Benchmarking of the Symbolic Machine Learning classifier with state of the art image classification methods

Application to remote sensing imagery

Martino Pesaresi
Vasileios Syrris
Andreea Julea

2015
Abstract
A new method for satellite data classification is presented. The method is based on symbolic machine learning (SML) techniques and is designed for working in complex and information-abundant environments, where it is important to assess relationships between different data layers in model-free and computational-effective modalities. In particular, the method is tailored for operating in earth observation data scenarios connoted by the following characteristics: i) they are made by a large number of data granules (scenes), ii) they are made by heterogeneous sensors and iii) they are mapping a large variety of different geographical areas in different data collection conditions. The volume, variety and partially unstructured nature of these scenarios can be associated with the characteristics of Big Data. The results of an experiment observing the behavior of the SML classifier by injecting increasing levels of noise in the training set are discussed. Spatial generalization, random thematic noise and spatial displacement noise are tested. Seven supervised classification algorithms have been considered for comparison: Maximum Likelihood, Logistic Regression, Linear Discriminant Analysis, Naive Bayes, Decision Tree, Random Forest and Support Vector Machine. According to the results of the experiment, the SML classifier performed very well providing outputs with comparable or better quality than the other classifiers. Furthermore, the better performances were released with a much less expensive computational cost. Consequently, the SML classifier was evaluated as the best available solution in the specific data scenario under consideration. Few applicative examples of the new SML classifier using Spot5, Sentinel1, and Sentinel2 data inputs are provided.
# Contents

1 Introduction ........................................... 2

2 Rationale .............................................. 3
   2.1 Scenarios ........................................ 4
   2.2 Causal explanatory paradigm ...................... 4
   2.3 Learning by examples ............................. 5
   2.4 Symbolic Learning ................................ 6

3 The new classification method ..................... 7
   3.1 Basic components ................................ 7
   3.2 Data quantization-sequencing .................... 7
   3.3 Association Rule Analysis ......................... 9
   3.4 Complexity analysis and method assumptions .... 10
   3.5 SML classifier .................................. 12

4 Experimental set .................................. 15
   4.1 Objectives ....................................... 15
   4.2 Test site ......................................... 16
   4.3 Data sets ......................................... 16
   4.4 Workflow ......................................... 17

5 Experimental design ............................... 19
   5.1 Scale generalization noise (A) .................... 19
   5.2 Random thematic noise (B) ........................ 19
   5.3 Spatial displacement noise (C) .................... 20
   5.4 ENDI cutoff ..................................... 20
   5.5 Classifiers parametrization ....................... 20
   5.6 Quality measurements ............................. 20

6 Results .............................................. 22
   6.1 On data reduction parameters .................... 22
   6.2 Benchmarking ..................................... 24
   6.3 On ROC analysis .................................. 28
   6.4 On applicative examples ......................... 30

7 Discussion .......................................... 36

8 Conclusions ......................................... 38

Bibliography ........................................... 40
Chapter 1

Introduction

In this manuscript a new method for classification of remotely sensed data is presented. The method builds upon early work made in the 90s for the definition of a distance metric based on empirical frequency histogram. The Histogram Distance Index (HDI) was defined in [Pes00] for the evaluation of the discriminant performances of textural measures showing strong non-Gaussian behaviors; in [Cha07] the same formulation is called 'Wave Hedges' as reference to the work done by [Hed76]. In more recent years, a similar approach was explored for solving the problem of the objective evaluation of the contribution in information produced by different input sensors, the evaluation being functional to the improvement of the information fusion phase in complex multi-sensor information mosaics as the one described in [Pes13]. During this exploratory work, a more general potential for data classification was envisaged in [Pes13] and systematic experimental activities were started with the aim of evaluating these techniques for solving massive information extraction tasks in multi-sensor data scenarios. In particular, in [Pes+15] these techniques were applied for the production of the global assessment of the built-up areas from multi-sensor Landsat data collected in the past 40 years.

The manuscript reports on tests using different satellite image data inputs including Spot (optical 2.5m resolution), Sentinel1 (radar 5m resolution) and Sentinel2A (optical 10m resolution) data will be discussed.

The tests aimed: i) to understand what are the critical parameters in the Symbolic Machine Learning (SML) classifier, ii) to understand how they can be automatically estimated during the classification workflow, and iii) to compare the SML results with other available classification techniques.

The remainder of the paper is organized as follows. Section 2 presents the big remote sensing data framework and the characteristics of the developed symbolic machine learning approach. The introduction of Evidence-based Normalized Differential Index (ENDI) measure and the details about the strategy, complexity analysis and assumptions of the new classification method based on Association Analysis are given in section 3. The section 4 presents the experimental set components while the experimental protocol aiming at measuring the performance is detailed in section 5. The large amount of experimental results are presented in section 6 discussed in section 7 and the conclusions are drawn in section 8.
Chapter 2

Rationale

We are living in the information age or digital revolution that analogously to the industrial revolution will create profound changes in the Economy, Society, and Culture of humanity, according to several authors as [Cas11a] [Cas11b] [Cas10]. In this new age, the information is abundant and continuously reprocessed: according to [HL11] the world’s technological capacity to store information grew from 2.6 (optimally compressed) exabytes in 1986 to 15.8 in 1993, over 54.5 in 2000, and to 295 in 2007. The world’s technological capacity to compute information with humanly guided general-purpose computers grew from $3 \times 10^8$ MIPS (Million Instructions Per Second) in 1986, to $4.4 \times 10^9$ MIPS in 1993, $2.9 \times 10^{11}$ MIPS in 2000 and $6.4 \times 10^{12}$ MIPS in 2007.

Nowadays the information is abundant, redundant and also (in general) largely contradictory - typically not harmonized as a whole. It is rapidly changing, heterogeneous and only partially structured because created by multilateral actors with a multiplicity of different objectives. According to [DLM14], “Big data” are high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

In the new method proposed for remotely-sensed data, causal deterministic models are largely complemented by data-driven inductive inferential reasoning - analogously as argued by [MSC13]. The whole data volume (samples = all) is typically used for making the inferences: this is considered improving their reliability in noisy and unstructured information environments where the assumptions needed for the correct sampling and stratification procedures are often largely violated or too expensive to be implemented. The methodological focus is provided on computing-efficient search of associative rules between large comprehensive data series describing the whole universe. This is in contrast to computing-expensive search of laws explaining few carefully-sampled points in the same data universe. In these (big, complex) data scenarios, the resilience (model transferability) of the eventually-found laws to new data segments coming in the continuous information flux is considered generally low, consequently not paying off the cost of their formulation.
The remote sensing data considered in this study have three main characteristics: i) they are made by a large (say greater than $10^4$) number of data granules (scenes), ii) they are made by heterogeneous sensors and iii) they are mapping a large variety of different geographical areas in different data collection conditions. The volume, variety and partially unstructured nature of these data scenarios can be associated with the characteristics of Big Data as in [DLMM14]; this is especially true if we combine the complexity of the data with time-critical requirements that are typical in some application areas as crisis and disaster management. For the purpose of brevity, in the manuscript we will refer to these scenarios as big remote sensing data (BRSD) scenarios.

Earth observation is about geo-information or mapping the same surface from different points of view (sensors, time). According to the approach taken here, it is highly probable that on the same portion of territory mapped by each of the image data granules under processing, an alternative description of the same or similar information already exists from alternative sources. These sources may originate from remote sensors, or other geo-spatial repositories as digital cartography, crowd sources or social media. Those sources can be linked to the data under processing because they are geo-coded, so referring to the same earth surface samples. The general assumption is that we live in an information-abundant environment: those alternative sources can be exploited in order to discover systematic relations between the remote sensing data instances and the target information to be extracted.

From the point of view of the specific image information extraction task, the alternative sources describing the same portion of territory include some kind of noise, justifying the effort done in extracting new information. For instance, in [Peg+13] the reference data used for learning set was 1000m resolution, while image data input of 0.5m resolution was tested during the experiment. In that case of scale gap, one single training example was corresponding to 4 millions of image data instances to be classified. Similarly, in [Sex+13] a relevant scale gap between the image data to be classified (30m resolution) and the available training set derived from MODIS 250m resolution data was observable.

Semantic differences, spatial information differences (scale, projection), temporal changes are the most common sources of noise dealing with multi-source data integration. The aim of the experiment reported here was to observe the behavior of the SML classifier in relation to three main sources of noise: scale gaps, thematic errors, and spatial inconsistencies.

2.2 Causal explanatory paradigm

In Remote Sensing (RS) systems, the primary information collected by the sensor is the quantity of electromagnetic energy reflected or emitted by the targets or surface units on the ground in specific wavelengths or energy frequency ranges. According to [LKC+04], the standard paradigm for extracting information from RS data rely on physical explicit modeling of the causal relations between target's energy absorption-reflection-emitting properties and sensor technical characteristics mediated by the atmosphere column between sensor and target. The interpretation model should include target reflectance-
emission characteristics, target mixture composition and target spatial patterns influencing the final amount of energy collected by the sensor at a given spatial resolution. Moreover, obsolescence and atmosphere exposure processes of the target surface (e.g. oxidation) may influence target's reflectance characteristics and should be taken into account. Furthermore, weather conditions before and during the data collection, and atmosphere column composition (aerosol, humidity, dust, etc.) between sensor and target during data collection will greatly influence the amount of energy collected by the sensor and should be modeled and compensated. Finally, accurate radiometric calibration of the sensor is required in order to estimate the collected absolute amount of energy. In this standard approach, several parameters required by the causal models need to be estimated or assumed by injecting prior knowledge in the data interpretation models. Estimation of these parameters typically requires a significant amount of time of high level experts, the complement of additional models and assumptions (for example atmospheric and weather models) and field measurements surveys with specialized devices as spectrometers. In this paradigm, the causal modeling is rigidly tailored around a single sensor and requires high stability of the input data, together with stability of the reflectance characteristics of the target to be detected. A single given stable sensor data set and a known and robust physical model linking data instances to information may allow the efficient scaling of this paradigm to large volume of data. Nevertheless, this paradigm is difficult to apply in real BRSD scenarios: the main challenges are related to the high requirements of data input (stability, calibration), the cost for the collection of necessary ancillary data, and the cost of porting the model in the different sensors and semantic (target information) domains. These costs become rapidly prohibitive or simply not applicable in BRSD cases.

