GWENN-SS: a simple semi-supervised nearest-neighbor density-based classification method with application to hyperspectral images

Claude Cariou, Kacem Chehdi, Steven Le Moan

GWENN-SS: a simple semi-supervised nearest-neighbor density-based classification method with application to hyperspectral images

Claude Cariou[a], Kacem Chehdi[a] and Steven Le Moan[b]

[a] Univ Rennes, CNRS, IETR, UMR 6164, 6 rue de Kerampont, F-22300 Lannion, France
[b] Department of Mechanical and Electrical Engineering, Massey University, Palmerston North, New Zealand

ABSTRACT

In this communication, we address the problem of semi-supervised classification under conditions where (*i*) learning samples are available only for specific classes and potentially mislabeled, and (*ii*) the actual number of classes is unknown. For this, we propose a semi-supervised extension of a Nearest-Neighbor - Density Based clustering method, namely the Graph WatershEd using Nearest Neighbor (GWENN) method. We show how an incomplete, erroneous learning sample (LS) set can be incorporated in the algorithm in order to produce efficient labeling decisions partly guided by *a priori* information, and to discover new classes and correct mislabeled objects. The efficiency of the proposed method, named GWENN-SS, is demonstrated experimentally. We first evaluate its robustness with simulated data for which an erroneous and incomplete LS set is given. We then assess the reliability of GWENN-SS on real hyperspectral images and we show that it can outperform a recent similar semi-supervised approach.

Keywords: Classification, semi-supervised learning, nearest neighbor, density estimation, hyperspectral image.

1. INTRODUCTION

Classification is a core task in data processing, due to its potential to summarize the informational content of a data set, and to further facilitate its interpretation. This is particularly true in the remote sensing field, especially for large-scale hyperspectral images which represent a huge amount of detailed information both in the spatial and the spectral domain.

Here, we address the problem of semi-supervised classification. Semi-supervised classification represents a broad family of methods which require at least some prior information about the data set to partition, e.g. the number of classes it comprises, or a set of labeled samples used to learn the decision rules, which are then generalized to the other (unlabeled) samples. These classification approaches are useful for pixel interpretation in remote sensing images and contribute to reducing the substantial cost associated with acquiring the ground truth. However they require high-quality learning samples as well as representative samples from each significant class in the scene. Because ground truth acquisition is error-prone,[1] there is a need for approaches which can call into question prior labeling, and also discover new clusters beyond the ones pointed out by the learning samples. With this objective, it can be useful to consider the help of unsupervised clustering approaches.

In the present work, we investigate the case where learning samples are provided, along with the data objects to classify. Therefore, a (small) number of objects in the data set are given *a priori* a specific label. More precisely, we consider the semi-supervised classification problem under conditions where:

- some learning samples are available for specific classes, but may be mislabeled;
- the actual number of classes is unknown, i.e. some classes are not represented in the learning set.

Further author information:
E-mail: claude.cariou@univ-rennes1.fr, Telephone: +33 296 469 039

Such conditions are of real practical interest to end-users who may face the availability of a biased, mislabeled ground truth[1] but still want to provide classification maps based on data-driven objective criteria, including the discovery of potential classes not represented in the learning sample set.

In the classification literature, such a problem lies in the category of Positive and Unlabeled - biased Negative data classification,[2] in reference to the multi-class extension of the binary classification problem in which (*i*) some classes are not represented by learning samples (Positive and Unlabeled - PU) and (*ii*) the other classes have potentially non representative or even no learning samples (biased Negative - bN). This framework is quoted as semi-supervised novelty detection by Blanchard *et al*.[3] Also, Scott *et al*[4] extend the problem to mislabeled learning samples occurring in binary classification. However all these approaches are based on some mixture model for the underlying data distribution, governed by a prior class distribution and model-based identifiable class-conditional distributions.

On the other hand, nearest-neighbor - density-based (NN-DB) clustering methods such as ModeSeek,[5] knnDPC,[6] knnClust,[7] GWENN,[6,8] are simple, deterministic, yet effective unsupervised methods, which can outperform state-of-the-art clustering methods like Kmeans and Fuzzy C-Means. These methods require the availability of a K-NN graph, weighted by the set of distances of objects to their K nearest neighbors. NN-DB methods are well adapted to discover non-convex clusters as often observed in high dimensional data, especially in hyperspectral images (HSIs). They have been improved to allow partitioning pixels in very large HSIs owing to a multiresolution scheme,[8] and it was also shown that a mode weighting improvement of GWENN provided encouraging results over a very large HSI.[6] Among the NN-DB methods, the improved GWENN method has shown to provide a good trade-off between clustering quality and computational complexity.

