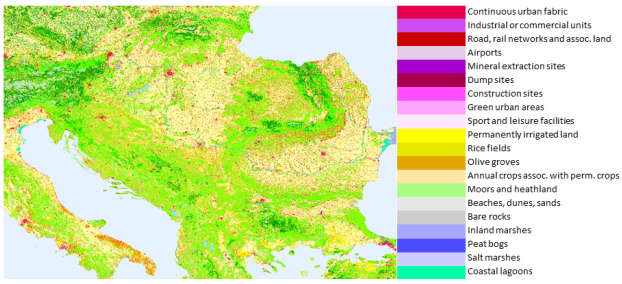


Fig. 3: CLC map and legend for the h19v04 tile of 2000 (CLC2000).

In order to construct the multi-label dataset, the CLC labels matrix was divided into non-overlapping blocks using a 5×5 grid, since the MODIS pixel size is approximately 25 times the size of a CORINE pixel. As a result, a binary vector per sample is produced, where a value of one indicates that a label is present while a value of zero denotes that a label is absent. We select 20 labels as depicted in Table I and exclude examples composed of only one label in order to acquire a challenging scenario for the multi-label learning algorithms.

It is important to highlight that not all multi-label datasets are equal, even if they have the same number of instances or labels. Therefore, to obtain a better understanding of the characteristics of our dataset, we estimated certain statistical metrics [9]. Let $\mathcal{S}$ be the multi-label dataset consisting of $|s|$ multi-label examples $(\boldsymbol{x}_i, Y_i), i = 1, \ldots, |s|$. *Label Cardinality* (LC) calculates the average number of class labels associated with each instance in the dataset: $\text{LC}(\mathcal{S}) = \frac{1}{|s|} \sum_{i=1}^{|s|} |Y_i|$. LC is independent of the number of labels m and it is used to denote the number of alternative labels that characterize the $|s|$ examples of a multi-label dataset. The larger the value of LC, the more difficult is to obtain good classification performance. Besides LC, we also calculated *Label Density* (LD), which is the cardinality normalized by the number of labels m: $\text{LD}(\mathcal{S}) = \frac{1}{|s|} \sum_{i=1}^{|s|} \frac{|Y_i|}{m}$. LD quantifies how dense (or sparse) the multi-label dataset is. Moreover, we consider the *Distinct Labelsets* (DL) metric, which expresses the number of different label combinations observed in the dataset and it is of key importance for methods that operate on label subsets: $\text{DL}(\mathcal{S}) = \left|\{Y_i \mid \exists \boldsymbol{x}_i : (\boldsymbol{x}_i, Y_i) \in \mathcal{S}\}\right|, i = 1, \ldots, |s|$. Table II summarizes the aforementioned statistics for the h19v04 tile of CLC2000 including some benchmark multi-label datasets from a variety of domains along with their corresponding statistics.

TABLE I: Selected ground-truth CLC labels from CORINE.

No.	CLC Code	Description
1	111	Continuous urban fabric
2	121	Industrial or commercial units
3	122	Road & rail networks & assoc. land
4	124	Airports
5	131	Mineral extraction sites
6	132	Dump sites
7	133	Construction sites
8	141	Green urban areas
9	142	Sport and leisure facilities
10	212	Permanently irrigated land
11	213	Rice fields
12	223	Olive groves
13	241	Annual crops assoc. with perm. crops
14	322	Moors and heathland
15	331	Beaches, dunes, sands
16	332	Bare rocks
17	411	Inland marshes
18	412	Peat bogs
19	421	Salt marshes
20	521	Coastal lagoons

TABLE II: Statistical characteristics of the proposed and other publicly available multi-label datasets.

Name (domain)	$\|s\|$	m	LC	LD	DL
land cover (rem. sensing)	12291	20	2.037	0.102	246
yeast [39] (biology)	2417	14	4.237	0.303	198
scene [18] (image)	2407	6	1.074	0.179	15
bibtex [55] (text)	7395	159	2.402	0.015	2856
emotions [17] (music)	593	6	1.869	0.311	27

C. Experimental and evaluation settings

In our analysis, we consider the algorithmic implementations included in the MULAN[5] Java library, an open source platform for the evaluation of multi-label algorithms that works on top of the WEKA[6] framework. We make an initial split of the training to testing examples in the order of 7 : 3, although we are particularly interested in classification with very limited training examples, since obtaining real labeled data is a costly process.

A multi-label classifier produces a set of predicted labels, but many implementations first predict a score for each label, which is then compared to a threshold to obtain the set. Ultimately, there exist two major tasks in supervised learning of multi-label data: *multi-label classification* aiming at producing a bipartition of the labels into a relevant (positive) and an irrelevant (negative) set, and *label ranking* seeking

[3] http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2000-raster-3

[4] http://www.qgis.org/en/site/

[5] http://mulan.sourceforge.net/

[6] http://www.cs.waikato.ac.nz/ml/weka/

to map instances to a strict order over a finite set of predefined labels [44]. Consequently, performance evaluation is significantly more complicated compared to the conventional supervised single-class case and several metrics are required in order to properly evaluate an algorithm. We assume two major performance metric categories: *example-based measures* which are calculated separately for each test example and averaged across the test set, and *label-based measures* which evaluate the system's performance for each label separately, returning the micro/macro-averaged value across all labels [10].

Formally, let $\mathcal{T} = \{(\boldsymbol{x}_i, Y_i) \mid i = 1, \dots, p\}$ be the multi-label evaluation dataset. With respect to the first category of metrics, we consider the following measures:

- *Hamming Loss* calculates the percentage of misclassified example-label pairs, considering the prediction error (an irrelevant label is predicted) and the missing error (a relevant label is not predicted) given by:

$$\text{Hamming Loss} = \frac{1}{p}\sum_{i=1}^{p}\frac{|Y_i \Delta Z_i|}{m}, \tag{10}$$

  where Δ stands for the symmetric difference between the two sets. The value is between 0 and 1, with a lower value representing a better performance. Due to the typical sparsity in multi-labeling, Hamming loss tends to be a lenient metric.
- *Subset Accuracy* evaluates the fraction of correctly classified examples:

$$\text{Subset Accuracy} = \frac{1}{p}\sum_{i=1}^{p} I(|Y_i| = |Z_i|), \tag{11}$$

  where I is the indicator function taking values $I(\text{true}) = 1$ and $I(\text{false}) = 0$. Subset accuracy is a strict accuracy metric, since it classifies a sample as correct if all the predicted labels are identical to the true set of labels.
- *One-error* is a ranking based metric which computes how many examples have irrelevant top-ranked labels according to:

$$\text{One-error} = \frac{1}{p}\sum_{i=1}^{p}\delta(\arg\min_{\lambda\in\mathcal{L}} r_{\boldsymbol{x}_i}(\lambda)), \tag{12}$$

  where $\delta(\lambda) = 1$ if $\lambda \notin Y_i$ and 0 otherwise.
- *Coverage* reports the average distance which needs to be traversed in order to cover all the relevant labels of the example from the ranked label list:

$$\text{Coverage} = \frac{1}{p}\sum_{i=1}^{p}\max_{\lambda\in Y_i} r_{\boldsymbol{x}_i}(\lambda) - 1\ . \tag{13}$$