2.3 Learning by examples

Alternative paradigms to the classical automatic image interpretation have been proposed: they try to bypass some of the modeling steps required in the standard causal paradigm, by making them implicit in the machine learning and classification phase that include the collection of target examples. Models linking sensor-observed energy reflectance characteristics with the presence of the target information are inductively learned by providing examples to the machine. A large number of machine learning approaches are currently used in state-of-the-art image information retrieval: they include supervised learning techniques like logistic regression, linear regression, decision trees and support vector machines; meta (ensemble) learning algorithms like boosting and bagging; unsupervised learning algorithms for clustering, dimensionality reduction, and density estimation; and feature learning algorithms as the auto-encoder, just to name a few popular ones (see section 3.4). Historically those algorithms have been designed and tested on small to medium data instances (if compared with RS high resolution data scenarios) with a moderate to large number of attributes or features. According to [Zy14], the following challenges are still open in geospatial Big Data machine learning and consequently in BRSD scenarios discussed here: i) big data volume needs to deal with a number of training examples beyond current in-memory processing capability, ii) big data variety creates feature dimensionality beyond computing capability, iii) big data vari-
et al increases learning complexity so as not be achievable in a meaningful time frame, and iv) big data velocity places requirements on the time efficiency of both learning and classification which are not attainable.

2.4 Symbolic Learning

In the meaning proposed here, symbolic (categorical, nominal) refers to the representation of the input data used for the classification, as opposed to continuous or quantitative. Symbolic data processing and information extraction techniques are traditionally used in the Knowledge Discovery in Data (KDD) or Data Mining (DM) applicative areas. Association Rule Analysis and Associative Classification are some of the most popular paradigms in mature DM and KDD application areas that require the analysis of large databases. They are largely used in market basket analysis, medical diagnosis, protein sequences, public health surveillance, census data, fraud detection in web and credit card business, just to name the most relevant applications. In particular, the Association Rules in genes (or Genetic Association), is extensively used in bio-informatics for revealing biologically relevant associations between different genes or between environmental effects and gene expression [Bec+02; CH03; Geo+05; AGG10].

These paradigms are rarely applied in the remote sensing community due to the fact that the physical causal-explanatory models require continuous numeric representation of the input signal. A search for "association analysis and remote sensing" in the Scopus database reveals few results that are, interestingly, all dealing with systematic comparison of heterogeneous input geo-spatial data. In [DSS09], an analysis with the aim of discovering geo-spatial discriminating patterns from remote sensing datasets was applied, while in [MSP10], data mining techniques are applied for resources evaluation. More recently, [LZL13] applied association analysis in order to understand relations between NDVI changes and topographic factors. Finally, in [GZ14] heterogeneous land cover products are compared with association analysis techniques for analysis and accuracy improvement of contradicting land cover types.
Chapter 3

The new classification method

In this section, we submit details about the components and principles of the new Symbolic Machine Learning associative classifier: the two step strategy, the Evidence-based Normalized Differential Index measure and the method assumptions.

3.1 Basic components

By analogy with the Genetic Association applied in bio-informatics, the associative classifier (AC) proposed in this experiment searches for relevant, systematic relations between image data instances and spatial information encoded in the selected training sets. The classifier strategy is based on two fundamental steps relatively independent i) reduce the data instances to a symbolic representation, and ii) calculate the association between the symbolic data instances subdivided in X (input data) and Y (training set data) parts. The first step can be called also data reduction and is required only if the input data has a quantitative nature. Typically, the first step involves an unsupervised clustering technique, while the second is solved by supervised classification. As already mentioned, the two steps are relatively independent: different techniques can be chained in the two components. The specific implementation of the basic components explored in the experiment are described in the next sections.

3.2 Data quantization-sequencing

Numerous methods are available in literature for approaching the issue of data reduction, taxonomies and cluster analysis. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields including remote sensing data analysis, information retrieval, and bio-informatics. An extensive introduction and overview to the topic can be found in [Bai94] and [Eve+11]. In this experiment, a new specific clustering approach uses data (re-)quantization and data sequencing. Two main criteria were used to select the proposed data quantization-sequencing method. The first criterion comes from the exigence of an objective method - as much as is possible independent from assumptions about specific statistical distributions...
of the input data. In BRSD scenarios, this is important because just the analysis of the statistical distribution of specific input data volumes may become computationally very expensive or even prohibitive. This is especially true in rapidly changing and highly heterogeneous BRSD scenarios, including continuous streams of image data that would force the application of heterogeneous ad-hoc data reduction approaches. The second criteria is linked to the necessity to adopt a sparse representation of the input data features and discriminant functions. Such coding allows the compression of the feature space, increasing the computational efficiency in BRSD analytical scenarios. In relatively recent years, in the computer vision community such representations have been successfully proposed for solving image compression, pattern recognition and image classification tasks; see for example \cite{ZGK11}, \cite{Wri+10} and \cite{Mai+09}.

Let $D^{m \times n \times F}$ be a data set with $m \times n = mn$ spatial samples or pixels and $F$ features or descriptors. Let $X^{mn \times F}$ be a 2-dimensional data matrix, $X = [x_1, x_2, \ldots, x_i, \ldots, x_F]$, with $F$ expressing the number of used features and $x_i \in \mathbb{Z}^{mn}$. Let $\hat{X}$ be the set of all the unique data instances of $X$. The $\hat{X}$ elements are the unique data sequences and they have cardinality $|\hat{X}| \leq mn$; their number depends on the specific number of symbols $s_i$ used to encode the $x_i$ values and on the number of descriptors or features $F$. The average support of $\hat{X}$ can be estimated as:

$$\text{supp}_\mu = \frac{|X|}{|\hat{X}|}$$ (3.1)

The $|\hat{X}|$ influences both the classifier’s computational and generalization performance. Too small $\text{supp}_\mu$ may lead to over-fitting issues at the successive phase of association analysis. Thus, very often a re-discretization of $x$ values must be considered in order to keep under control this issue.

Several quantization or re-discretization methods including uniform, adaptive and statistical approaches may be found in literature as reviewed by \cite{Lin+02}. They are all compatible with the general SML framework. In the application presented later on, a uniform quantization approach was implemented, with the quantized data instances defined as follows:

$$X_{q,F} = [\lfloor x + 0.5 \rfloor / q_i : x \in x_i]^{mn \times F}$$ (3.2)

with $q_i = \max(x_i)/s_i$ and $i = 1, \ldots, F$. Accordingly, the respective quantized unique data sequences are denoted by $\hat{X}_{q,F}$. The justification of this choice relies on the fact that it is the computationally least expensive and the most objective method, being independent from the specific statistical data distributions found in the different datasets under processing. The $F$ number of available descriptors is dictated by the sensor characteristics and optionally complemented by additional image features (such as textural or morphological features) extracted from the raw image data. The $s_i$ number of eligible symbols per each descriptor is derived by the combination of the raw digital number encoding the specific data $D$ and the quantization parameter $q_i$. A similar approach can be taken for textural and morphological attributes. Due to the symbolic nature of the classifier, any categorical data as for example the output of other classification exercises, land cover etc. can be integrated seamlessly as attribute of the $X$ input set.

Figure 3.1 shows an example of quantization-sequencing applied to the Landsat data used in the experiment. The whole six bands were used as input features
with a quantization in 8 levels, generating a total of 3388 individual sequences in the specific scene under processing.

![Table and figure illustrating the quantization and classification process.](image)

Figure 3.1: An example of sequences and $\Phi_E$ measures generated using Landsat data and Corine Land Cover reference. In this example the six bands of the Landsat data were quantized in 8 levels producing 3388 unique sequences in the scene: they are ordered from top to bottom by the values of the input features. On the left of the image the sequences are provided, while on the right the corresponding $\Phi_E$ measure value is reported for each sequence combination. The color code is a scale from blue (low) to yellow-green (medium) up to dark red (high) values. Eight different classes are evaluated by $\Phi_E$ using Corine LC as examples: (a) water, (b) agricultural, (c) generic forest, (d) broad-leaved forest, (e) coniferous forest, (f) generic artificial surfaces, (g) industrial areas, and (h) residential areas.

### 3.3 Association Rule Analysis

Let $A = \{A_1, A_2, \ldots, A_j, A_{j+1}\}$ be a set of $j+1$ distinct attributes. For any two disjoint attribute-value subsets $\{X\}$ and $\{Y\}$ of $A$, the patterns of the form $X \rightarrow Y$ are called association rules, where $X$ and $Y$ are disjoint sets (i.e., $X \cap Y = \emptyset$). The attribute-value sets $X$ and $Y$ are called antecedent and consequent of the association rule respectively. In the Associative Classification (AC) framework
CHAPTER 3. THE NEW CLASSIFICATION METHOD

started by [LHM98], the Class Association Rules (CARs) are the association rules with a class label attribute as the only consequent \( X = \{ A \setminus C \} \) and \( Y = A_{j+1} = C \), where \( C \) is the class label attribute set.

The proposed method consists in using Associative Classification to quantify the association \( X \rightarrow Y \) between multi-attribute data sequences, \( X \), and available high-abstraction semantic layers as reference, \( Y \), using a measure inspired by a probability-based objective interestingness measure. The objective measures are those that are not application-specific or user-specific and depend only on raw data. In our remote sensing application, the attribute set \( \{ A \setminus C \} \) is \( \hat{X}_q \) from a multi-band (multi-attribute) satellite image and the class label attribute set \( \{ C \} \) is \( \{ Y \} \) obtained from binary semantic reference sources (training sets).

A large set of interestingness measures can be found in literature: extensive reviews can be found in [Tew+14] and [GH06]. In the experiment, a new measure is introduced: it is named Evidence-based Normalized Differential Index (ENDI). It is a generalization of a measure from the confidence family to the case where a multiple set of positive/negative not-mutually-exclusive examples are provided. Three variations of the ENDI measures are proposed and tested in the experiment: they are named \( \Phi_a^E \), \( \Phi_b^E \), and \( \Phi_{ab}^E \).