In this communication, we investigate the potential to incorporate *a priori* information in the NN-DB clustering procedures, in the form of learning samples (LSs). More precisely, focusing on GWENN, we propose to introduce LS in the original algorithm. To this end, GWENN is modified in a way to drive the labeling decisions with the help of these LS. The resulting algorithm, named semi-supervised GWENN (GWENN-SS) shows two interesting properties: (*i*) it can yield a number of clusters greater than the number of labels available within the LS set; this is a great advantage over most supervised methods which are generally not able to infer the presence of classes not referenced by labeled LS; (*ii*) it is able to correct for mislabeled LS; this property is also important to the end-user who may question the reliability of available ground truth data for learning.

The paper is organized as follows: in Section 2, we provide a brief overview of related works in semi-supervised classification, pointing out the need for more flexible methods which can (*i*) extend the classification to unknown classes and (*ii*) correct for learning samples errors; in Section 3, we detail the implementation of the proposed method; Section 4 describes an experimental study including its application to pixel classification in hyperspectral images; we conclude in Section 5.

2. RELATION TO PREVIOUS WORKS

In a recent work,[6] a set of nearest-neighbor density-based clustering methods was proposed and compared in the context of large scale hyperspectral image pixel clustering. The conclusions of this work were that NN-DB methods are interesting alternatives to a classical (semi-supervised) clustering methods such as FCM, and provide results which are faster, more accurate, and deterministic. These methods only rely on the computation of a nearest neighbor (NN) graph from the original data set, obtained by a greedy NN search procedure. This graph is governed by a single parameter K, i.e. the number of nearest neighbors (NNs) which is set by the user. Also, a model of local (point-wise) density is adopted in Ref. 6, still depending on the same parameter K:

$$\rho(\mathbf{x}_m) \propto \frac{K}{\sum_{\mathbf{x}_j \in K\mathrm{NN}(\mathbf{x}_m)} d(\mathbf{x}_m, \mathbf{x}_j)} \quad , \quad 1 \leq m \leq N, \tag{1}$$

where N is the number of objects, and $K\mathrm{NN}(\mathbf{x}_m)$ is the set of nearest neighbors to $\mathbf{x}_m$ according to the Euclidean distance $d(.,.)$. Using the K-NN graph and the density model above, the four NN-DB methods described in Ref. 6, represent different partitioning strategies by aggregation of objects based on the interaction between NNs' local densities. These approaches have two important features: the first one is that they do not require the

explicit knowledge of the number of clusters which form the partition, and the second one is the stability of the number of clusters within a large range of K.

Among the NN-DB clustering methods, GWENN[8] and its variant GWENN-WMODE[6] actually only differ in the way the labeling decisions are taken. By accounting for pointwise local densities as in Eq. (1), GWENN-WMODE offers a higher quality of clustering results w.r.t. the baseline method GWENN. These approaches sequentially assign labels to objects sorted by decreasing density, based on the mode (or weighted mode) of their nearest neighbors which were previously labeled.

The motivation of this work is therefore to address semi-supervised classification with such approach, by proposing a labeling decision rule based on some available prior knowledge (labeled learning samples) while keeping the spirit of density-based decisions.

3. PROPOSED SEMI-SUPERVISED METHOD: GWENN-SS

In the present work, we seek to introduce *a priori* information into GWENN (more precisely GWENN-WMODE) in the form of a learning set, i.e. a subset of the objects to classify having previously assigned labels. This prior labeling is assumed to be available from external information sources, typically the result of a field work in the case of a remotely imaged scene. However, such a prior information may not be free of errors, and we want our approach to be able to correct such errors automatically, by calling into question the initial prior labels. In the sequel, we will refer to this information as a LS set.

For this, we propose that the semi-supervised algorithm follows the principles below:

- The order of objects' processing must remain consistent with their specific density. By doing this, we do emphasize and give priority to the data consistency rather than to the quality of the LS set.
- The labeling decision relative to an object is only based on prior information available from its NNs, i.e. their labels, whether these are imposed through the LS set, or previously assigned by the algorithm.