- *Ranking Loss* evaluates the average fraction of labels pairs that are ordered incorrectly:

$$\text{Ranking Loss} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i||\bar{Y}_i|}\left|\mathcal{G}_i\right|, \tag{14}$$

  where $\mathcal{G}_i$ is a set equal to $\{(\lambda', \lambda'') : r_{\boldsymbol{x}_i}(\lambda') > r_{\boldsymbol{x}_i}(\lambda'')\}$ for $(\lambda', \lambda'') \in Y_i \times \bar{Y}_i$. Here, $\bar{Y}_i$ denotes the complementary set of Y_i with respect to $\mathcal{L}$. In other words, ranking loss measures the ability to capture the relative order between labels.
- *Average Precision* expresses the percentage of labels ranked above a particular relevant label:

$$\text{Av. Prec.} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{|Y_i|}\sum_{\lambda\in Y_i}\frac{\left|\{\lambda'\in Y_i : r_{\boldsymbol{x}_i}(\lambda') < r_{\boldsymbol{x}_i}(\lambda)\}\right|}{r_{\boldsymbol{x}_i}(\lambda)}\ . \tag{15}$$

From information retrieval, we know that the evaluation metrics for a binary classification problem are based on the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) test samples. Based on the above, one can compute *Precision* as TP/(TP+FP), *Recall* as TP/(TP+FN), and the *F-Measure* as the harmonic mean between precision and recall. Extending this concept to multi-label problems, we can derive the corresponding quantities for each label λ_j calculating the micro-averaging operation [56]: $B_{\text{micro}} =$

$$B\left(\sum_{j=1}^{m}\text{TP}_{\lambda_j}, \sum_{j=1}^{m}\text{TN}_{\lambda_j}, \sum_{j=1}^{m}\text{FP}_{\lambda_j}, \sum_{j=1}^{m}\text{FP}_{\lambda_j}\right), \tag{16}$$

as well as the macro-averaging operation [56]:

$$B_{\text{macro}} = \frac{1}{m}\sum_{j=1}^{m} B(\text{TP}_{\lambda_j}, \text{TN}_{\lambda_j}, \text{FP}_{\lambda_j}, \text{FN}_{\lambda_j}), \tag{17}$$

where B is one of the previous mentioned classification metrics, and TP_{λ_j}, TN_{λ_j}, FP_{λ_j}, FN_{λ_j} is the value of TP, TN, FP and FN after the binary evaluation for λ_j. Conceptually speaking, micro-averaging gives equal weight to each example and is an indicator of large classes, whereas macro-averaging to each label and gives a sense of effectiveness on small classes [57].

Finally, we consider the *Area Under the Curve* (AUC) metric which is calculated from the Receiver Operating Characteristic (ROC) curve. In case all annotations contain confidence values, the AUC score describes the overall quality of performance, independently of individual threshold configurations regarding specific trade-offs between TP and FP [58]. More precisely, let the True Positive Rate (TPR) be defined as TP/(TP+FP) and the False Positive Rate (FPR) as FP/(FP+TN). Then, each point on the ROC curve corresponds to a pair (TPR, FPR) for one threshold, and the area under this ROC curve is called micro-AUC, derived as:

$$\text{AUC}_{\text{micro}} = \frac{\left|\{(\boldsymbol{x'}, \boldsymbol{x''}, \lambda', \lambda'') \mid r_{\boldsymbol{x'}}(\lambda') \geq r_{\boldsymbol{x''}}(\lambda'),\}\right|}{|\mathcal{R}^+||\mathcal{R}^-|}, \tag{18}$$

for $(\boldsymbol{x'}, \lambda') \in \mathcal{R}^+$, and $(\boldsymbol{x''}, \lambda'') \in \mathcal{R}^-$, where $\mathcal{R}^+ = \{(\boldsymbol{x_i}, \lambda) \mid \lambda \in Y_i, 1 \leq i \leq p\}$ corresponds to the set of relevant, and $\mathcal{R}^- = \{(\boldsymbol{x_i}, \lambda) \mid \lambda \notin Y_i, 1 \leq i \leq p\}$ to the set of irrelevant labels [10]. Subsequently, the macro-averaged AUC is the average AUC of the separate ROC curves for each class and can be defined as follows: $\text{AUC}_{\text{macro}} =$

$$\frac{1}{m}\sum_{j=1}^{m}\frac{\left|\{(\boldsymbol{x'}, \boldsymbol{x''}) \mid r_{\boldsymbol{x'}}(\lambda_j) \geq r_{\boldsymbol{x''}}(\lambda_j), (\boldsymbol{x'}, \boldsymbol{x''}) \in \mathcal{Z}_j \times \bar{\mathcal{Z}}_j\}\right|}{|\mathcal{Z}_j||\bar{\mathcal{Z}}_j|}, \tag{19}$$

where $\mathcal{Z}_j = \{\boldsymbol{x}_i \mid \lambda_j \in Y_i, 1 \leq i \leq p\}$ is the set of test instances with label $\lambda_j \in \mathcal{L}$, and $\bar{\mathcal{Z}}_j = \{\boldsymbol{x}_i \mid \lambda_j \notin Y_i, 1 \leq i \leq p\}$

is its complementary set of test instances without $\lambda_j \in \mathcal{L}$. A value of 1 corresponds to a perfect system.

VI. Experimental Results

In this Section, we examine the performance of each approach under the following challenging scenarios: (a) classification performance with respect to the number of training examples; (b) classification performance when the training data correspond to a specified *geographic region* and the testing data come from a neighboring region; and (c) classification performance when the training data correspond to a specified *time instance* and the testing data come from the same location but another instance. Each experiment has its own distinct value, whereas the objective is to evaluate the multi-label classification framework when applied under real-life conditions where limited training data are available.

A. Classification performance with respect to training set

The objective of the first set of experiments is to evaluate the generalization capabilities of each learning algorithm. To that end, we evaluate the performance of each method as a function of the number of training examples using the NDVI and the LST features. This is a critical parameter, since it is directly related to the cost and manpower required for the classification and understanding of newly acquired remotely sensed images. We consider a varying number of training examples ranging from 25 to 5000, averaged over 10 realizations.

We selected the C4.5 [59] DT learning algorithm as the base-level single-label classifier in all problem transformation and ensemble techniques, while individual parameters of each method were instantiated according to recommendations from the literature. In specific, the ML-kNN and IBLR algorithms are parametrized by the size of the neighborhood, for which we adopted the value of $k = 10$, while ML-kNN needs further a smoothing parameter γ controlling the effects of priors on the estimation, where we select a value of $\gamma = 1$ leading to a Laplace smoothing prior [41]. For the ensemble methods, the key parameter is the number of component classifiers (models), whereas RAkEL also requires the definition of the size of the labelsets. The number of models was set to 10 for ECC, and to $2m$ for RAkEL with a size-3 subset.

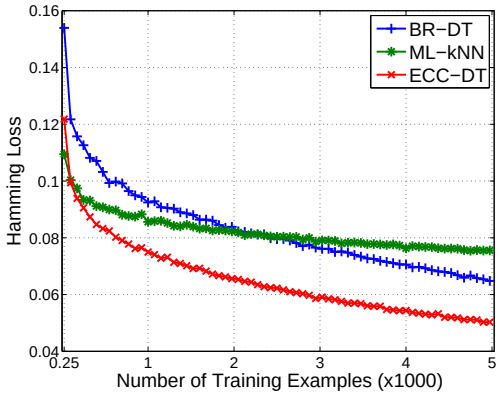
In Fig. 4, we present the performance of multi-label classification for tiles h19v04 and h18v04 using data from the CLC2000 inventory, where the Hamming loss and the micro-averaged AUC associated with the BR-DT, the ML-kNN, and the ECC-DT algorithms are presented (one algorithm from each category). These two metrics are highly representative, since the Hamming loss belongs to the example-based metrics and can give us an overall intuition of the misclassified instance-label pairs, whereas the AUC is a label-based metric evaluating the quality of predictions for each label independently.