The ENDI confidence measure \( \Phi_a^E \) of the (antecedent) data instances \( X(= \hat{X}_q) \) provided the positive \( Y^+ \) and negative \( Y^- \) (consequent) data instances is defined as follows:

\[
\Phi_a^E (X, Y^+, Y^-) = \frac{f_{pos} - f_{neg}}{f_{pos} + f_{neg}} \tag{3.3}
\]

where \( f_{pos} \) and \( f_{neg} \) are the frequencies of the joint occurrences among \( X \) (antecedent) data instances and the positive and negative (consequent) data instances respectively. In the case where only one mutually-exclusive binary training set is provided, the \( \Phi_a^E \) measure can be reduced to the measure known as Charade/Ganascia Index proposed by [Gan91] or Descriptive Confirmed-Confidence as renamed by [Kod01]. The ENDI confidence measure \( \Phi_b^E \) is defined analogously to \( \Phi_a^E \) by substituting the frequencies \( f_{pos} \) and \( f_{neg} \) with the empirical probabilities \( p_{pos} \) and \( p_{neg} \) calculated as \( p_{pos} = f_{pos}/N_{pos} \) and \( p_{neg} = f_{neg}/N_{neg} \), with \( N_{pos}, N_{neg} \) denoting, respectively, the numerosity of the positive and negative training samples. The ENDI confidence measure \( \Phi_{ab}^E \) is defined as the median decision between the \( \Phi_a^E \) and the \( \Phi_b^E \) measures. In the formulation adopted in the experiment \( \Phi_{ab}^E = (\Phi_a^E + \Phi_b^E)/2 \).

Figure 3.1 shows an example of \( \Phi_a^E \) measure evaluated on a set of 3388 sequences derived from Landsat data. The classes used as reference are derived from Corine LC information of the same year of the image data.

3.4 Complexity analysis and method assumptions

Given that \( m \) denotes the number of data instances/examples/observations, \( n \) denotes the number of features/dimensions/variables, \( v \) the number of distinct values per variable, \( v < < m \) (for simplification reasons, we assume \( v = v_j, j = 1, \ldots, n \) and \( Y = \prod_{j=1}^n Y_j \)), and \( |Y| \) the number of classes, the total time complexity for the SML training is composed of: (a) taxonomy of the data set values, and (b) data frequencies computation, giving a cost of \( O(m \cdot n) \). The
brute-force approach yields a worst-case upper bound of order \( O(m^2 \cdot n) \). However, in satellite imagery, the practical dynamic range is reaching the level of 16-bits at most; in addition and after the taxonomy application, the range of values can be reduced considerably. Exploiting techniques like bit arrays and hash tables, a pragmatic time complexity can reach the order of \( \Theta (v^n) \), with integer \( v \leq 65,535 \). In terms of memory requirements, the worst case scenario is realised when all the data instances of the initial data set are unique, \( O(m \cdot (n + |Y|)) \); the extra structure of size \( m \cdot |Y| \) contains the occurrences of every unique data instance observed in each of the classes. According to the previous analysis, the pragmatic space complexity is \( \Theta (|Y| \cdot v^n) \). At the classification stage, a new \( n \)-datum receives a score based on its frequency in every class and gets a class label depending on the threshold applied over the score, resulting in \( O(|Y|) \); there is also an additional cost \( O(V \cdot n) \) to match the \( n \)-datum with an element of the unique data instances matrix, a cost that becomes \( O(1) \) in case of hashing; the space complexity remains the same, \( O(|Y| \cdot v^n) \).

Seven supervised classification algorithms which have been widely used in the statistical learning and data mining community have been considered for comparison: Maximum Likelihood (ML, [RJ99]), Logistic Regression (LR, [MN89]), Linear Discriminant Analysis (DA, [GHT07]), Naive Bayes (NB, [MRS08]), Decision Tree (DT, [Bre84]), Random Forest (RF, [Bre01]) and Support Vector Machine (SVM, [Vap95]); the list of references is just indicative and not exhaustive. For instance, even though Artificial Neural Networks (ANN) are broadly used classifiers ([LW07]) and in practical scenarios turn to be powerful tools, able to approximate any arbitrary function, this high performance doesn’t come for free. Particularly for large data sets they require long time to be trained and fine-tune a large number of parameters. Anyway, the main reason ANNs are missing from this indicative list is that the definition of the network architecture (topology, transfer functions) is more like art than a rigorous process and requires an extensive trial and error approach.

Table 3.1 shows the worst-case time and space complexity of the seven classifiers together with the SML algorithm. The complexities refer to batch-mode learning and execution on a single processor (no parallel implementations). The training space complexity reports approximately the memory cells needed for the storage of the data structures that are retained during the training process. Since only batch-mode learning is considered, all the algorithms have an initial space complexity equal to the size of the training set, i.e. \( O(m \cdot n) \). Respectively, the space complexity of the classification stage refers to the storage of the model parameters and the necessary data that are derived from the supervised learning process. The time complexity of the classification stage refers explicitly to a single-datum processing.

Although the worst-case complexity aspect seems prohibitive in some of the classifiers, heuristics and approximations can boost these algorithms and make them scale-up smoothly in large-scale learning scenarios. For example, in case of DT an average time complexity of order \( O(V \cdot n \cdot (\log (V))^2) \) can be achieved under some heuristic considerations ([DHS00]). Even if the theoretical worst-case complexity for the NB training is \( O(m^2 \cdot n) \), the implementation of effective data structures and mappings (hash tables, skip lists, trees, bitmaps, etc) decreases notably both the practical time and space complexity. Similar approaches can be applied on DTs, reducing further their complexity. Through approximations, a time complexity that is linear in \( m \) can be shown for SVMs ([FKC05]); the
CHAPTER 3. THE NEW CLASSIFICATION METHOD

12

m penalty factor is small and computational cost of solving the SVM problem grows at least like $\Theta(\log(m^3))$. The time complexity is also induced by the optimization method that is used in the case of LR and SVM and the re-sampling method (cross-validation or bootstrapping) employed for branch pruning and feature selection in case of DT and RF. Lastly, a factor that is usually overlooked, yet influences the algorithms complexity is the similarity/distance/kernel/scoring function; for instance, computing exponential kernels turns to be quite expensive. Also, due to this multivalued function, a standardization/normalization process of the numerical values needs to run before the training (SVM, ANN, LR with regularization).

ML, DA and NB are very easy to construct. The parameters estimation is straightforward without the need for iterative complicated algorithms; nevertheless, the model accuracy is strongly affected by their intrinsic assumptions (table 3.2). The RF is an ensemble method and as such, it is more computationally demanding; its complexity is in general proportional with that exhibited by the DT. In the cases of DT, RF and SVM, the parameters tuning takes place through a re-sampling method like cross-validation. These last three non-parametric methods together with SML exhibit a generalization capability that depends on the class separability and the chosen function that draws the separating hypersurfaces/hyper-boxes. This fact affects considerably the algorithms complexity; for instance, the classification time complexity of DT can become $O(\log(V))$; that of SVM becomes $O(p\cdot n)$, with $p$ being the number of support vectors that is usually expected to be $p \ll m$, and that of SML becomes $O(1)$.

### Table 3.1: Worst-case computational complexity in asymptotic notation

<table>
<thead>
<tr>
<th></th>
<th>Time</th>
<th>Space</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>$O(m\cdot n^2)$</td>
<td>$O(</td>
<td>V</td>
<td>\cdot n^2)$</td>
</tr>
<tr>
<td>LR</td>
<td>$O(m\cdot n^2)$</td>
<td>$O(m\cdot n)$</td>
<td>$O(n)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>DA</td>
<td>$O(m\cdot n^2)$</td>
<td>$O(</td>
<td>V</td>
<td>\cdot n^2)$</td>
</tr>
<tr>
<td>NB</td>
<td>$O(m\cdot n) \sim O(h(v)\cdot n)$</td>
<td>$O(h(v)\cdot n\cdot</td>
<td>V</td>
<td>)$</td>
</tr>
<tr>
<td>DT</td>
<td>$O(V^2\cdot n\cdot \log(V))$</td>
<td>$O(V\cdot n\cdot</td>
<td>Y</td>
<td>)$</td>
</tr>
<tr>
<td>RF</td>
<td>$O(P\cdot V\cdot n\cdot \log(V))$</td>
<td>$O(P\cdot V\cdot n)$</td>
<td>$O(P\cdot V)$</td>
<td>$O(P\cdot V\cdot n)$</td>
</tr>
<tr>
<td>SML</td>
<td>$O(m^2\cdot n)$</td>
<td>$O(g(v, n))$</td>
<td>$O([V] + O(g(v, n)\cdot</td>
<td>Y</td>
</tr>
</tbody>
</table>

$m$ data instances; $n$ dimensions; $|Y|$ classes; (RF) $P$ randomized trees; $h(v)$ is an effective mapping of the distinct values $v$ to an index; for the $n$-dimensional case, the mapping takes the form of a bivariate function $g(v, n)$. It holds $h(v) \leq g(u, n) \leq m$.

3.5 SML classifier

In terms of generalization performance and in the absence of assumptions, the No Free Lunch Theorem (DHS00) states that no pattern classification method is inherently superior to any other or even to random guessing. Following the Occam’s razor, one should start testing classifiers following a direction from low to high complexity models. According to the complexity analysis, NB and SML constitute the first choices since they are simple, easily and rapidly constructed, achieving relatively low memory footprint and quite fast classification. They can handle nominal and discrete values; for continuous numerical values, a sort of clustering/quantization needs to proceed. In contrast with NB, SML doesn’t
CHAPTER 3. THE NEW CLASSIFICATION METHOD

<table>
<thead>
<tr>
<th>Alg.</th>
<th>Mode</th>
<th>Parameters</th>
<th>Assumptions</th>
</tr>
</thead>
</table>
| ML   | single-pass | mean vectors and covariances | linear decision boundary; 
|      |      |            | $x^k$ follow different multivariate normal distributions |
| LR   | iterative | coefficient estimates | absence of multicollinearity among the variables; the resulting logit transformation is linear |
| DA   | single-pass | mean vectors and covariances | linear decision boundary; $x^k$ follow multivariate normal distributions with different mean vectors and same covariance |
| NB   | single-pass | class conditional likelihood | linear decision boundary; 
|      |      |            | all the variables are mutually independent and each follows a multinomial model within a class |
| DT   | two-pass | split points | feature space partitioning via hyper-rectangles; solution strongly dependent to locally-optimal decisions made at each node |
| RF   | iterative | split points | same as DT |
| SVM  | iterative | support vectors and weights | classes are linearly separable either into the feature space, or into a high-dimensional space defined implicitly by the kernel function |
| SML  | single-pass | threshold | none |

The $x^k$ denote the $k \leq m$ examples of class $y \in Y$.

Table 3.2: Main characteristics of the classification techniques under consideration

make any assumption about dependence/correlation of the features, neither presupposes specific underlying distributions in general. Typical non-parametric classifiers like DT, SVM and ANN use a flexible number of parameters that grows as the learning space becomes complex. SML instead doesn't have parameters that control the learning capacity. It can learn non-linear boundaries and define non continuous partitions in the feature space based only on the estimated joint probability. As in DTs, the knowledge SML discovers has a comprehensible representation in the form of rules.