The pseudo-code of GWENN-SS is given in Algorithm 1. The algorithm first computes the K-NN graph (distances and indices of NNs), then computes the local density using Eq. (1), and sorts the latter to provide an order of object processing. Then the method has two sequential passes, a main one producing a rough classification result, and a secondary one aiming to refine this result. In the main pass, the key modification w.r.t. the original GWENN-WMODE relies in the labeling decision based on available NNs having non-zero labels, either because they have been previously processed, or because they belong to the LS set. However, in the second pass, the constraint on LS labels is totally relaxed from the decisions, and only the labels issued from the main pass and the pointwise local densities are accounted for in the classification refinement. Notice that the pointwise labeling decision (`wmode`) is similar to the one in GWENN-WMODE and is based upon the mode of the labels assigned to its K NNs, weighted by their respective densities, i.e.:

$$c(\mathbf{x}_m) = c_m = \arg\max_{c \in \mathbf{c}_Q} \sum_{\mathbf{x} \in Q} \mathbf{1}_{[c(\mathbf{x})=c]} \cdot \rho_{\mathbf{x}} \quad , \tag{2}$$

where c is a class label, and $\rho_{\mathbf{x}}$ is defined in Eq. (1). However, differently from GWENN-WMODE, Q is the set of each object's NNs which have been previously processed or belong to the LS set.

4. EXPERIMENTAL STUDY

In this section, we describe a few experiments to assess the proposed semi-supervised method. In each experiment, we considered two ground truths. One of them is the actual, absolute ground truth, assumed free of any labeling error; it will be used for the evaluation of the classification results only. The other one is a simplified ground truth which comprises mislabeled objects or pixels, and also ignores some classes of the actual ground truth, i.e. no labeled object or pixel of these classes are available in the training set. This simplified ground truth is actually the LS set used for semi-supervised classification.

4.1 Experiment 1: Toy dataset

We first conducted an assessment of GWENN-SS on a toy dataset consisting of $N = 600$ 2-D data objects distributed following a Gaussian mixture model. The data points and their actual membership are shown in Fig. 1-(a). The true means of the three Gaussian distributions are $(0, 0), (0, 1), (1, 1)$ and each of them has a standard deviation of 0.3 in each dimension, and zero covariance. Fig. 1-(b) displays the LS set. In this LS set, one of the three classes is not represented, and 20% of the remaining points are randomly selected as learning samples. In addition, these 80 learning samples are subject to 30% labeling error. For the unlabeled points, Fig. 1-(b) also shows the magnitude of the local density as shades of gray.

The K-NN graph of the data set is first computed with $K = 40$. Notice that the choice of the optimal neighborhood size K is not an issue in the present work. GWENN-SS (with weighted mode) is then applied to the data set with the erroneous LS set, providing the classification result shown in Fig. 1-(c). In can be seen that the original classes of the LS set can be correctly recovered, as well as the extra class which was not referenced in the LS set. The classification accuracy, assessed by computing a confusion matrix between the output classification and the actual GT,[9] reports an overall accuracy of 86.83%, which is quite satisfactory in view of the large overlap between classes and the LS error rate.

4.2 Experiment 2: *Salinas* HSI

The *Salinas* HSI was acquired by the AVIRIS sensor in 1998 over agricultural fields. It has a spatial size of 512×217 pixels, and 204 spectral bands. The reference map is composed of 16 vegetation classes as shown on Figure 2. A full KNN graph (with $K = 1000$, i.e. less than 1% of the number of objects) was computed over the whole HSI (111104 pixels).

Erroneous GT maps were constructed as follows: firstly, we selected a number of 10 classes among the 16 referenced ones (C2, C3, C4, C5, C6, C7, C8, C9, C15, C16). Secondly, for each of these classes, we manually selected one pixel position around which we grew square patches of various sizes, from 7×7 to 39×39, while maintaining the original class label in each patch. With increasing patch size, the overlap with spatially neighboring different classes of the actual GT map is likely to occur. This is illustrated in the first row of Figure 3 where footprints of the patches are superimposed on the actual GT map: for small size patches (7×7), all patches comprise pixels from the correct, expected class, whereas for larger patches (27×27 or 39×39), the learning set includes pixels which do not belong to the same original class, according to the actual GT map, but also pixels which are not labeled at all.