One can observe that for both metrics the performance increases monotonically with the gradual increase of the training set size, with a fast rate during initial cases and then with a slower one. These results indicate that the algorithms indeed learn and exploit information from the training data in order to ameliorate their predictions. Looking closer at each metric, we observe that the performance of the three classifiers according to the AUC is quite stable across tiles examined on the same labels (especially for a large number of training examples), whereas slight differences are attributed to the variation of the intrinsic spatio-spectral characteristics of each tile. Analogous results arise when considering the Hamming loss metric as well. In this case, we further observe that when a large number of training examples is employed, BR-DT outperforms ML-kNN, indicating that in some cases "letting the data speak for themselves" can allow naive algorithms to beat more complex approaches. Overall, ECC-DT outperforms the other two algorithms in this experimental setup, partly due to its internal mechanisms that benefit from label dependencies. This behavior has been also observed in other scenarios of multi-label classification [38].

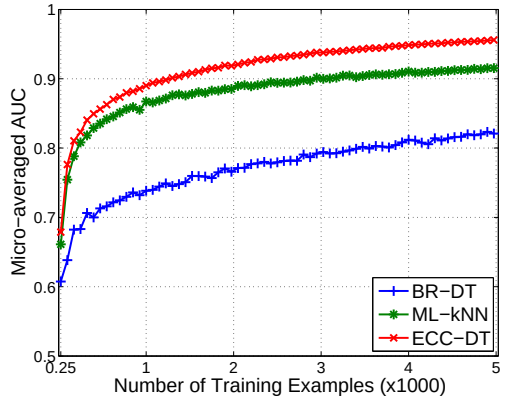
Considering the h19v04 tile of CLC2000 as our reference region, an overview of the performance is presented in Table III, where we have included all the evaluation metrics. For the evaluation of experiments, we performed 10 different 10-fold cross validation experiments and report the average results over these 100 executions. Considering the problem transformation methods, we observe that BR-DT is better than LP-DT in all exampled-based metrics except subset accuracy, which is notably high among all methods, suggesting that LP-DT is able to faithfully capture the underlying statistics of the labels. For the label-based metrics, the results are more balanced, since BR-DT achieves superior precision and F-measure, whereas LP-DT is slightly better with respect to AUC. Regarding the algorithm adaptation methods, we observe that IBLR has a small lead, but overall ML-kNN and IBLR are on equal footing, since the observed variation in prediction accuracy manifested in most of the evaluation metrics is of limited statistical significance. Last, analyzing the ensemble methods, RAkEL-DT has a clear advantage when it comes to metrics such as subset accuracy and recall, however, ECC-DT achieves superior performance for precision and AUC.

In general, one can argue that the ensemble methods confirmed their reputation as one of the most powerful class of multi-label classification algorithms, since they achieve a better and more robust performance compared to other methods. On the opposite, the higher performance comes at a significant higher computational cost, as it is shown in the runtimes reported in Table III on a typical workstation. Between the remaining two categories, *i.e.*, algorithm adaptation and problem transformation methods, results are balanced and largely dependent on the metric which one seeks to optimize.

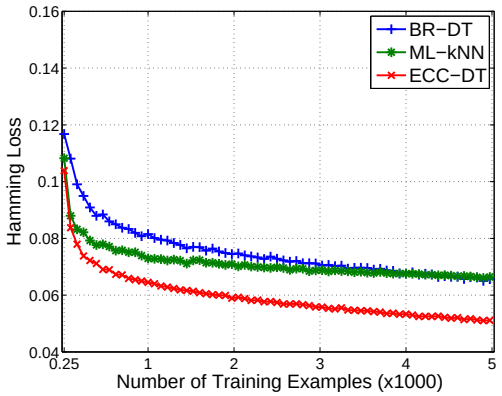
To better demonstrate the behavior of each algorithm and how differences in error metrics are perceived, we introduce the "*multi-label confidence map*". Each row of the map corresponds to a specific label (CLC code), while columns encode particular examples, *i.e*, spatial locations. Fig. 5 presents the ground-truth multi-label map for h19v04 of CLC2000. This image is a binary matrix where a value of 1 indicates the presence of a specific label, while 0 denotes the absence of the label. For instance, considering the first example (column), labels 1 and 8 are active, indicating that CLC labels with codes 111 and 141 exist in this pixel.



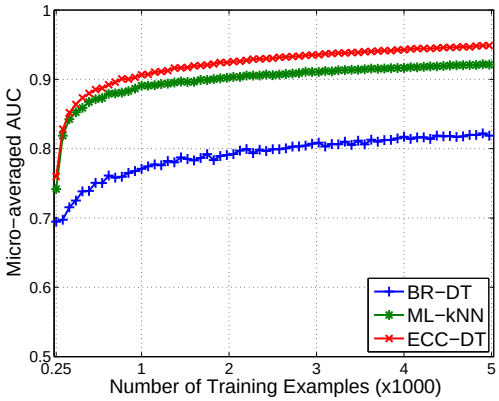
(a) Hamming Loss for h19v04 tile of CLC2000



(b) Micro AUC for h19v04 tile of CLC2000



(c) Hamming Loss for h18v04 tile of CLC2000



(d) Micro AUC for h18v04 tile of CLC2000

Fig. 4: Classification performance w.r.t. the number of training examples for two different tiles, where the complex interactions between training data size and classification performance are illustrated. In general, the performance gains are more dramatic when increasing smaller sets of training samples, while the benefits of introducing more training data are moderate.

TABLE III: Performance (mean $\pm$ std) of each multi-label learning algorithm over 10 different 10-fold cross validation experiments. For each metric, $\uparrow$ indicates "higher the better", whereas $\downarrow$ indicates "lower the better". Ensemble methods perform overall better than problem transformation and algorithm adaptation techniques.