The SML is an algorithm that is adapted over a dataset through one parameter, the setting that controls the data granularity. In practice, SML can achieve generalization only if a taxonomic schema clusters the data instances in meaningful granules. This is the sensitive part of the method and is under study. Low granularity (few features, small encoding alphabet) may lead to underfitting, while high data granularity (big number of features, many encoding symbols) can have the opposite effect (overfitting). Techniques like clustering, quantization and data encoding are under consideration and test. At the mo-
ment, this issue seems to be the weak/sensitive point since it can be interpreted as a large-scale optimization problem. However, empirical evidence \cite{Pes+15} shows that even by adopting brute-force solutions like uniform quantization applied to each feature independently, SML achieves significant classification performance. Another open issue is the comparison of different interestingness measures and scoring functions and the assessment of their contribution in terms of class separability and robustness against noise.

In remote sensing image classification particularly, SML is suggested for problems characterized by i) ill-defined and spectrally heterogeneous target classes (as built-up areas), ii) lack of seasonal standardization and/or radiometric calibration), iii) low resolution training set the downscaling of which brings spatial misalignment and thematic noise, and iv) necessity for fusion with other information layers.

For the above mentioned reasons, we claim that SML can constitute an efficient candidate for large-scale processing in BRSD scenarios.
Chapter 4

Experimental set

This section presents details about the test site, the input and reference data sets, and the workflow followed in the experiments.

4.1 Objectives

The main objectives of the test reported here were to answer to the following questions: i) what are the critical parameters to be set in the new SML classifier, ii) how they can be automatically estimated during the classification workflow, and iii) how the performances of the SML can be compared with other available classification techniques.

The three objectives are focused around the specific task of automatic recognition of built-up areas in satellite imagery using exclusively radiometric information: in particular, radiometric information collected by the Landsat multispectral sensor having 30m of spatial resolution. The task was recognized as a kind of worst-case-scenario: urban and more generally settlement areas when observed by satellite sensors are connoted by large spectral mixture and inner heterogeneous spectral characteristics. In average, the characteristic scale of the entities composing the contemporary settlements is of 10 meters. Early empirical evidences confirming the average scale of settlement entities were collected by [Sma03; Sma05]. The observed spectral variability is mainly produced by i) the relation between sensor spatial resolution and the characteristic scale of the settlement entities, ii) the presence of different materials in the settlement areas, their variable local spatial arrangements and mixture, and iii) the variable illumination conditions coupled with variable building sizes causing variable shadowing largely affecting the observed radiometric behavior.

Apart from the above general-interest reasons related to the discrimination of complex information showing large heterogeneity and mixture issues, a more specific reason was considered during the experimental design of the test. This is related to the estimation of some operative parameters that are used in [Pes+15] for the assessment of the global human settlements from 40-years record of Landsat satellite data.
4.2 Test site

The experiment described here was set in the Tuscany region, situated in central Italy with an area of about 23,000 km$^2$ and an estimated population for 2013 of about 3.8 million inhabitants. Surrounded and crossed by major mountain chains, and with few plains, the region has a relief that is dominated by hilly zones and is used for agriculture. Hills make up nearly two-thirds (66.5%) of the region’s total area, and mountains (of which the highest are the Apennines), a further 25%. Plains occupy 8.4% of the total area mostly around the valley of the River Arno. Tuscany shows a long history of important human settlement colonization dating back to the late second millennium BC (roughly 1350-1150 BC). Nowadays, the settlement structure is dominated by a set of very compact-core towns shaped in the medieval time: the modern industrial expansion is sometime remarkable and typically less compact as urban structure. A large agricultural area is connoted by diffused traditional small settlements, hamlets, and isolated buildings. In the area, the residential buildings are typically built with a clay tile roof, while industrial and services buildings may show metal, concrete and any other modern material for roofing. Depending on agricultural practices and timing, plugged or recently seeded agricultural fields with clay soil and dry moisture may show significant similarity with clay roof tiling reflectance characteristics.

4.3 Data sets

The reference information set consists of 951,466 building footprints available in vector format and covering the whole Tuscany region. The buildings are described with a cartographic generalization of 1:10,000 in a topographic map source made by the Tuscany region administration and updated in 2000.

Therefore, the input satellite data used in the experiment were collected by the Landsat 7 Enhanced Thematic Mapper Plus (ETM+) sensor on 20 June 2000, in the path 192 and row 030 of the satellite. The data were pre-processed by the USGS Global Land Survey (GLS) and made available for public use in the GLS 2000 collection.

The building footprints in vector format were translated to raster format at the same projection of Landsat data and to the much greater spatial resolution of 1x1 meter. The spatial resolution of the reference data was then fitted to the resolution of the satellite image data (30 meters) by an average operator: as a consequence in each 30m resolution sample pixel the proportion covered by building footprint coded in 1m resolution raster was reported. In the test reported here any pixel of the 30m resolution reference data with a value strictly greater than zero is considered as positive example of building footprints. The approach is consistent with the built-up area class defined as the union of all the spatial units collected by the specific sensor and containing a building or part of it as applied in [Pes+13] and [Pes+15].
CHAPTER 4. EXPERIMENTAL SET

4.4 Workflow

In the experiment, the reference information is used for both training and testing. The training set is biased by introducing increasing levels of noise, while the test set remains constant allowing the monitoring of the classifier performances against the ideal point.

The experiment is made by two main parts: the first is designed for a quantization and noise assessment of SML classifiers, while the second is designed for benchmarking SML with other classifiers.

The first part of the experiment is organized around three main noise injection tests called A, B and C and defined as follows: A) scale generalization noise, B) random thematic noise and C) spatial displacement noise. Each noise test has a specific number of levels of noise tested: specifically 10, 11 and 19, for the test A, B, and C, in this order. They are described in the next section. Each test is performed with the whole six bands available in input from the Landsat sensor, and 8 different levels of the quantization parameter \( q \). As a consequence, each classifier was tested with \((10 + 11 + 19) \times 8 = 320\) different combinations of quantization and noise injection parameters.

As explained in section 3.3, three \( \Phi_E \) measures are employed, named \( a \), \( b \), and \( ab \). On each \( \Phi_E \) measure, 5 different automatic threshold approaches are tested:

- \( c_0 \) - unsupervised threshold at the zero level
- \( c_1 \) - area-based foreground histogram fitting
- \( c_2 \) - training statistics on foreground domain
- \( c_3 \) - training statistics on background domain
- \( c_4 \) - training statistics on median point of foreground and background domains

They are described in the next section. As a consequence of the above, a total of \( 3 \times 5 = 15 \) SML outputs are tested for each combination of quantization and noise injection. This brings to \( 320 \times 15 = 4800 \) total number of tests involving SML classifiers during the experiment. The quality of the training set was also evaluated during the noise injection phase in order to understand the value added of the classifier. The total number of tests during this part of the experiment is therefore equal to \( 320 \times (15 + 1) = 5120 \).

In the second part of the experiment, a specific value of the quantization parameter \( q \) was selected for the SML based on the good results in the first phase. The results of SML classifiers are then compared with the results of other classifiers, using the same input image data and the same biased training set. The input image data supplied to the other classifiers (OTH) were not re-quantized and provided at the same resolution available in the raw data (256 levels). Preliminary tests on effects of the quantization of input data on the OTH classifiers showed a constant decrease of performance associated to increasing values of the quantization parameter \( q \). Consequently, a strategy aiming to maximize the performances of OTH classifiers was adopted. The OTH classifiers that were subject to the experiment are the ones described in
Section 3.4 All the A, B and C noise injection tests were performed with the OTH classifiers, producing \((10 + 11 + 19) \times 7 = 280\) outputs and evaluations.

The total number of outputs and evaluations produced by the experiment in the first and second part is equal to \(5120 + 280 = 5400\).

Figure 4.1: Examples of noise injected in the training set. Top line: A) scale generalization noise - white and red, respectively, the unbiased and biased references - from left to right: generalization with a box of size 480, 960, and 1920 meters.

Middle line: B) random thematic noise - from left to right 0, 0.25 and 0.50 bias.

Bottom line: C) spatial displacement noise - white and red, respectively, the unbiased and biased references - from left to right displacement of 2, 18 and 36 pixels (60, 540, and 1080 meters). The image samples represent an area of with the side of about 15 km on the ground.
Chapter 5

Experimental design

Trying to put in place the proposed classification technique among some well-established classifiers, we designed an experimental protocol aiming at measuring the performance in terms of robustness and training/classification speed. The suggested experimental protocol attempts to capture three typical scenarios of signal disturbance as explained analytically in the next subsections. The binary reference set is notated from now on as $R$.

5.1 Scale generalization noise (A)

The process aims to simulate the information collected by sensors with low resolution. Using the variable $s(i) = 2s(i-1)$, with $s(1) = r$ and $i = 2, \ldots, 10$ to denote the scale, initially $R$ is being scaled up by a factor $s/r$ through a box-shaped kernel and then resized back to the original resolution by using nearest-neighbor interpolation. The produced raster data play the role of the degraded training sets (except from the trivial first case) and become binary after thresholding by 0.2. The variable $r$ expresses the original spatial resolution, which is 30m in this specific Landsat context. In the specific implementation discussed here, 10 different degraded sets were produced: the worst degraded set was produced by a generalization corresponding to a square kernel with the side 15,360 km (see upper line of Figure 11).

5.2 Random thematic noise (B)

By inserting artificially different levels of "salt & pepper" noise, we aim at measuring the susceptibility to alternative kinds of random disturbances that may occur over the training set. The vector of noise density values is given by the relationship: $d(i) = d(i-1) + 0.05$, with $d(1) = 0$ and $i = 2, \ldots, 11$. In this case, the degraded training sets derive from $R \land M(d_i)$, where $M(d_i)$ are the "On/Off" masks produced by the "salt & pepper" noise insertion. In the implementation discussed here, a total of 11 degraded sets were produced. The worst case includes a 50% random noise.
5.3 Spatial displacement noise (C)

The third test concerns the systematic translation of the training set with respect to the real reference set. Defining a vector of shifts as \( v(i) = v(i-1) + 2 \), with \( v(1) = 0 \) and \( i = 2, \ldots, 19 \) along the image matrix diagonal, the degraded training sets are being produced by

\[
[x_{kj}]_{m \times n}^i = \begin{cases} 
0, & \forall k < v(i) \land \forall j < v(i) \\
1, & x_{kj} \in R, \text{otherwise}
\end{cases}
\]

where \( m \) and \( n \) denote the number of rows and columns of \( R \) respectively. In the implementation discussed here a total of 19 levels of displacement were applied. The worst case included a displacement of 36 pixels, corresponding to 1080 meters.