The erroneous GT maps with various patch sizes were used as input for GWENN-SS and ss-Kmeans++.[10] ss-Kmeans++ is a recent semi-supervised extension of Kmeans++ which can also deal with semi-supervised information, i.e. LS sets. In our experiments, GWENN-SS was applied first for each patch size, providing as output a classification map with a number of classes C greater than the number of labels available in the training set (10 classes). ss-Kmeans++ was then applied using the same erroneous GT map and setting the number of output classes to C accordingly. Since ss-Kmeans++ involves random initialization of class centroids, we performed 10 runs of this method for each patch size. The corresponding accuracy indices (Overall Accuracy, Average Accuracy, kappa index) were derived as above.[9] Other internal indices (Mean Square Error and CVR[11]) were also computed to assess and compare the consistency of classification results. Table 1 displays the classification performance indices obtained with the two methods as a function of the patch size. The number of classes obtained at the output of GWENN-SS is given, as well as the per-class error in the labels of the LSs induced by the increase in patch size. These errors do not account for unlabeled LSs. The C4 Class is the most impacted among all the LS classes since it has a narrow extent along one spatial dimension in the corresponding GT map; the labeling error for this class goes from 0% for the 7×7 patch to more than 60% for the 39×39 patch. Notice that incorrect labels issued from several neighboring regions of the actual GT contribute to the class-specific error: for instance, the 39×39 patch of class C5 incorporates LSs from classes C4, C6 and a few samples of class C8 as of the actual GT map.

The analysis of Table 1 first shows a neat superiority of GWENN-SS versus ss-Kmeans++ in terms of external indices (OA, AA, kappa) whatever the patch size. The evolution of these indices with the patch size exhibit optimum OA and kappa for a patch size of 27×27 for GWENN-SS, whereas for ss-Kmeans++ an optimum

is found for a size of 31×31. The CVR index is also significantly better (lower) for GWENN-SS than for ss-Kmeans++ in all cases, which reveals a better ability of GWENN-SS to form coherent classes in the high dimensional space. However, the MSE index is lower for ss-Kmeans++; this result indicates that MSE, and to a more general extent the classification approaches which aim to optimize it (among which centroid-based methods) are not appropriate to HSI data due to the fact the class-conditional distributions are elongated and/or non-convex. In most cases, GWENN-SS is able to retrieve the correct labels of the actual GT, as well as to detect new classes which are consistent with it. Figure 3 shows some classification maps given by ss-Kmeans++ and GWENN-SS for various patch sizes. Notice that the maps shown for ss-Kmeans++ correspond to the result providing the best kappa index among the 10 runs of this method. Focusing on classes C1 and C2, where only C2 is present in the LS set, one can notice that for patch sizes 7×7 and 27×27, ss-Kmeans++ is able to discover a new class corresponding to C1, but with some confusion with C2; however, for a larger patch size (39×39), both classes are merged into C2. Comparatively, GWENN-SS provides much well-conditioned results, and can unveil the C1 class whatever the patch size, with very little confusion. As another example, class C3 is incorrectly merged with other classes according to the actual GT map by ss-Kmeans++; besides, even if GWENN-SS could not maintain the correct label for this class for the 7×7 patch, a consistent class was discovered in the corresponding region. Moreover, for larger patches, the label assigned was correctly maintained and propagated only to unlabeled pixels, without interference with other classes. Another important feature of GWENN-SS is its ability to produce classification maps which are spatially almost regular, in comparison with ss-Kmeans++. This property is all the more surprising as the classification algorithm does not account for spatial relationships between adjacent pixels. While it has not been proved yet, we conjecture that this is another consequence of the capability of a NN-DB classification method like GWENN to cope with elongated/non-convex distributions.