Measure	Multi-Label Learning Algorithm					
	BR-DT	LP-DT	ML-kNN	IBLR	RAkEL-DT	ECC-DT
Hamming Loss $\downarrow$	0.044 ± 0.002	0.047 ± 0.002	0.067 ± 0.001	0.067 ± 0.001	$\mathbf{0.030 \pm 0.002}$	0.033 ± 0.001
Subset Accuracy $\uparrow$	0.496 ± 0.013	0.626 ± 0.015	0.305 ± 0.010	0.318 ± 0.014	$\mathbf{0.638 \pm 0.014}$	0.597 ± 0.013
One-error $\downarrow$	0.189 ± 0.013	0.289 ± 0.015	0.269 ± 0.015	0.268 ± 0.012	$\mathbf{0.098 \pm 0.010}$	0.108 ± 0.009
Coverage $\downarrow$	4.793 ± 0.236	5.743 ± 0.176	2.987 ± 0.073	3.008 ± 0.054	2.436 ± 0.179	$\mathbf{1.939 \pm 0.075}$
Ranking Loss $\downarrow$	0.120 ± 0.008	0.177 ± 0.008	0.072 ± 0.004	0.073 ± 0.003	0.044 ± 0.006	$\mathbf{0.030 \pm 0.002}$
Average Precision $\uparrow$	0.802 ± 0.010	0.742 ± 0.012	0.762 ± 0.009	0.763 ± 0.006	$\mathbf{0.900 \pm 0.008}$	0.896 ± 0.005
Macro Precision $\uparrow$	0.770 ± 0.013	0.731 ± 0.015	0.679 ± 0.033	0.669 ± 0.019	0.874 ± 0.010	$\mathbf{0.897 \pm 0.015}$
Macro Recall $\uparrow$	0.724 ± 0.016	0.729 ± 0.013	0.419 ± 0.008	0.451 ± 0.016	$\mathbf{0.772 \pm 0.011}$	0.692 ± 0.011
Macro F-Measure $\uparrow$	0.743 ± 0.013	0.727 ± 0.012	0.483 ± 0.012	0.520 ± 0.016	$\mathbf{0.814 \pm 0.009}$	0.766 ± 0.011
Macro AUC $\uparrow$	0.864 ± 0.007	0.878 ± 0.010	0.913 ± 0.005	0.919 ± 0.005	0.953 ± 0.005	$\mathbf{0.968 \pm 0.003}$
Micro Precision $\uparrow$	0.795 ± 0.009	0.768 ± 0.013	0.746 ± 0.014	0.735 ± 0.010	0.881 ± 0.009	$\mathbf{0.897 \pm 0.006}$
Micro Recall $\uparrow$	0.768 ± 0.011	0.766 ± 0.012	0.516 ± 0.013	0.531 ± 0.009	$\mathbf{0.820 \pm 0.011}$	0.761 ± 0.011
Micro F-Measure $\uparrow$	0.782 ± 0.009	0.767 ± 0.012	0.610 ± 0.009	0.616 ± 0.009	$\mathbf{0.850 \pm 0.009}$	0.823 ± 0.008
Micro AUC $\uparrow$	0.882 ± 0.008	0.891 ± 0.008	0.942 ± 0.003	0.940 ± 0.002	0.967 ± 0.004	$\mathbf{0.980 \pm 0.002}$
Training Time (sec)	48.19 ± 1.800	$\mathbf{19.46 \pm 0.564}$	37.40 ± 2.628	49.03 ± 1.855	243.3 ± 5.418	761.8 ± 134.8
Testing Time (sec)	0.436 ± 0.035	$\mathbf{0.378 \pm 0.094}$	4.220 ± 0.333	4.730 ± 0.472	0.507 ± 0.058	0.935 ± 0.178

In Figures 6 and 7, we visualize the performance of the BR-DT, ML-kNN, and ECC-DT classifiers with the help of the multi-label confidence map, where each pixel in the map takes a confidence value ranging from 0 to 1 (results averaged over 50 realizations and then scaled to $[0, 1]$ interval). Values closer to 0 indicate that the label is less likely to be enabled, whereas values closer to 1 indicate that this label has a higher probability of being active.

Using these maps, we can visually verify the performance due to the use of more training examples, by examining for instance the label 2 (second row) associated with the CLC code 121. When the algorithms utilize 128 training examples, they assume that almost all samples have this label enabled, something that is not in accordance with the ground-truth data presented in Fig. 5. On the contrary, when the algorithms use 1024 training examples, we can see that their revised predictions become more accurate and reliable. This observation suggests that for the specified label many false positive examples arise. False positives are taken into account by the precision metric, where BR-DT and ECC-DT outperform ML-kNN for this number of training examples. Another illustrative paradigm occurs when we take into account the label 20 with CLC code 521 (last row). In this case, we observe that the classification algorithms cannot detect that the specified label is active in some pixels when presented with 128 training examples, however, the prediction improves dramatically with 1024 training examples. In other words, we have the case of recall and false negatives, where we show that the BR-DT and the ECC algorithms achieve almost similar performance whereas ML-kNN exhibits a significant performance lag.

B. Classification on different spatial regions and temporal instances

In this Section, we examine the performance of the algorithms when the training examples are acquired from a specified region at a given year, while testing takes place either on a neighboring region, or on a different time instance. We initially examine the classification efficiency in a neighboring geographic region, since accurate prediction of the labels of another region suggests that we can leverage training examples of a particular label to evaluate its presence in unexplored locations, avoiding the high cost of hand-collecting new annotated training examples. We consider an experimental setup where three different types of training sets are used, namely a training set from the same tile (h18v04), a training set from another tile (h19v04), and a mixed training set containing all training examples from the reference tile (h19v04) and only a few (*i.e.*, 1024) training examples from the target tile (h18v04). The ECC-DT ensemble classifier was selected for this set of experiments.

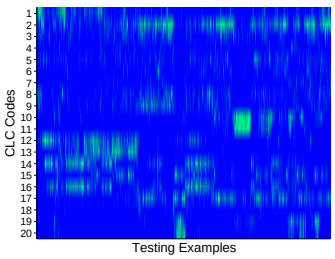
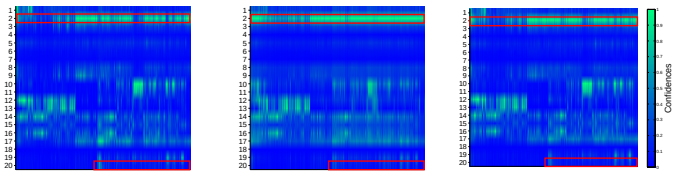
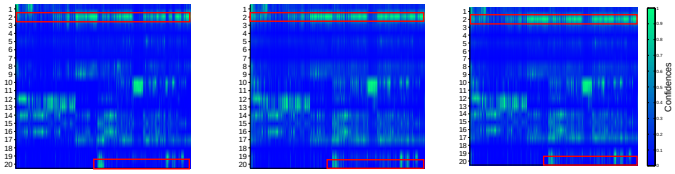


Fig. 5: Ground-truth multi-label map for h19v04 of CLC2000 corresponding to a binary matrix indicating which labels are active for each example, *i.e.*, spatial location. Each horizontal line corresponds to a specific label as illustrated in Table I, while each vertical line corresponds to a specific testing example out of the 3687.



(a) BR-DT (b) ML-kNN (c) ECC-DT

Fig. 6: Multi-label confidence maps for h19v04 of CLC2000 with 128 training samples. The red boxes outline areas where there is significant deviation between the predicted and the ground-truth labels. They highlight that some labels in classification are more sensitive than the others.



(a) BR-DT (b) ML-kNN (c) ECC-DT

Fig. 7: Multi-label confidence maps for h19v04 of CLC2000 with 1024 training samples. Similar to before the red boxes outline areas where there is significant deviation between the predicted and the ground-truth labels. Comparing to the case of 128 training examples shown above, we observe less errors according to the ground-truth map in Fig. 5.

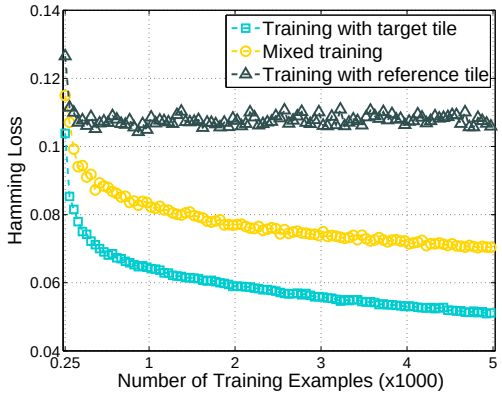


Fig. 8: Classification performance with respect to the amount of training data for a tile originating from a different spatial location using ECC-DT. The results indicate that training using data from different spatial locations can have a dramatic effect on performance, while exploiting a mixed training set composed of data from both the corresponding location as well as from a different location can achieve very high performance.