5.4 ENDI cutoff

For each of the three measures \( \Phi_E, E = a, b \) or \( ab \), we test five alternative forms defined as follows (non data values excluded from the calculations):

\[
d_l = \begin{cases} 
1, & \forall x_{kj} > T_l : x_{kj} \in \Phi_E \\
0, & \text{otherwise}
\end{cases}
\]

with threshold values given by

\[
T_l = \left[ 0, \arg\min_i y_i - \sum_{i} R \cdot m_1, m_0, m_0 + (m_1 - m_0) / 2 \right]
\]

where \( y_i \) are the values of the cumulative sum over the quantized values of \( \Phi_E \), \( m_0 = \text{mean} (x_{lp} : (x_{lp} \in \Phi_E) \land (r_{lp} \in R) \land (r_{lp} = 0)) \) and \( m_1 = \text{mean} (x_{lp} : (x_{lp} \in \Phi_E) \land (r_{lp} \in R) \land (r_{lp} = 1)) \), for \( l = 0, \ldots, 4 \). Particularly, in the case of \( l = 0 \), the \( > \) operator becomes \( \geq \).

5.5 Classifiers parametrization

Concerning the classifiers ML, LR, DA, NB and the trees, two ways were tested for the prior probability estimation: (a) from the relative frequencies of the classes in the training and (b) using equal values. About the DT and RF, we selected the Gini’s diversity index as split criterion, the classification error as pruning criterion, and the 3-fold cross-validation for the tree growing in case of DT and bootstrapping in case of RF. For the latter, we chose to grow 20 trees in every cycle and select randomly 5 variables. Regarding the SVM, after executing an 10-fold cross-validation, we ended up with the Gaussian radial basis kernel with scaling factor of 0.6, the penalty factor of 0.9, the tolerance of 0.001 with which the Karush-Kuhn-Tucker conditions are checked and the sequential minimal optimization method for the training phase.

5.6 Quality measurements

The classification outputs produced in the experiment were systematically evaluated with an exhaustive set of performance measures. In this manuscript, the
description of the results is focused on the performance measure that we call Informedness \( I \) as introduced by [Pow03]. It is a popular skill score for contingency table prediction verification that was first proposed by [Pei84]. Since then, it has been rediscovered and renamed on multiple occasions in different domains: Youden’s J statistics or Youden’s index introduced by [You50] and used in medical statistics for diagnostic accuracy, Hanussen-Kuipers Skill Score or Kuipers’ performance index proposed by [HK65] and used in atmospheric science for meteorological forecasts, True Skill Statistic as named by [Flu87] and used also in meteorology. Weighted Relative Accuracy in its skew-insensitive dichotomous form from [LFZ99] and used in knowledge discovery, DeltaP’ identified as a psychological measure by [PP04] and used in linguistics, Powers Kappa or Informedness introduced in a betting scenario (Bookmaker) by [Pow03] and used in computational linguistics. In this manuscript, we use the Informedness denomination which gained popularity in the recent years. As stated in [Pow11], the measure can be directly quantifiable as the probability that an informed decision is made rather than a random guess (chance), specifying how much a model informs the specified condition (reference) or how informed a predictor is for the specified condition (reference).

The \( I \) measure is formulated as:

\[
I = TPR - FPR = \frac{TP \times TN - FP \times FN}{(TP + FN)(FP + TN)} = \frac{(TPR - \text{Bias})}{(1 - \text{Prev})}
\]  \hspace{1cm} (5.3)

with the usual meaning of the terms from the binary confusion matrix context: \( TP \)-true positives, \( TN \)-true negatives, \( FP \)-false positives, \( FN \)-false negatives, \( TPR \)-true positive rate or recall, \( FPR \)-false positive rate, \( \text{Bias} = (TP + FP)/N \), \( \text{Prev} = \text{Prevalence} = (TP + FN)/N \), \( N \)-total number of samples.

With a value range of [-1,1], for perfect predictions it has the value "1" (\( FN = FP = 0 \), no miss classification exists), for random predictions it has the value 0 (\( TPR = FPR \), same proportion of positive predicted in the real positive and real negative) and for predictions inferior to the random prediction it has a negative value. For constant prediction it also receives a value 0.

The measure is fair for both negative and positive events and is chance corrected: this is in contrast to the measures Recall (TPR), Precision (Positive Predictive Value) and F-factor that ignore performance in correctly handling negative examples and they fail to take account the chance level performance. In contrast to the kappa coefficient introduced by [Coh60] (also known as Heidke skill score proposed by [Hei26]), Informedness is an unbiased accuracy measure and is not affected by the prevalence (independent of the absolute/relative sizes of real positive and real negative classes) or bias (independent of the absolute/relative sizes of predicted positive and predicted negative classes). In accordance with [Pow11], it can be viewed as a renormalization (debiased version) of recall (TPR) after subtracting the chance level of TPR (the \( \text{Bias} \)). Unlike the case of Cohen’s kappa coefficient, the contribution made to Informedness by a correct “no” or “yes” prediction increases since the event is more or less likely, respectively. Being an equitable measure, it penalizes over-forecasting the most likely events and rewards correct predictions of rare events. All these properties are justifying the choice of Informedness as the appropriate performance measure for the evaluation of a classification system when a reference (Gold Standard) exists and especially in our context of imbalanced reference datasets.
Chapter 6

Results

This section summarizes and interprets the results of experimental works for the assessment of SML classifiers functionality and the comparison between SML and other classical type of classifiers. In addition, few examples of the SML approach applied to Spot and Sentinel satellite data are presented and commented.

6.1 On data reduction parameters

In this section, the results of the first part of our experiment, that deals with the quantization and noise assessment of the developed SML classifiers are presented and analysed. The table 6.1 contains information about the dependence of the new SML classifiers characteristics to the quantization and noise parameters.

<table>
<thead>
<tr>
<th>$q$</th>
<th>$N_{lev}$</th>
<th>$N_{seq}$</th>
<th>$AvgSupp$</th>
<th>Test A</th>
<th>Test B</th>
<th>Test C</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>256</td>
<td>2.E+07</td>
<td>1.E+00</td>
<td>0.4576</td>
<td>0.7347</td>
<td>0.2188</td>
<td>0.4203</td>
</tr>
<tr>
<td>2</td>
<td>128</td>
<td>9.E+06</td>
<td>2.E+00</td>
<td>0.4190</td>
<td>0.6131</td>
<td>0.1899</td>
<td>0.3636</td>
</tr>
<tr>
<td>4</td>
<td>64</td>
<td>2.E+06</td>
<td>1.E+01</td>
<td>0.4557</td>
<td>0.5702</td>
<td>0.3450</td>
<td>0.4346</td>
</tr>
<tr>
<td>8</td>
<td>32</td>
<td>2.E+05</td>
<td>1.E+02</td>
<td>0.4921</td>
<td>0.6103</td>
<td>0.4636</td>
<td>0.5111</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>3.E+04</td>
<td>9.E+02</td>
<td>0.4509</td>
<td>0.5648</td>
<td>0.4382</td>
<td>0.4762</td>
</tr>
<tr>
<td>32</td>
<td>8</td>
<td>3.E+03</td>
<td>7.E+03</td>
<td>0.3931</td>
<td>0.4771</td>
<td>0.3955</td>
<td>0.4173</td>
</tr>
<tr>
<td>64</td>
<td>4</td>
<td>5.E+02</td>
<td>4.E+04</td>
<td>0.3276</td>
<td>0.3887</td>
<td>0.3122</td>
<td>0.3306</td>
</tr>
<tr>
<td>128</td>
<td>2</td>
<td>1.E+02</td>
<td>2.E+06</td>
<td>0.3077</td>
<td>0.3600</td>
<td>0.2865</td>
<td>0.3183</td>
</tr>
</tbody>
</table>

Table 6.1: Influence of quantization parameter $q$ on SML_c4ab characteristics and informedness values for each type of noise injection.

Figure 6.1 presents the influence of noise and quantization level variation on the performance in terms of Informedness of the classifier SML_c4ab, in the three tested scenarios of signal disturbance. The notation SML_c4ab denotes that the measure $\Phi_{c4}$ was used with the $c4$ threshold option. For all three cases of noise injection (A - scale generalization, B - random thematic and C - spatial displacement), we can notice that the increase of the noise level produces in general a decrease of Informedness value with different shapes and slopes in function of noise injection type (accentuated for low levels of noise and almost imperceptible for medium and high levels). In addition, we observe
that, as the value of the quantization parameter $q$ increases, the behaviour of the classifier becomes more independent of the training set influence, allowing the consideration of the quantization parameter as a generalization one. We can note that the Informedness independence versus the training sets and noise levels benefits from higher values of generalization, despite of a small decreasing of the performance value with increasing of $q$ (for $q > 8$).

The experimental results of the three noise scenarios are synthesized in the last diagram, where an average behavior of all noise levels is presented for each test, according to last columns of table 6.1. It is obvious that the curves show a maximum at $q = 8$, corresponding to a quantization in 32 levels of the input image data sequences. This is the value of the generalization parameter being considered for the benchmarking study including OTH classifiers. The maximum performance is obtained in the case B of random thematic noise injection.
CHAPTER 6. RESULTS

6.2 Benchmarking

In the second part of our experiments, we make a comparison between the new SML and other classical type of classifiers, OTH. The results of this study are included in tables 6.2 - 6.5 that present the general benchmark in the three cases of noise injection (A, B and C) and in a general case obtained by averaging the performances. In these tables, the values of the characteristics for SML_c0b and SML_c4b are identical. This is not a random situation since it can be demonstrated that the threshold calculated analytically for the case of c4b, according to equation (5.2), is $T = 0$, the same value as in the unsupervised case c0b. This is only true if the reference data are mutually exclusive as is the case of the specific results discussed here. Generally, with non-mutually-exclusive training sets they may show values not exactly coinciding. In the tables, we have kept the two variants for their different significances.