We have also compared these results with those obtained by a standard SVM classifier trained with the same LS set, to see the influence of introducing mislabeled pixels. For this we used a Gaussian RBF SVM model with aperture and penalty parameters estimated from 5-fold cross-validation. The model was trained with preprocessed data, each data feature (i.e. data value for each spectral band) being normalized in the $[0, 1]$ range. For a patch size of 7×7, 27×27 and 39×39, the cross-validation provided OAs of respectively 98.37%, 96.82% and 93.98%. This result is consistent, the decrease in OA being related to the increase in LS error with patch size. The corresponding classification results are also shown in Figure 3. Of course this approach is totally supervised since the predicted labels are bounded to the LS labels solely. In this sense, the prediction of some classes will depend on the closeness of the data features to the ones belonging to classes referenced in the LS set. This is the case for the pixels belonging to C1 according to the actual GT map, which are predicted as belonging to C2; note that this result was also obtained with ss-Kmeans++ for large patches as seen above. Moreover it is interesting to observe the behavior of the SVM prediction for large size patches (27×27 and 39×39) on the obtained classification maps: for classes C4 and C5, which are the most erroneous, the predicted labels tend to form larger regions around the corresponding LS patches, and the spatial coherence of these regions with regards to the actual GT map is lost. Besides, this method better preserves the existence of class C8 than GWENN-SS or ss-Kmeans++, most likely because the corresponding LS patches do not contain mislabeled pixels. Actually, classes C8 and C15 (*Vineyard untrained* and *Grapes untrained*) comprise pixels with very close spectral signatures, and are hardly separable.[9]

4.3 Experiment 3: *Guardamar* HSI

In this experiment, we used a HSI acquired with our AISA Eagle sensor in 2010 over the city of Guardamar, region of Murcia, Spain, for purposes of invasive cane detection and discrimination. Its size is 300 × 300, and it is composed of 62 spectral bands covering the [400, 960] nm range at 10 nm spectral width per band. Figure 4-(a) shows a color composition of the HSI. The two classes referenced in the GT and shown in Figure 4-(b) are *Phragmites australis* (yellow) and *Arundo donax* (cyan). The semi-supervised methods were fed with LS patches of size 31×31 for each class. Note that the LS patch for class *Arundo donax* does not overlap the corresponding GT area.

As previously, we applied GWENN-SS and ss-Kmeans++ to this HSI. GWENN-SS (with $K = 100$) provided a total of 18 classes, and this result was used to initialize ss-Kmeans++. The OA and AA given by GWENN-SS are respectively 72.93% and 68.41%, whereas for ss-Kmeans++ the average OA and average AA (± standard deviations) are only 31.54±12.71% and 20.08±7.79%. Figures 4-(c) and (d) display the classification maps

obtained by the two approaches. Note that the ss-Kmeans++ result shown is the best among the 10 runs of the method. Visually, the classes of interest are correctly retrieved by GWENN-SS compared to ss-Kmeans++. However, both methods were able to clearly identify new putative classes such as water, roads, trees, cultivated and bare soil.

5. CONCLUSION

In this communication, we addressed the problem of semi-supervised classification under conditions where a learning sample set is incomplete (i.e. it does not reference all the classes for be found), and erroneous (i.e. some learning samples are mislabeled and possibly far in distance from their original class). This problem was tackled by a new method based on a Nearest-Neighbor - Density-Based (NN-DB) approach. More precisely, we investigated the potential of GWENN – originally an unsupervised clustering method –, to incorporate learning samples in its core labeling decision rule. It was found that the modified labeling rule can be set up in a straightforward manner, realizing a fair balance between density-based decisions and the prior information brought by the LS set. In an experimental study we first assessed the robustness of the proposed semi-supervised method, GWENN-SS, using simulated data. Preliminary results show that GWENN-SS is able to incorporate the LS data in a coherent way (i.e. to consistently grow referenced classes), while allowing the discovery of new classes without any other supervision. The classification of pixels in hyperspectral images is then considered, and GWENN-SS is shown to outperform a recent state-of-the-art method (semi-supervised Kmeans++) in terms of classification accuracy under the provision of erroneous learning samples, and despite the correct number of classes is not given in advance. One key result of this work is that more efficient classification strategies may actually reside in-between supervised methods (i.e. the classes are solely defined by the class membership of the learning samples) and unsupervised clustering (i.e. no learning sample available, and no other *a priori* information). In this perspective, NN-DB methods are probably good potential candidates to open the way for more robust classification approaches which better fit the end-user requirements.