Analyzing Fig. 8, we observe that when the training and the testing sets are associated with the same tile, a high classification performance is achieved. Naturally, there is significant degradation in performance when the training set is associated to another tile. We observe that although the performance improves initially, it soon reaches a performance plateau which is significantly worse than when the data from the same tile are used in both training and testing. This is where the third

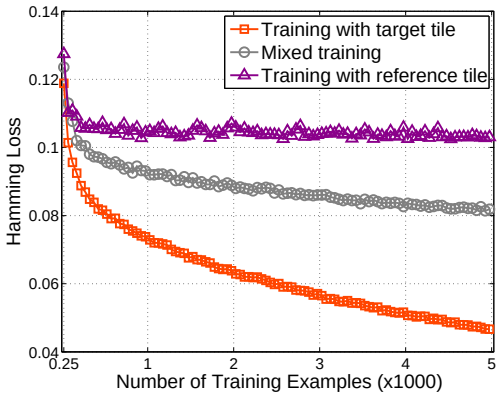


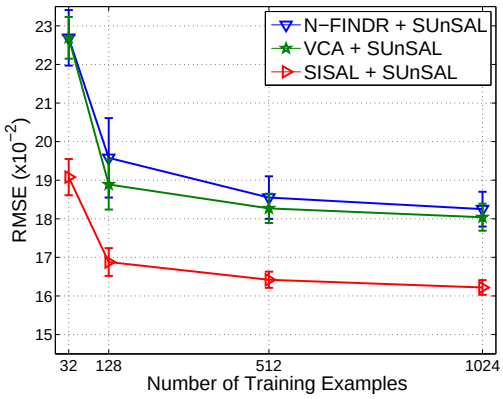
Fig. 9: Classification performance with respect to amount of training data from a single tile at different time instances using ECC-DT. Similar to Fig. 8, the performance significantly improves by considering mixed training conditions.

described training set comes into play. As we can see, the performance in mixed training conditions is quite close to the performance achieved in the benchmark case. This behavior suggests that one can exploit already acquired annotated data, limiting the effort required for collecting new labeled data, and still achieve a high classification performance.

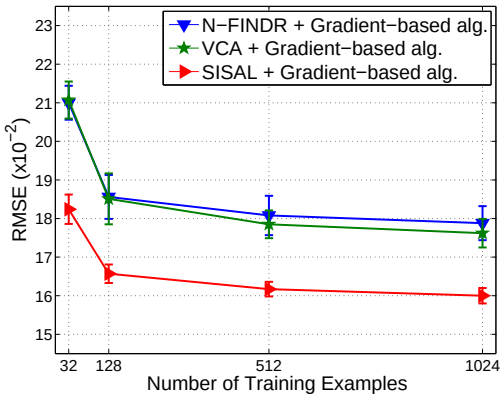
Fig. 9 examines the predicting performance for data from the h19v04 tile in CLC2006 under three different training sets, namely using a training set from the same tile (h19v04) in the same year (2006), using a training set from the same tile (h19v04) in another year (2000), and a training set composed of all data from the reference tile enhanced by 1024 training examples from the target tile. In this case, the objective is to forecast the presence/absence of specific labels in order to understand the temporal evolution of land cover for this region. This is an immensely important scenario, since obtaining up-to-date field-based annotation is extremely challenging, causing very low update rates that characterize the CLC. The problem holds for land cover maps in general, leading to data that are outdated at release time. Similar to the previous case, we observe that the prediction performance when utilizing examples from the reference tile reaches a plateau for the two metrics, but the performance gradient is smoother when using our proposed mixed training approach.

C. Comparison with spectral unmixing

Spectral unmixing and multi-label classification in remote sensing can both operate under the scenario that an observed spectral vector can be actually composed of one or more materials (in contrast to single-label classification). Nevertheless, a direct comparison between the two methods is very difficult for various reasons. Firstly, spectral unmixing is an unsupervised method whereas multi-label classification adheres to the supervised learning paradigm and strongly utilizes the provided labels. Furthermore, the objective of spectral unmixing is the estimation of the abundance of each endmember in an observed spectral vector, while multi-label classification aims at estimating a bipartition and a ranking of all labels. With the above in mind, we proceed to the comparison between spectral unmixing and multi-label classification for real remotely sensed multispectral data. The algorithms were supplied with a priori knowledge regarding the number of endmembers, which is assumed to be equal to the number of the labels m, in order to be able to compare with the ground-truth. Moreover, in order to satisfy the sum-to-one constraint, we convert the 1's indicating the label existence to probabilities that sum up to one (*i.e.* if two labels are present in a pixel, we assign to the corresponding positions the value of 0.5). The authors' implementation with suggested settings were used for all algorithms.



(a) SUnSAL assumes LMM



(b) Gradient-based algorithm assumes PPNMM

Fig. 10: RMSE with respect to the number of examples for h19v04 of CLC2000 over 30 realizations. The approximation for all the examined unmixing chains improves, suggesting the considered dataset can be used for unmixing purposes.

In order to unmix the reference tile h19v04 of CLC2000, we initially have to decompose the measurements into a library $\mathbf{A} \in \mathbb{R}^{d \times m}$, where d is the number of bands/features and m is the number of endmembers/labels. In this step, three state-of-the-art algorithms are considered, namely the N-FINDR, the VCA, and the SISAL. For the fractional abundance estimation we evaluate two state-of-the-art methods, namely the SUnSAL, which uses sparse regression under the LMM, and a gradient-based algorithm developed in [37] which assumes the PPNMM. The quality of the unmixing procedure is measured by comparing the estimated $\hat{\boldsymbol{a}}$ and the "actual" abundance vector $\boldsymbol{a}$, in terms of the error defined by the

Root Mean Square Error (RMSE) $= \sqrt{\frac{1}{mp}\sum_{i=1}^{p}\|\boldsymbol{a}_i - \hat{\boldsymbol{a}}_i\|^2}$, where $\boldsymbol{a}(i)$ and $\hat{\boldsymbol{a}}(i)$ is the actual and the estimated abundance vectors of the i-th testing pixel.

The performance of unmixing using the different algorithms is shown in Fig. 10. We observe that the error for all the unmixing chains reduces with respect to the number of training examples, suggesting that the proposed ground-truth based labeled dataset can be used for unmixing tasks. More specifically, the SISAL method, which does not rely on the pure pixel assumption achieves a lower RMSE, which is also characterized with a lower variance. In addition, the gradient-based algorithm assuming the PPNMM captures better the existing nonlinearities and leads to a better approximation of $\boldsymbol{a}$ than SUnSAL, especially for a small number of training examples. A possible explanation of this behavior, following the reasoning in [60], is that the LMM assumption may be inappropriate for images containing sand, mineral mixtures, trees and vegetation areas, elements that are all contained in the selected labels (CLC codes $331, 131, 141, 223, 241$).

Given the best performing unmixing strategy, *i.e.*, SISAL for endmember extraction and the gradient-based algorithm for abundance estimation, we proceed to a comparison between spectral unmixing and multi-label classification versus the ensemble methods, utilizing multi-label classification metrics. In order to produce a binary predictions matrix strictly encoding the presence or absence of a label, one must convert the positive-valued abundances to binary values. To achieve this, we performed a sorting of the estimated abundances in a descending order and selected only the endmembers that exceed a threshold T. All the corresponding estimated abundances above this threshold are set to 1 while the rest are set to 0. Table IV presents the experimental results with respect to the threshold (50% and 95%) and the number of training examples.

TABLE IV: Performance (mean $\pm$ std) of the ensemble versus unmixing methods over 30 realizations.