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>$I_{Avg}$</th>
<th>$I_{StdDev}$</th>
<th>Omission Error</th>
<th>Commission Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SML_c4ab</td>
<td>0.6002</td>
<td>0.0710</td>
<td>0.2374</td>
<td>0.0487</td>
</tr>
<tr>
<td>SML_c0b</td>
<td>0.5991</td>
<td>0.0701</td>
<td>0.1826</td>
<td>0.0643</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.5941</td>
<td>0.0728</td>
<td>0.2956</td>
<td>0.0429</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.5906</td>
<td>0.0410</td>
<td>0.1483</td>
<td>0.0409</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.5876</td>
<td>0.0385</td>
<td>0.2394</td>
<td>0.0328</td>
</tr>
<tr>
<td>LHK</td>
<td>0.5792</td>
<td>0.0306</td>
<td>0.2396</td>
<td>0.0322</td>
</tr>
<tr>
<td>SML_c4b</td>
<td>0.5991</td>
<td>0.0701</td>
<td>0.1826</td>
<td>0.0643</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.5941</td>
<td>0.0728</td>
<td>0.2956</td>
<td>0.0429</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.5906</td>
<td>0.0410</td>
<td>0.1483</td>
<td>0.0409</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.5876</td>
<td>0.0385</td>
<td>0.2394</td>
<td>0.0328</td>
</tr>
<tr>
<td>LHK</td>
<td>0.5792</td>
<td>0.0306</td>
<td>0.2396</td>
<td>0.0322</td>
</tr>
<tr>
<td>SML_c4a</td>
<td>0.5201</td>
<td>0.0801</td>
<td>0.4305</td>
<td>0.0612</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.5100</td>
<td>0.0400</td>
<td>0.3811</td>
<td>0.0359</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.4728</td>
<td>0.0099</td>
<td>0.2347</td>
<td>0.0231</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.4578</td>
<td>0.1470</td>
<td>0.5042</td>
<td>0.1607</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.4334</td>
<td>0.1502</td>
<td>0.5387</td>
<td>0.1860</td>
</tr>
<tr>
<td>LHK</td>
<td>0.4434</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c4a</td>
<td>0.4341</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.4339</td>
<td>0.1502</td>
<td>0.5343</td>
<td>0.1850</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.4728</td>
<td>0.0099</td>
<td>0.2347</td>
<td>0.0231</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.4578</td>
<td>0.1470</td>
<td>0.5042</td>
<td>0.1607</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.4334</td>
<td>0.1502</td>
<td>0.5387</td>
<td>0.1860</td>
</tr>
<tr>
<td>LHK</td>
<td>0.4434</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c4a</td>
<td>0.4341</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.4339</td>
<td>0.1502</td>
<td>0.5343</td>
<td>0.1850</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.4728</td>
<td>0.0099</td>
<td>0.2347</td>
<td>0.0231</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.4578</td>
<td>0.1470</td>
<td>0.5042</td>
<td>0.1607</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.4334</td>
<td>0.1502</td>
<td>0.5387</td>
<td>0.1860</td>
</tr>
<tr>
<td>LHK</td>
<td>0.4434</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c4a</td>
<td>0.4341</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.4339</td>
<td>0.1502</td>
<td>0.5343</td>
<td>0.1850</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.4728</td>
<td>0.0099</td>
<td>0.2347</td>
<td>0.0231</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.4578</td>
<td>0.1470</td>
<td>0.5042</td>
<td>0.1607</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.4334</td>
<td>0.1502</td>
<td>0.5387</td>
<td>0.1860</td>
</tr>
<tr>
<td>LHK</td>
<td>0.4434</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c4a</td>
<td>0.4341</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.4339</td>
<td>0.1502</td>
<td>0.5343</td>
<td>0.1850</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.4728</td>
<td>0.0099</td>
<td>0.2347</td>
<td>0.0231</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.4578</td>
<td>0.1470</td>
<td>0.5042</td>
<td>0.1607</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.4334</td>
<td>0.1502</td>
<td>0.5387</td>
<td>0.1860</td>
</tr>
<tr>
<td>LHK</td>
<td>0.4434</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c4a</td>
<td>0.4341</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.4339</td>
<td>0.1502</td>
<td>0.5343</td>
<td>0.1850</td>
</tr>
<tr>
<td>ML1K</td>
<td>0.4728</td>
<td>0.0099</td>
<td>0.2347</td>
<td>0.0231</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.4578</td>
<td>0.1470</td>
<td>0.5042</td>
<td>0.1607</td>
</tr>
<tr>
<td>DA1K</td>
<td>0.4334</td>
<td>0.1502</td>
<td>0.5387</td>
<td>0.1860</td>
</tr>
<tr>
<td>LHK</td>
<td>0.4434</td>
<td>0.1501</td>
<td>0.5390</td>
<td>0.1850</td>
</tr>
</tbody>
</table>

Table 6.2: General benchmarking of classifiers upon Informedness ($I$) in experiment A of noise injection; Omission and Commission Errors and their Standard Deviations.

The results presented in tables 6.2 - 6.5 are visualised in the figures 6.2 - 6.4. The tables 6.2 - 6.3 - 6.4 show the performances of the different classifiers in the A, B and C noise injection tests. They average the performances for all the level of noise injected in the specific test. Table 6.5 shows the aggregated results for all the noise injection tests. By ranking the classifiers by their Informedness, the SML classifiers show generally similar or superior performances than the OTH. Only in the C noise injection (displacement) case, the OTH top rank with DA and LR. According to these results, the SML_c4ab classifier scores very well; it is always in the top 5 for all noise injection cases (see Fig. 6.2). Interestingly, this method is looking for the median threshold between the statistics calculated on foreground and background domains (option c4) and it is using the average between the $\Phi_E$ and the $\Phi_E^d$ measures. Consequently, it is seeking for the maximum of compromise between the different measure and
Table 6.3: General benchmarking of classifiers upon Informedness ($I$) in experiment B of noise injection; Omission and Commission Errors and their Standard Deviations

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>$T_{Avg}$</th>
<th>$T_{StdDev}$</th>
<th>Omission Error</th>
<th>Commission Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train Set</td>
<td>0.2216</td>
<td>0.1970</td>
<td>0.7423</td>
<td>0.1880</td>
</tr>
<tr>
<td>SVM_c0a</td>
<td>0.0598</td>
<td>0.0813</td>
<td>0.9391</td>
<td>0.0825</td>
</tr>
<tr>
<td>Grand Total</td>
<td>0.4670</td>
<td>0.1440</td>
<td>0.7547</td>
<td>0.1735</td>
</tr>
</tbody>
</table>

Table 6.4: General benchmarking of classifiers upon Informedness ($I$) in experiment C of noise injection; Omission and Commission Errors and their Standard Deviations

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>$T_{Avg}$</th>
<th>$T_{StdDev}$</th>
<th>Omission Error</th>
<th>Commission Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train Set</td>
<td>0.2216</td>
<td>0.1970</td>
<td>0.7423</td>
<td>0.1880</td>
</tr>
<tr>
<td>SVM_c0a</td>
<td>0.0598</td>
<td>0.0813</td>
<td>0.9391</td>
<td>0.0825</td>
</tr>
<tr>
<td>Grand Total</td>
<td>0.4670</td>
<td>0.1440</td>
<td>0.7547</td>
<td>0.1735</td>
</tr>
</tbody>
</table>

thresholding options.

The majority in top 10 (from 5 to 9 presences, usually on the first positions) shows the superiority of SML classifiers in this test. Beside the OTH classifiers, the Discriminant Analysis (DA) shows the best average performances across all the noise injection cases and levels. In the noise injection case A, the ML
and SVM classifiers show good results, while in the noise B case they all show poor results as compared to the SML. As already mentioned, in the noise case C the DA, LR and ML show good results. By observing the omission and commission errors relative to the built-up class, a more variate situation can be described. Because of the imbalanced reference set in this experiment, they can differ radically from the ranking by Informedness. Generally, all the classifiers show imbalanced error rates with commission error rates quite high. From table 6.5 the average omission and commission error rates obtained in the experiment by all the classifiers are equal to 0.31 and 0.74, respectively. Also in the case of the SML_c4ab top ranking in Informedness a commission error rate in the order of 0.81 was estimated. In some applications, it would be important to show a more balanced behavior of omission and commission errors. From this perspective, the SML_c1a, SML_c1b, and SML_c1ab classifiers show clearly better error rates performances in the noise injection A and C tests, while in the noise test B the SML_c0b, SML_c2a, and SML_c0a classifiers are performing well. Figure 6.3 shows the general behavior of classifiers on error rates. On the whole, the classifiers SML_c0ab, SML_c2a, and SML_c1a show the most balanced error rates in all the tests for all the noise levels. Regarding the balance of the error rates the OTH classifiers perform worse than the SML.

Figure 6.4 depicts the robustness to noise of our SML classifiers in comparison to the OTH classifiers studied in our experiment, for each of the three noise injection tests. In the scale generalization noise test (A), the OTH classifiers show behaviours that are independent of the training set and of the level of noise while SML ones present more influenced curves by the training set (a decreasing slope exists but with a reduced value). For the random thematic noise scenario (B), the OTH classifiers still behave quite independently from

### Table 6.5: General benchmarking of classifiers upon Informedness (I) in experiment. All averaged results of A, B and C tests of noise injection; Omission and Commission Errors and their Standard Deviations