REFERENCES

[1] Chehdi, K. and Cariou, C., "True-false ground truths: what interest?," in [*Proc. SPIE Image and Signal Processing for Remote Sensing XXII*], Bruzzone, L. and Bovolo, F., eds., **10004** (Sept. 2016).

[2] Hsieh, Y.-G., Niu, G., and Sugiyama, M., "Classification from positive, unlabeled and biased negative data.," in [*ICML*], Chaudhuri, K. and Salakhutdinov, R., eds., *Proceedings of Machine Learning Research* **97**, 2820–2829, PMLR (2019).

[3] Blanchard, G., Lee, G., and Scott, C., "Semi-supervised novelty detection.," *Journal of Machine Learning Research* **11**, 2973–3009 (2010).

[4] Scott, C., Blanchard, G., and Handy, G., "Classification with asymmetric label noise: Consistency and maximal denoising.," in [*COLT*], Shalev-Shwartz, S. and Steinwart, I., eds., *JMLR Workshop and Conference Proceedings* **30**, 489–511, JMLR.org (2013).

[5] Duin, R. P. W., Fred, A. L. N., Loog, M., and Pekalska, E., "Mode seeking clustering by knn and mean shift evaluated," in [*SSPR/SPR*], *Lecture Notes in Computer Science* **7626**, 51–59, Springer (2012).

[6] Cariou, C. and Chehdi, K., "Nearest-neighbor density-based clustering methods for large hyperspectral images," in [*Proc. SPIE Image and Signal Processing for Remote Sensing XXIII*], Bruzzone, L. and Bovolo, F., eds., **10427** (Oct. 2017).

[7] Tran, T. N., Wehrens, R., and Buydens, L. M. C., "Knn-kernel density-based clustering for high-dimensional multivariate data," *Computational Statistics & Data Analysis* **51**, 513–525 (Nov. 2006).

[8] Cariou, C. and Chehdi, K., "A new k-nearest neighbor density-based clustering method and its application to hyperspectral images," in [*IEEE Intern. Geoscience and Remote Sensing Symposium*], 6161–6164 (2016).

[9] Cariou, C. and Chehdi, K., "Unsupervised Nearest Neighbors Clustering with Application to Hyperspectral Images," *IEEE J. Selected Topics in Signal Processing* **9**, 1105 – 1116 (Sept. 2015).

[10] Yoder, J. and Priebe, C. E., "Semi-supervised k-means++," *Journal of Statistical Computation and Simulation* **87**(13), 2597–2608 (2017).

[11] Ver Steeg, G., Galstyan, A., Sha, F., and DeDeo, S., "Demystifying information-theoretic clustering," in [*International Conference on Machine Learning*], (2014).