Measure	# Tr.	Spectral Unmixing		Multi-Label Classification	
		$T = 50\%$	$T = 95\%$	RAkEL-DT	ECC-DT
Hamming Loss ↓	128	0.24 ± 0.01	0.59 ± 0.02	0.11 ± 0.00	0.10 ± 0.00
	1024	0.26 ± 0.02	0.71 ± 0.03	0.08 ± 0.00	0.07 ± 0.00
Micro Precision ↑	128	0.09 ± 0.02	0.10 ± 0.01	0.47 ± 0.02	0.55 ± 0.02
	1024	0.10 ± 0.02	0.10 ± 0.01	0.67 ± 0.01	0.72 ± 0.01
Micro Recall ↑	128	0.16 ± 0.03	0.61 ± 0.05	0.33 ± 0.02	0.28 ± 0.02
	1024	0.19 ± 0.04	0.74 ± 0.06	0.50 ± 0.01	0.43 ± 0.01

Overall, Table IV demonstrates that the multi-label methods are considerably better than the spectral unmixing ones in terms of the classification measures. Regarding the performance of unmixing with respect to the selected threshold, Table IV demonstrates that increasing the threshold leads to higher Hamming error, and that it dramatically increases the recall, due to the fact that larger values of the threshold produce a larger number of false positives and a lower number of false negatives. With respect to the number of training examples, we observe that only the recall metric is increased suggesting that the architecture is able to capitalize on the training examples by identifying a larger portion of true labels.

A higher level snapshot of the each method's behavior can be obtained by comparing the Hamming loss metric, which shows that the percentage of misclassified example-label pairs is much higher for unmixing compared to the ensemble multi-label learning algorithms. In general, the results presented in Table IV demostrate that multi-label classifiers, even if they are not able to produce fractional abundance estimations, achieve much higher and more robust binary predictions, even under noisy environments.

D. Applying the scheme with hyperspectral data

Hyperspectral imaging platforms can provide finer spatial resolution imagery than multispectral systems, typically at the cost of a smaller field-of-view. As a result, they are limited in their capacity to provide global land cover estimation. For instance, the Hyperion sensor aboard EO-1 has a spatial resolution of 30m^2, acquiring images at 242 spectral bands, however, it does not provide global coverage. As a consequence, we cannot directly introduce the concept of multi-label classification of Hyperion imagery by utilizing CORINE land cover data which have a spatial resolution of 100m^2.

Nowadays however, novel datasets have been compiled that provide ground-truth data at a much higher spatial resolution than 30m^2. Such data do not consider widespread coverage (*e.g.*, whole continents like Europe) as the process of labeling is extremely costly and time-consuming. Nevertheless, they do provide detailed maps of more specific geographic areas (*e.g.*, cities, forests, etc.). For instance, a high resolution land cover dataset for New York City (NYC) of 2010 with a spatial resolution of 1m (3 feet) has been recently released[7]. We investigate the application of our scheme with the NYC dataset combined with the Hyperion data, where we consider the study area encoded as EO1H0130322010245110KF_SGS_01 by Hyperion from September 2, 2010, provided in GeoTIFF format. In Fig. 11 we consider the performance of four multi-label algorithms by utilizing the 198 calibrated bands from Hyperion.

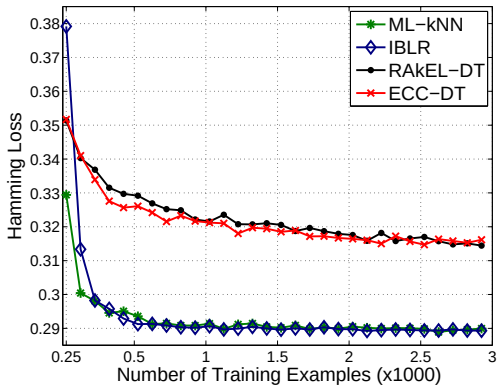


Fig. 11: Performance with respect training examples using hyperspectral data of the Hyperion sensor.

The results suggest that multi-label classification is also a viable choice for the exploitation of hyperspectral data for land

[7]https://nycopendata.socrata.com/Environment/Landcover-Raster-Data-2010-/9auy-76zt

cover estimation. Observing the achieved performance with the performance in the case of multispectral data, we note that for the NYC dataset with hyperspectral data, performance is worse compared to CLC prediction with multispectral data. However, we should note that the results are not directly comparable due to many differences in datasets: different types and number of labels, distinct intrinsic characteristics of the regions (topologies), different years, and different types of features that are extracted from hyperspectral compared to multispectral imagery.

Apart from these significant reasons, additional effects come into play in the case of multi-label classification of hyperspectral data. First of all, we could only manage to find one scene from the whole 2010 to characterize just the same small portion of the NYC. On the other side, MODIS has a high temporal resolution with a sun-synchronous orbit and thus we are able to generate the time-series which is more appropriate for capturing the variation of land cover characteristics through a whole year. As a conclusion, it is not only the spectral resolution which is critical for the classification quality, but also the temporal resolution, especially for land cover estimation.

A second important parameter is the ratio of scale incompatibility. Whereas each pixel of MODIS is approximately 25 times the size of the CORINE, in the NYC land cover case the Hyperion pixel is approximately 900 times the land cover pixel, since the NYC dataset has a spatial resolution of $1m^2$ and the Hyperion sensor of $30m^2$. The dramatic change in scale, is definitely a key factor that highly affects the performance of classification.

Comparing the performance of the different classification schemes, we observe that algorithm adaptation methods, *i.e.*, ML-kNN and IBLR, achieve better performance compared to the RAkEL-DT and ECC-DT ensemble methods. A reason for this behavior is that the performance of the classifiers is directly related to a well-known problem in estimation theory and machine learning, the curse of dimensionality, whereby increasing the dimensionality of the data space, makes the processes of data modeling more challenging due to the sparse coverage of high-dimensional spaces with limited examples. The problem in multi-label learning is even bigger, since features represent all the classes of the whole dataset, whereas many of them are not relevant to a specified class.

E. Parameter sensitivity analysis

In order to provide a comprehensive analysis, in this subsection we investigate the effects and influence in performance attributed to the parameter selection process of each considered algorithm. To minimize the effects introduced by other sources of variation, we fix the number of training examples to 1024 and report results averaged over 50 realizations in order to obtain an informed view of the sensitivity of each method.

Regarding the kNN-based methods, the key parameter that must be defined concerns the number of neighbors that are employed. We observe in Fig. 12 that the worst choice for Hamming loss corresponds to using 30 neighbors, while the point of optimal performance is attained with 5 neighbors for both methods. The results suggest that further increasing the number of the neighbors leads to performance deterioration, since valuable information is replaced with noise obtained, in addition to the computational overhead. Between the two classifiers, we note that IBLR has a more robust behavior compared to ML-kNN.

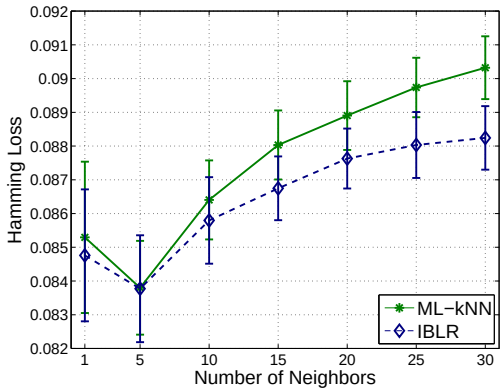


Fig. 12: Performance with respect to the number of neighbors for h19v04 of CLC2000 with 1024 training examples. Both adaptation algorithms exhibit similar performance with respect to this parameter, however IBLR has a slightly higher and more robust behavior compared to ML-kNN.