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>I Avg</th>
<th>I StdDev</th>
<th>Omission Error</th>
<th>Commission Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>SML_c4ab</td>
<td>0.6098</td>
<td>0.0567</td>
<td>0.2018</td>
<td>0.0331</td>
</tr>
<tr>
<td>SML_c0b</td>
<td>0.6064</td>
<td>0.0600</td>
<td>0.1897</td>
<td>0.0314</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.5914</td>
<td>0.0546</td>
<td>0.1409</td>
<td>0.0308</td>
</tr>
<tr>
<td>DAT</td>
<td>0.5882</td>
<td>0.0442</td>
<td>0.1982</td>
<td>0.0652</td>
</tr>
<tr>
<td>SML_c2b</td>
<td>0.5872</td>
<td>0.0566</td>
<td>0.2659</td>
<td>0.0602</td>
</tr>
<tr>
<td>MLTK</td>
<td>0.5842</td>
<td>0.0522</td>
<td>0.2881</td>
<td>0.0534</td>
</tr>
<tr>
<td>LR1K</td>
<td>0.5822</td>
<td>0.0476</td>
<td>0.1988</td>
<td>0.0481</td>
</tr>
<tr>
<td>SVM1K</td>
<td>0.5599</td>
<td>0.0686</td>
<td>0.3226</td>
<td>0.0628</td>
</tr>
<tr>
<td>SML_c2ab</td>
<td>0.5495</td>
<td>0.0566</td>
<td>0.3675</td>
<td>0.0610</td>
</tr>
<tr>
<td>SML_c3ab</td>
<td>0.5266</td>
<td>0.0626</td>
<td>0.0926</td>
<td>0.0269</td>
</tr>
<tr>
<td>SML_c3b</td>
<td>0.4773</td>
<td>0.1269</td>
<td>0.4669</td>
<td>0.1750</td>
</tr>
<tr>
<td>SML_c4b</td>
<td>0.4771</td>
<td>0.1269</td>
<td>0.4662</td>
<td>0.1749</td>
</tr>
<tr>
<td>SML_c1a</td>
<td>0.4709</td>
<td>0.1270</td>
<td>0.4663</td>
<td>0.1783</td>
</tr>
<tr>
<td>NBTK</td>
<td>0.4733</td>
<td>0.0120</td>
<td>0.2252</td>
<td>0.0440</td>
</tr>
<tr>
<td>SML_c2a</td>
<td>0.4705</td>
<td>0.0919</td>
<td>0.4884</td>
<td>0.1014</td>
</tr>
<tr>
<td>RTTK</td>
<td>0.4441</td>
<td>0.1896</td>
<td>0.2281</td>
<td>0.0601</td>
</tr>
<tr>
<td>Train Set</td>
<td>0.3204</td>
<td>0.3121</td>
<td>0.5144</td>
<td>0.2386</td>
</tr>
<tr>
<td>SML_c4ab</td>
<td>0.4111</td>
<td>0.1292</td>
<td>0.5645</td>
<td>0.1305</td>
</tr>
<tr>
<td>TRTK</td>
<td>0.3285</td>
<td>0.1189</td>
<td>0.2892</td>
<td>0.0163</td>
</tr>
<tr>
<td>SML_c0a</td>
<td>0.1785</td>
<td>0.1527</td>
<td>0.8154</td>
<td>0.1663</td>
</tr>
<tr>
<td>Grand Total</td>
<td>0.5066</td>
<td>0.1490</td>
<td>0.3107</td>
<td>0.2122</td>
</tr>
</tbody>
</table>
the training set but with worse performances in general. The SML classifiers have in majority a compact behaviour, also independent from the training set and displaying higher performance values for high noise levels (improving the quality of the training set for these extreme noise levels). In the case of spatial displacement injected noise (C), the Informedness values of the OTH classifiers are again quite independent of the ones presented by the degraded training set but have small variations according to the noise level. The SML classifiers behaviours are smoother and have only a reduced correlation with the behaviour of the training set, but their spread is higher than in the B test results. It is to be noted that in the case of SML classifier, there is a group of classifiers (SML_c4ab, SML_c0b, SML_c4b, SML_c2b and SML_c2a) showing always top rankings.

Figure 6.2: Performance (averaged Informedness) for SML and OTH classifiers in the A, B, C cases of noise injection and for an averaged behavior of all noise levels, ALL (bottom right)
6.3 On ROC analysis

In this section, the overall results of a Receiving Operator Characteristics (ROC) analysis are presented, assessing the absolute discrimination power of the decision maps as produced in the feature space by two classifiers, the SML and the maximum likelihood (ML), the latter being based on a well-established statistical model. In this assessment, the best cutoff value of the decision map was searched by optimization techniques using the unbiased reference set as test set. The results discussed here are derived from optimization of the Informedness metric; consequently they express the best Informedness achievable given the specific decision map discriminating the foreground from background information in the feature space. In the case of the SML, the decision map under test are the $\Phi_a^E$, $\Phi_b^E$, and $\Phi_{ab}^E$ measures as defined in the section 3.3. In the case of the ML classifier, six trials have been made. Specifically the trials involved two options concerning the amount of training samples, and three options regarding the use of the posterior probabilities estimated by the ML classifier. Concerning the sampling, a stratified sampling schema with 1000 samples for each class (1K) and a sampling schema involving the whole available samples (EntireSet) were tested. Concerning the posterior probabilities, the probability of the background ($c_0$) and foreground ($c_1$) classes were tested. Moreover, a combination of the posterior probabilities ($c_0$, $c_1$) was also taken into account by imitating the basic $\Phi_E^k$ formula and substituting the negative and positive frequencies, correspondingly. In the test this is encoded as ENDI. Table 6.6 shows the aggregated results of the assessment for all the noise injection experiments (A, B, C) and all the levels of noise. The Informedness scores are summarized in Figure 6.5. Few main results can be observed i) the $\Phi_a^E$, $\Phi_b^E$, and $\Phi_{ab}^E$ measures show the very same results and they obtain the best performances in the experiment; ii) the tested alternative sampling schema are mostly irrelevant in the ML case; anyway the random stratified sampling provide always slightly better results than the use of the whole (imbilanced) set; iii) the best combination of the ML posterior probabilities was the ENDI schema, while the worst ML performances were provided by the use of the $c_1$ probability alone.
Figure 6.4: Robustness of SML and OTH classifiers in function of level and injection type of noise

About the first point, it is worth noting that the experiment confirms the fact that the decision map produced in the feature space by SML techniques is more
precise (or adherent of the real distribution of the data) than the one estimated using a statistical parametric approach as ML. This is creating an absolute greater discrimination power of the SML decision map. Secondly, it is interesting to note that the different ENDI options are delivering the same decision map and consequently the choice of the correct threshold for classification purposes becomes more important.

![Figure 6.5: Ranking by best Informedness obtained by ROC analysis of SML and ML decision maps vs. unbiased test sets](image)

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Opt Overall Acc Avg</th>
<th>Opt Overall Acc StdDev</th>
<th>Opt Kappa Avg</th>
<th>Opt Kappa StdDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMLronicsb</td>
<td>0.8600</td>
<td>0.0670</td>
<td>0.9603</td>
<td>0.0031</td>
</tr>
<tr>
<td>SMLronicsab</td>
<td>0.8658</td>
<td>0.0671</td>
<td>0.9603</td>
<td>0.0031</td>
</tr>
<tr>
<td>SMLronicsa</td>
<td>0.8646</td>
<td>0.0684</td>
<td>0.9603</td>
<td>0.0031</td>
</tr>
<tr>
<td>ML1Kronics</td>
<td>0.8377</td>
<td>0.0408</td>
<td>0.9640</td>
<td>0.0014</td>
</tr>
<tr>
<td>MLEntireSetronicsb</td>
<td>0.8620</td>
<td>0.0670</td>
<td>0.9603</td>
<td>0.0031</td>
</tr>
<tr>
<td>MLEntireSetronicsab</td>
<td>0.8658</td>
<td>0.0671</td>
<td>0.9603</td>
<td>0.0031</td>
</tr>
<tr>
<td>MLEntireSetronicsa</td>
<td>0.8646</td>
<td>0.0684</td>
<td>0.9603</td>
<td>0.0031</td>
</tr>
<tr>
<td>ML1Kcl0</td>
<td>0.5837</td>
<td>0.0408</td>
<td>0.9540</td>
<td>0.0014</td>
</tr>
<tr>
<td>MLEntireSetcl0</td>
<td>0.5739</td>
<td>0.0375</td>
<td>0.9549</td>
<td>0.0007</td>
</tr>
<tr>
<td>ML1Kcl1</td>
<td>0.3274</td>
<td>0.0217</td>
<td>0.9516</td>
<td>0.0002</td>
</tr>
<tr>
<td>MLEntireSetcl1</td>
<td>0.3185</td>
<td>0.0179</td>
<td>0.9523</td>
<td>0.0004</td>
</tr>
<tr>
<td>ML1Kcl2</td>
<td>0.1856</td>
<td>0.1232</td>
<td>0.9521</td>
<td>0.0003</td>
</tr>
<tr>
<td>MLEntireSetcl2</td>
<td>0.1841</td>
<td>0.1207</td>
<td>0.9521</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Table 6.6: ROC analysis for optimal Informedness, optimal overall Accuracy and optimal Kappa measure (averaged values and standard deviations)

### 6.4 On applicative examples

In this section, few examples of SML applied to different type of satellite data and different input image features are presented and shortly commented. The examples include the processing of i) Spot 2.5m resolution data for updating topographic cartography, ii) Sentinel1 5m resolution data for built-up areas recognition, and iii) Sentinel2A 10m resolution multispectral data for built-up areas and water surfaces.

**Spot 2.5m**

Figure 6.6 shows some results of a test aiming to evaluate the use of Spot image data for the updating of cartography at scale 1 : 10,000. The input image data
were available through the Copernicus programme as 2.5m resolution panchromatic data covering the towns of Pisa and Lucca, in the Tuscany region of Italy. They were obtained by pan-sharpening of multispectral 10m resolution data and local filtering for visual enhancing\footnote{Specifications of view services for GMES Core003 VHR2 coverage \url{https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/specifications-view-service-gmes-core003-vhr2-coverage}}. From the point of view of automatic image information extraction task, all these operations are adding noise to the original data. Moreover, the specific pre-processing parameters are unknown, so they must be modeled as black box by the classifier. In this example, building footprints at scale $1 : 10,000$ of year 2000 are compared with image data of year 2011 with the purpose to extract automatically the changes between the two years. The image features provided as input to the SML classifier are six including morphological and multispectral attributes used during the sequencing phase. The first three attributes encode the characteristic-saliency-leveling (CSL) multi-scale morphological decomposition scheme based on connected components of the image, as defined in [POG12] and using a parameter $\lambda = [0, 80, 800, 8000]$ pixels. The input of the CSL is the luminance calculated as maximum between the available red and green channels of the Spot data. The three multi-spectral attributes are the digital numbers encoding the green, red, and infrared bands available in the pan-sharpened image. The data was originally encoded in 8 bits (256 levels): except for the morphological characteristic attribute of the CSL that is already symbolic, all the other attributes have been re-quantized in 8 levels ($q = 32$) during the sequencing. Figure 6.6 upper left is the overview of the image data processed, while the upper right is a detail of the same data. On the bottom the results: in white there are the stable buildings, also used as training set by the SML process. In red and green regions, there are the strong disagreements between the $\Phi_E$ and the expectation from the reference set $R$. Specifically, set to $\pm 0.85$ the cutoff defining a strong $\Phi_E$, the red $\leftarrow (\Phi_E < -0.85) \cap (R = \text{true})$ and green $\leftarrow (\Phi_E > 0.85) \cap (R = \text{false})$. In other words, red encodes dismantled buildings while green encodes new buildings in the time interval between the cartography update and the image data. Note how the use of morphological descriptors and relatively high resolution input imagery during the classification allows to map the changes on single built-up structures with approximation of their shape (lower right in the image).