Algorithm 1 GWENN-SS

Require:
$\mathcal{X} = \{\mathbf{x}_m\}, \mathbf{x}_m \in \mathbb{R}^n, m = 1, \ldots, N$; // The set of data objects to classify
K, the number of NNs;
A vector of LS labels $\mathbf{l} = [\ell_1, \ldots, \ell_N]^t$, with $\ell_i \in \mathcal{L} = \{0, \ldots, C\}$; $\ell_i = 0 \Rightarrow \mathbf{x}_i$ is unlabeled $\forall i$;
Ensure: The vector of objects' labels $\mathbf{c} = [c_1, \ldots, c_N]^t$; the set of cluster exemplars $\mathcal{E} = \{\mathbf{e}_c\}_{1 \leq c \leq C}$;
1) Compute $\mathbf{D}$, the $N \times K$ array of distances (in ascending order) between each object and its KNNs.
2) Compute $\mathbf{J} = \{\mathbf{j}_m\}_{m=1,\ldots,N}, \mathbf{j}_m = \left[j_m^1, j_m^2, \ldots, j_m^K\right]$, the $N \times K$ array of indices of each object's KNNs.
3) Compute $\boldsymbol{\rho} = [\rho_i]_{1 \leq i \leq N}$, the vector of local densities around each object, using $(\mathbf{D}, \mathbf{J})$.
4) Compute $\boldsymbol{\rho}' = \texttt{DescendSort}(\boldsymbol{\rho})$, keep KNNs indices $\mathbf{i} = [i_1, i_2, \ldots, i_N]^t : \boldsymbol{\rho}' = \boldsymbol{\rho}(\mathbf{i})$.
5) Set $c_i \leftarrow \ell_i, \forall i : \ell_i \neq 0$
$Q = \{i : \ell_i \neq 0\} \cap \mathbf{j}_{i_1}$ // Intersection between non-zero LSs and the NNs of the denser object
if $Q \neq \emptyset$ **then**
  $c_{i_1} \leftarrow \texttt{wmode}(\mathbf{c}, Q, \rho(Q))$; // $\mathbf{x}_{i_1}$ takes the weighted mode of its LS NNs
else
  $C \leftarrow C + 1$;
  $c_{i_1} \leftarrow C$; // $\mathbf{x}_{i_1}$ takes a new label
end if
$\mathcal{E} = \{i_1\}$; // Exemplar set
$P = \emptyset$; // The set of previously processed objects
for $m = 2 : N$ **do** // Main pass
  if $\ell_{i_m} \neq 0$ **then**
    $c_{i_m} \leftarrow \ell_{i_m}$;
    if $\mathbf{c}_{\mathcal{E}} \cap \ell_{i_m} = \emptyset$ **then**
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{i_m\}$; // Update exemplar set
    end if
  else
    $P \leftarrow P \cup i_{m-1}$;
    $Q = \{P \cup \{i : \ell_i \neq 0\}\} \cap \mathbf{j}_{i_m}$; // Previously processed, or LS NNs of $\mathbf{x}_{i_m}$
    if $Q \neq \emptyset$ **then**
      $c_{i_m} \leftarrow \texttt{wmode}(\mathbf{c}, Q, \rho(Q))$;
    else
      $C \leftarrow C + 1$; // Update number of clusters
      $c_{i_m} \leftarrow C$; // $\mathbf{x}_{i_m}$ takes a new label
      $\mathcal{E} \leftarrow \mathcal{E} \cup \{i_m\}$; // Update exemplar set
    end if
  end if
end for
$\mathbf{c}' \leftarrow \mathbf{c}$;
for $m = 1 : N$ **do** // Second pass
  $Q = \mathbf{j}_m$;
  $c'_m \leftarrow \texttt{wmode}(\mathbf{c}, Q, \rho(Q))$;
end for
$\mathbf{c} \leftarrow \mathbf{c}'$;

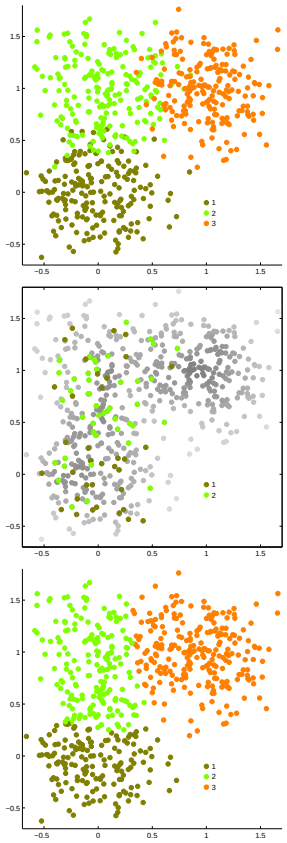

Figure 1. Toy data set classification. (a): data objects with actual GT; (b): data objects with erroneous learning set; (c): classification with GWENN-SS.

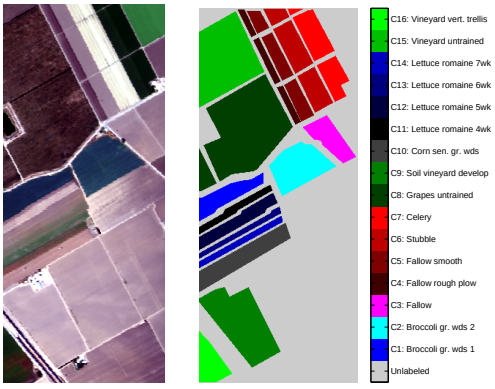

Figure 2. *Salinas* HSI. (a): Color composite (bands 30, 20, 10); (b): Actual GT map.