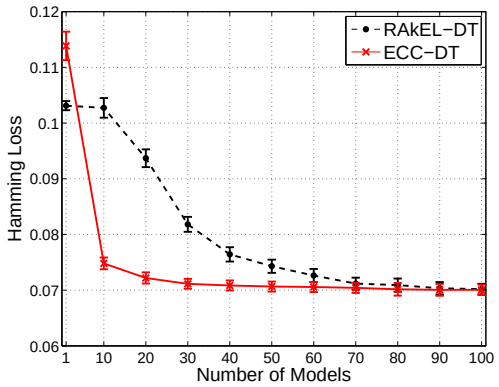


Fig. 13: Classification performance with respect to the number of models for h19v04 of CLC2000 with 1024 training examples. Varying the number of models has a major effect on the classification performance of ensemble methods where ECC achieves superior performance compared to RAkEL.

Considering the powerful class of ensemble techniques, we investigate how the number of component base classifiers involved in the chain affects the performance. As illustrated in Fig. 13, ECC-DT and RAkEL-DT differ significantly with respect to the internal design, since the former achieves a performance close to optimal with a small number of classifier chain models, whereas RAkEL-DT is learning progressively with an increased number of models. With an adoption of a large number of models (more than 60), we can observe that the RAkEL-DT approximates the performance achieved by ECC-DT. However, the superiority of ECC-DT for multi-label classification with land cover data is particularly evident when a small number of training examples is considered.

VII. Conclusions

In this work, we presented a radically different approach in satellite-based land cover identification, where we cast the problem as an instance of multi-label learning. Multi-label classification in this specific domain provides supplementary solutions to the important problem of spectral unmixing, however, unlike state-of-the-art schemes, the proposed formulation utilizes publicly available labels in conjunction with contemporary satellite data, and provides a real-world answer to maintaining up-to-date land cover maps.

We considered an extensive set of experiments, employing state-of-the-art multi-label learning algorithms under diverse and challenging scenarios. The experimental results suggest that a small number of training examples is sufficient for achieving satisfying performance in the situation where training and testing data from a specified region on a given time instance is considered. However, the performance deteriorates when testing takes place on a different spatial region or from another instance in time. We demonstrate that by encompassing a limited number of examples of the target-tile at the target-time, the performance improves remarkably, offering a solid answer to the issues related to the cost and time required for gathering annotated ground-truth data. It should be noted that the proposed formulation can fully exploit the existence of ground-truth data, which means that this approach cannot be applied in cases where labeled data are unavailable, *e.g.*, unmixing of the Mars surface data.

In addition to the value of this work in the remote sensing community, we have also effectively introduced a new class of datasets composed of satellite and geographic data, offering the research community the possibility to evaluate different multi-label classification schemes on alternative remote sensing datasets, which provide a more appropriate formulation compared to the single-label cases that have been explored in the literature so far.

Acknowledgment

This work was partially funded by the PHySIS project (contract no. 640174) and the DEDALE project (contract no. 665044) within the H2020 Framework Program of the European Commission.

References

[1] A. Di Gregorio, *Land Cover Classification System—Classification concepts and user manual for Software version 2*. Rome, Italy: Food and Agriculture Organization of the United Nations, 2005.

[2] I. McCallum, M. Obersteiner, S. Nilsson, and A. Shvidenko, "A spatial comparison of four satellite derived 1km global land cover datasets," *Int. J. of App. Earth Observation and Geoinformation*, vol. 8, no. 4, 2006.

[3] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International journal of Remote sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[4] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. Atli Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *Signal Processing Magazine, IEEE*, vol. 31, no. 1, pp. 45–54, Jan 2014.

[5] N. Keshava, "A survey of spectral unmixing algorithms," *Lincoln Laboratory Journal*, vol. 14, no. 1, pp. 55–78, 2003.

[6] N. Keshava and J. Mustard, "Spectral unmixing," *Signal Processing Magazine, IEEE*, vol. 19, no. 1, pp. 44–57, Jan 2002.

[7] J. Li and J. M. Bioucas-Dias, "Minimum Volume Simplex Analysis: A fast algorithm to unmix hyperspectral data," in *Geoscience and Remote Sensing Symposium, 2008. IEEE International*, vol. 3, July 2008.

[8] C. Salvaggio and C. J. Miller, "Comparison of field- and laboratory-collected midwave and longwave infrared emissivity spectra/data reduction techniques," pp. 549–558, 2001.

[9] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[10] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, Aug 2014.

[11] J. M. Bioucas-Dias, A. Plaza, S. Member, N. Dobigeon, M. Parente, Q. Du, S. Member, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, pp. 354–379, 2012.

[12] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier, "On label dependence and loss minimization in multi-label classification," *Machine Learning*, vol. 88, no. 1-2, pp. 5–45, 2012.

[13] M. Bossard, J. Feranec, J. Otahel *et al.*, "CORINE land cover technical guide: Addendum 2000," 2000.

[14] C. Justice, E. Vermote, J. Townshend, R. DeFries, D. Roy, D. Hall, V. Salomonson, J. Privette, G. Riggs, A. Strahler, W. Lucht, R. Myneni, Y. Knyazikhin, S. Running, R. Nemani, Z. Wan, A. Huete, W. Van Leeuwen, R. Wolfe, L. Giglio, J.-P. Muller, P. Lewis, and M. Barnsley, "The moderate resolution imaging spectroradiometer (MODIS): land remote sensing for global change research," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 36, no. 4, 1998.

[15] A. Santos, A. Canuto, and A. Neto, "A comparative analysis of classification methods to multi-label tasks in different application domains," *Int. J. Comput. Inform. Syst. Indust. Manag. Appl*, vol. 3, 2011.

[16] R. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, pp. 135–168, 2000.

[17] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions." in *ISMIR*, vol. 8, 2008.

[18] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pat. Rec.*, vol. 37, no. 9, 2004.

[19] D. Pflugmacher, O. N. Krankina, W. B. Cohen, M. A. Friedl, D. Sulla-Menashe, R. E. Kennedy, P. Nelson, T. V. Loboda, T. Kuemmerle, E. Dyukarev, V. Elsakov, and V. I. Kharuk, "Comparison and assessment of coarse resolution land cover maps for Northern Eurasia," *Remote Sensing of Environment*, vol. 115, no. 12, pp. 3539 – 3553, 2011.

[20] G. Büttner, J. Feranec, G. Jaffrain, L. Mari, G. Maucha, and T. Soukup, "The CORINE land cover 2000 project," *EARSeL eProceedings*, vol. 3, no. 3, pp. 331–346, 2004.

[21] C. Giri, Z. Zhu, and B. Reed, "A comparative analysis of the Global Land Cover 2000 and MODIS land cover data sets," *Remote Sensing of Environment*, vol. 94, no. 1, pp. 123 – 132, 2005.

[22] J. Guo, J. Zhang, Y. Zhang, and Y. Cao, "Study on the comparison of the land cover classification for multitemporal modis images," in *Earth Observation and Remote Sensing Applications*, 2008, pp. 1–6.

[23] V. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93 – 104, 2012.

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

[25] P. Teillet, K. Staenz, and D. William, "Effects of spectral, spatial, and radiometric characteristics on remote sensing vegetation indices of forested regions," *Remote Sensing of Environment*, vol. 61, no. 1, 1997.

[26] *Remote Sensing Satellites*. John Wiley & Sons Ltd, 2014, pp. 524–576.

[27] R. DeFries, M. Hansen, and J. Townshend, "Global discrimination of land cover types from metrics derived from AVHRR pathfinder data," *Remote Sensing of Environment*, vol. 54, no. 3, pp. 209 – 222, 1995.

[28] X. Yang and C. P. Lo, "Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia metropolitan area," *International Journal of Remote Sensing*, vol. 23, no. 9, pp. 1775–1798, 2002.