\textbf{Sentinel1 5m}

Figure 6.7 shows results for a Sentinel 1 (S1) strip-map mode image of the year 2015. The first row shows the input data (HV polarization), the middle row shows the training set used for the SML, and the bottom row shows the results of the built-up classification. An overview (left) and a particular (right) of the same area is shown in the image: they have an approximate size of 140x140km and 10x10km, respectively. The S1 data was provided with HH and HV polarization and 5m of spatial resolution after the geo-coding. The training set used in the example are the class 1.1.1 \textit{Continuous urban fabric}, 1.1.2 \textit{Discontinuous urban fabric} and 1.1.3 \textit{Industrial or commercial units} of the Corine Land Cover product of year 2000\footnote{Copernicus Land Monitoring Services \url{http://land.copernicus.eu/pan-european/corine-land-cover}}. In this example, the noise present in the training
set is composed by three main parts: time gap (15 years), scale/generalization gap (100m, 25 hectares minimal mapping unit size vs 5m of S1), and semantic differences. On the latter, it is worth noting that the LC classes (especially 1.1.2) include large not built-up surfaces that are reported by 5m resolution sensor scale. During the sequencing phase of the SML, seven attributes where encoded: they are the HII and HV back-scattering amplitude values, two derived textural measurements, and three measurements describing the orography as extracted by filtering the digital terrain model. Specifically, the *pantex* textural measurement was applied as defined in \[PGK08\] with a window size of 9x9 pixels (45x45 m$^2$) and a displacement vector distance of 1 pixel with 8 directions. The orography was described by the *aspect*, *slope* and *top* parameters estimated from the Shuttle Radar Topographic Mission (SRTM) data in the 90m resolution version downloaded from the Global Land Cover Facility (GLCF) server. The parameter *top* was estimated using morphological opening tophat filtering with a structuring element of 3x3 pixels. All the seven attributes of the sequencing were encoded in 16 levels during the SML. The processed

\[3\text{SRM on the GLFC} \text{http://glcf.umd.edu/data/srtm/}\]
scene present a large part on mountainous area on the North of Milano who is placed in the center. Mountainous areas with remarkable slope and crest-lines may produce false alarm in automatic detection of built-up areas both based on back-scattering amplitude or derived textural criteria. These commission errors are usually reduced by post-classification masking techniques, based on domain expert knowledge and metadata information. Note how the SML classifier find the systematic relation between the orography pattern and the specific illumination angle of the back-scattering radar data and compensate it without the necessity to set specific ad-hoc post-classification masking rules.

Sentinel2 10m

Figure 6.8 shows some results of a preliminary test classifying the very first Sentinel2A (S2A) data made available by the European Space Agency (ESA) in June 2015. In the image, a small portion of S2 data processed is represented: it covers the town of Porto Viro, in the Po river delta (Italy). The image has a size of of 25x25 km$^2$. The purpose of the test was to evaluate the S2 capacity for solving global high-resolution land cover tasks. The bands 2, 3, 4 and 8 of the multispectral S2 sensor were used as input for the example. They are placed in the central wavelength of 490, 560, 665, and 842 nm, respectively, with a spatial resolution of 10 meters. Some spatial misalignment between the input spectral bands was observed in these very early data and was not compensated: consequently it must be accounted as noise in the classification process shown here. The input features used for the classification are ten: they include the four input spectral bands, four derived textural measurements, and two morphological filters derived from the band 4. Specifically, a pantex textural measurement as defined in [PGK08] was applied to all the four spectral bands separately. The applied operative parameters were a window size of 5x5 pixels (50x50 m$^2$) and a displacement vector distance of 1 pixel with 8 directions. The two morphological features are derived from the opening and closing residuals (top-hat morphological transforms) using a structuring element of 5x5 pixels. All the ten attributes were quantized in 16 levels during the sequencing phase of the SML. The training set used for the example is shown in top right of the Figure 6.8. As in the previous example, they are extracted from Corine LC data source of year 2000, used in the gridded version at 100 meter resolution. In white are represented the LC classes related to built-up and in blue the classes related to water. As in the previous example, large time gaps, scale gap and semantic inconsistencies are included in the training set as regarding the classification of the specific satellite image. In the lower part of Figure 6.8 the SML outputs for the classes built-up and water are displayed in white, respectively, on the left and on the right. It is worth noting in the example how the SML outputs may significantly contradict the information provided in the training set if the evidences collected from the image data sufficiently support the decision. In both the built-up and the water example important information was discovered contradicting the training set: in the built-up case scattered settlement structures were discovered even if neglected by the reference, while in the water case a large water surface (top right) not present in the reference was discovered by the SML associative analysis. Moreover, thin water channels

[^4]: Sentinel-2 delivers first images [http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-2/Sentinel-2_delivers_first_images](http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-2/Sentinel-2_delivers_first_images)
Figure 6.7: Detecting built-up areas using Sentinel1-stripmap data input and SML. Left: from top to bottom the input image (HV polarization), the training set (Corine LC) and the classification results - the image is of 140 km side. Right: detail of the same information - the image of 10 km side.

and small water bodies are discovered as well. Analogously to the built-up case these thin structures are the value added provided by the better scale of the image data respect to the training set.
Figure 6.8: Detecting water and built-up areas using Sentinel 2A input data and SML. Top left: a detail of the S2 input data. Top right: training set from Corine LC, built-up areas in white and water in blue. Lower left: output of the SML classification for built-up areas. Lower right: output of the classification for water surfaces.
Chapter 7
Discussion

The experiment tested the capacity of the classifiers to detect complex information using noisy training sets as input. In particular, the target information to be detected was connoted by large spectral mixture and inner heterogeneous spectral characteristics (built-up areas). According to the results processed during the experiment, the new SML method outperforms parametric classifiers (Maximum Likelihood, Naive Bayes, Discriminant Analysis, Logistic Regression) and non-parametric classifiers (Decision Tree, Random Forest, Support Vector Machine) as well.

The strength of the SML classifier is the fact that is largely agnostic both with respect to the statistical distribution of the input data and to the relations between the image data and the information to be detected. The decision map translating image data instances to the class memberships is purely based on observed joint occurrences of image data and reference information in the specific universe under processing. In contrast to common practices in remote sensing data classification, very little or no assumptions need to be postulated for performing the classification task. Both assumptions on data statistical distribution or hypothesis on physical causal mechanisms linking observed data with target information can be largely avoided, allowing the machine to discover them during the association analysis process. This fact allows the radical reduction of the complexity in the classification design and parameter tuning phases, and consequently leads to an increase in the level of automatic control of the process. The same fact largely facilitates the generalization of the identical classification process to image data collected by different sensors and different local data collection conditions (season, illumination, building practices, materials, settlement patterns, natural background). This is a general feature that can be valid in any BRSD scenario involving the detection of complex information in the image data.

Furthermore, also in contrast to common practices, SML classifiers do not need to postulate the continuity of the functions discriminating the target information in the image feature space. This fact permits to handle more realistically and accurately classification tasks involving targets connoted by large spectral mixture and inner heterogeneous spectral characteristics. In the specific test, these are the typical characteristics of the human settlements areas as described by remotely sensing data, but they can be extended to any other target information for which causal-deterministic models are not effective or not available.
Finally, SML classifiers are computationally much more efficient if compared to other non-parametric approaches requiring random iterative searches in the image feature space and relatively complex optimization function assessment as Support Vector Machine (SVM) or Random Forest (RF). Computational efficiency in big data analytics scenarios largely facilitates the phase of the inception and design of the new classification process, allowing large number of low-cost tests involving the whole data universe instead of a sample of it.
Chapter 8

Conclusions

In this manuscript a new method for supervised classification of remotely sensed data was presented in analogy to classification techniques used in bio-informatics and genome expression analysis. The method is defined in the frame of Symbolic Machine Learning (SML) and is based on two main steps i) data quantization-sequencing and ii) associative analysis. The method is designed for working in Big Remote Sensing Data (BRSD) scenarios. They are connoted by co-presence of the following characteristics: i) they are made by a large number of data granules (scenes), ii) they are made by heterogeneous sensors and iii) they are mapping a large variety of different geographical areas in different data collection conditions.

An experiment was set in order to understand the behavior of the new method under specific data and noise conditions and to compare it with some other supervised classifiers that are well established in the remote sensing community. In particular, the new classifier was compared with Maximum Likelihood, Naïve Bayes, Discriminant Analysis, Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine classifiers. The classifiers were tasked to recognize built-up areas in Landsat satellite imagery using exclusively spectral input features; thus, in the test, the target information was connoted by large spectral mixture and inner heterogeneous spectral characteristics.

During the experiment three different noise injection procedures were performed affecting the training set used for the supervised classification. Specifically: scale generalization, random thematic noise, and spatial misalignment. They are the most common sources of noise in using open source global data as training sets.

According to the test reported here the new SML classifier performed very well providing outputs with quality comparable or better than the other classifiers. Furthermore, the better performances were released with a much less expensive computational cost. As a consequence, in accordance with the Occam’s razor principle, the SML classifier was evaluated as the best available solution in the specific data scenario under consideration.

Few examples of use of the new classifier were provided using different satellite sensors in input. Specifically: Spot 2.5m optical sensor, Sentinel 1 5m radar sensor, and Sentinel 2 10m optical sensor. Different set of input features were used in the examples including radiometric, textural, and morphological multi-scale descriptors derived from the image data. Additional Digital Surface Model
derived features were also integrated in one example. In these examples, two important characteristics of the new method were highlighted: i) the capacity to handle seamlessly heterogeneous input features including categorical and continuous information with different statistical distributions and different physical meanings and ii) the capacity to handle noisy input data. In the method presented here this is achieved without the necessity to set new ad-hoc explanatory models and without the necessity to implement expensive expert-driven parameter tuning processes.

The experiences described in this paper supported the decision to apply SML as core classification techniques for solving the global processing of Landsat data involving multi-sensor, multi-temporal, large-scale data collections as described in [Pes+15].
Bibliography


Europe Direct is a service to help you find answers to your questions about the European Union. Freephone number (*): 00 800 6 7 8 9 10 11

(*) Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed.

A great deal of additional information on the European Union is available on the Internet. It can be accessed through the Europa server http://europa.eu.

How to obtain EU publications

Our publications are available from EU Bookshop (http://publications.europa.eu/howto/index_en.htm), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

European Commission
EUR 27518 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen

Title: Benchmarking of the Symbolic Machine Learning classifier with state of the art image classification methods

Authors: Martino Pesaresi, Vasileios Syrris, Andreea Julea

Luxembourg: Publications Office of the European Union

2015 – 46 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series – ISSN 1831-9424 (online)


doi:10.2788/638672
JRC Mission

As the Commission’s in-house science service, the Joint Research Centre’s mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

Serving society
Stimulating innovation
Supporting legislation

doi:10.2788/638672