[29] J. Xia, P. Du, X. He, and J. Chanussot, "Hyperspectral remote sensing image classification based on rotation forest," *Geoscience and Remote Sensing Letters, IEEE*, vol. 11, no. 1, pp. 239–243, Jan 2014.

[30] M. E. Winter, "N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data," pp. 266–275, 1999.

[31] J. Nascimento and J. Bioucas-Dias, "Vertex component analysis: a fast algorithm to unmix hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 4, pp. 898–910, April 2005.

[32] J. Bioucas-Dias, "A variable splitting augmented lagrangian approach to linear spectral unmixing," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS '09. First Workshop on*, Aug 2009, pp. 1–4.

[33] J. Nascimento and J. Bioucas Dias, "Does independent component analysis play a role in unmixing hyperspectral data?" *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 1, Jan 2005.

[34] K. Themelis, A. Rontogiannis, and K. Koutroumbas, "A novel hierarchical bayesian approach for sparse semisupervised hyperspectral unmixing," *Signal Proc., IEEE Transactions on*, vol. 60, no. 2, 2012.

[35] J. Bioucas-Dias and M. Figueiredo, "Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing," in *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, June 2010, pp. 1–4.

[36] G. Tsagkatakis and P. Tsakalides, "Compressed hyperspectral sensing," pp. 940 307–940 307–9, 2015.

[37] Y. Altmann, A. Halimi, N. Dobigeon, and J.-Y. Tourneret, "Supervised nonlinear spectral unmixing using a postnonlinear mixing model for hyperspectral imagery," *Image Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 3017–3025, June 2012.

[38] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Deroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, Sep. 2012.

[39] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *In Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 681–687.

[40] A. Clare and R. King, "Knowledge discovery in multi-label phenotype data," in *Principles of Data Mining and Knowledge Discovery*, ser. Lecture Notes in Computer Science, L. De Raedt and A. Siebes, Eds. Springer Berlin Heidelberg, 2001, vol. 2168, pp. 42–53.

[41] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, p. 2007, 2007.

[42] M.-L. Zhang, "Ml-rbf: RBF neural networks for multi-label learning," *Neural Processing Letters*, vol. 29, no. 2, pp. 61–74, 2009.

[43] L. Rokach, A. Schclar, and E. Itach, "Ensemble methods for multi-label classification," *Expert Systems with Applications*, vol. 41, no. 16, 2014.

[44] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2010, pp. 667–685.

[45] O. Luaces, J. Dez, J. Barranquero, J. del Coz, and A. Bahamonde, "Binary relevance efficacy for multilabel classification," *Progress in Artificial Intelligence*, vol. 1, no. 4, pp. 303–313, 2012.

[46] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," vol. 85, no. 3, 2011, pp. 335–359.

[47] G. Tsoumakas, I. Katakis, and L. Vlahavas, "Random k-labelsets for multilabel classification," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 7, pp. 1079–1089, July 2011.

[48] M. D. Turner, C. Chakrabarti, T. B. Jones, J. F. Xu, P. T. Fox, G. F. Luger, A. R. Laird, and J. A. Turner, "Automated annotation of functional imaging experiments via multi-label classification," *Frontiers in neuroscience*, vol. 7, 2013.

[49] T.-H. Chiang, H.-Y. Lo, and S.-D. Lin, "A ranking-based KNN approach for multi-label classification." *ACML*, vol. 25, pp. 81–96, 2012.

[50] W. Cheng and E. Hüllermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Mach. Learn.*, vol. 76, no. 2-3, pp. 211–225, Sep. 2009.

[51] R. B. Myneni, C. Keeling, C. Tucker, G. Asrar, R. Nemani *et al.*, "Increased plant growth in the northern high latitudes from 1981 to 1991," *Nature*, vol. 386, no. 6626, pp. 698–702, 1997.

[52] A. J. Ramon Solano, Kamel Didan and A. Huete, "MODIS vegetation index user's guide (MOD13 series)," May 2010.

[53] C. Justice, J. Townshend, E. Vermote, E. Masuoka, R. Wolfe, N. Saleous, D. Roy, and J. Morisette, "An overview of MODIS land data processing and product status," *Remote Sensing of Environment*, vol. 83, no. 12, pp. 3 – 15, 2002.

[54] Z.-L. Li, B.-H. Tang, H. Wu, H. Ren, G. Yan, Z. Wan, I. F. Trigo, and J. A. Sobrino, "Satellite-derived land surface temperature: Current status and perspectives," *Remote Sens. of Environ.*, vol. 131, 2013.

[55] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Multilabel text classification for automated tag suggestion," in *In: Proceedings of the ECML/PKDD-08 Workshop on Discovery Challenge*, 2008.

[56] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, vol. 1, no. 1-2, pp. 69–90, 1999.

[57] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, USA: Cambridge University Press, 2008.

[58] S. Nowak, H. Lukashevich, P. Dunker, and S. Rüger, "Performance measures for multilabel evaluation: A case study in the area of image classification," in *Proceedings of the International Conference on Multimedia Information Retrieval*. New York, USA: ACM, 2010.

[59] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[60] Y. Altmann, N. Dobigeon, and J.-Y. Tourneret, "Nonlinearity detection in hyperspectral images using a polynomial post-nonlinear mixing model," *Image Processing, IEEE Transactions on*, vol. 22, no. 4, 2013.

Konstantinos Karalas received the B.Sc. and M.Sc. degree in Electronic and Computer Engineering from the Technical University of Crete, Chania, Greece in 2013 and 2015, respectively. As part of his M.Sc. research programme he joined the Institute of Computer Science - FORTH, Heraklion, Greece. His research interests are in the fields of signal processing, machine learning, and image classification.

Grigorios Tsagkatakis received his B.S. and M.S. degrees in Electronics and Computer Engineering from Technical University of Crete (TUC), Greece in 2005 and 2007 respectively. He was awarded his PhD in Imaging Science from the Center for Imaging Science at the Rochester Institute of Technology (RIT), USA in 2011. He is currently a postdoctoral research fellow at the Institute of Computer Science - FORTH, Greece. His research interests include signal processing and machine learning with applications in sensor networks and imaging systems.

Michael Zervakis holds a Ph.D degree from the University of Toronto, Department of Electrical Engineering, since 1990. He joined the Technical University of Crete on January 1995, where he is currently full professor at the department of Electronic and Computer Engineering. Prof. Zervakis is the director of the Digital Image and Signal Processing Laboratory (DISPLAY) and is involved in research on modern aspects of signal processing, including estimation and constrained optimization, multichannel and multi-band signal processing, wavelet analysis for data/ image processing and compression, biomedical imaging applications, neural networks and fuzzy logic in automation applications. He has been involved in more than 20 international projects and has published more than 90 papers in related areas of image/signal processing.

Panagiotis Tsakalides received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Greece, in 1990, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, in 1995. He is a Professor and the Chairman of the Computer Science Department at the University of Crete, and the Head of the Signal Processing Laboratory at the Institute of Computer Science (FORTH-ICS). His research interests lie in the field of statistical signal processing with emphasis in non-Gaussian estimation and detection theory, sparse representations, and applications in sensor networks, audio, imaging, and multimedia systems. He has coauthored over 150 technical publications in these areas, including 30 journal papers. Prof. Tsakalides has an extended experience of transferring research and interacting with the industry. During the last 10 years, he has been the project coordinator in 6 European Commission and 9 national projects totaling more than 4 M in actual funding for the University of Crete and FORTH-ICS.