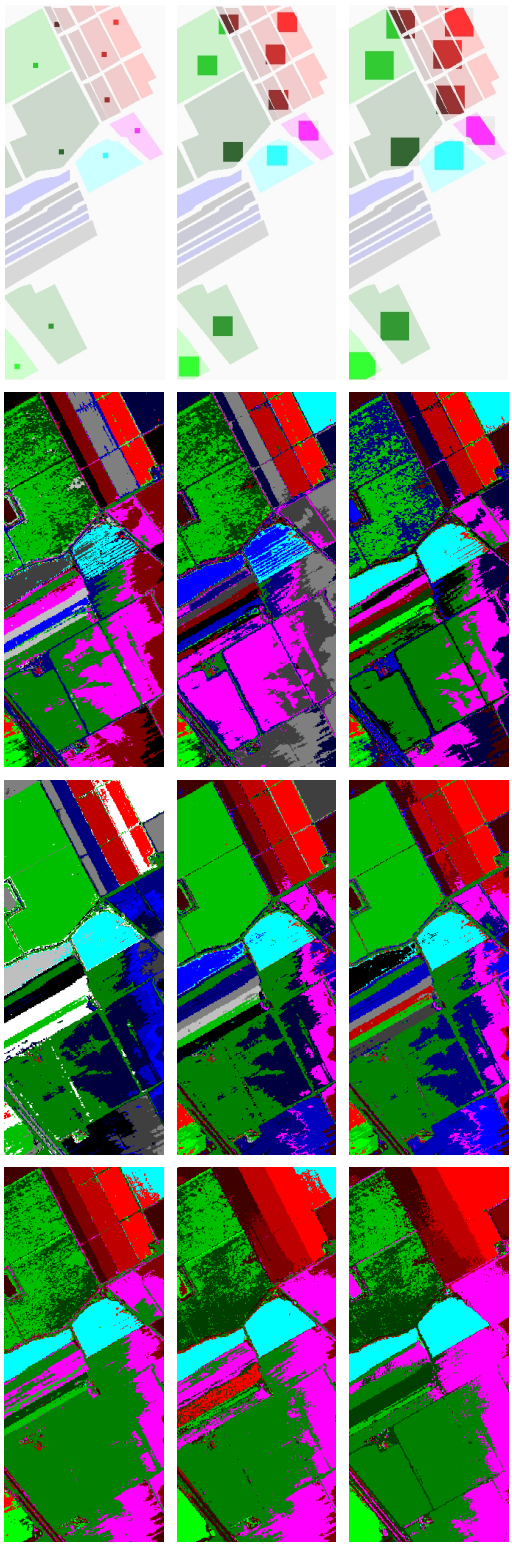

Figure 3. Pixel classification for the *Salinas* HSI. 1st row: footprints of LS patches, from left to right 7×7, 27×27, 39×39 pixels; 2nd row: ss-Kmeans++; 3rd row: GWENN-SS; 4th row: RBF-SVM.

Table 1. Classification results for various patch sizes.

Patch size	LS Error	Method	OA	AA	kappa	CVR	MSE*10^6
7×7	none	ss-Kmeans++	65.15±1.74	66.57±3.13	0.6132±0.0183	1.0436±0.0316	1.807±0.1001
18 classes		GWENN-SS	72.03	71.12	0.6831	0.8768	4.017
11×11	C4: 10.89%	ss-Kmeans++	65.91±2.04	65.55±2.00	0.6205±0.0223	1.0398±0.0508	1.9809±0.1049
17 classes		GWENN-SS	73.85	71.91	0.7033	0.8643	4.1837
15×15	C4: 24.28%	ss-Kmeans++	65.82±2.16	66.55±1.96	0.6200±0.0222	1.0057±0.0432	2.0418±0.1593
16 classes		GWENN-SS	74.54	73.12	0.7111	0.8305	4.2390
19×19	C4: 38.49%	ss-Kmeans++	67.35±0.90	66.54±1.75	0.6354±0.0111	0.9914±0.0545	2.0890±0.1181
16 classes	C5: 1.47%	GWENN-SS	74.66	73.20	0.7124	0.8317	4.2463
23×23	C4: 49.51%	ss-Kmeans++	65.78±1.51	66.59±1.85	0.6201±0.0155	1.0114±0.0417	1.9885±0.1723
17 classes	C5: 8.49%	GWENN-SS	75.26	74.28	0.7195	0.8597	4.1813
27×27	C4: 58.39% C5: 17.07%	ss-Kmeans++	66.64±2.48	67.15±2.55	0.6289±0.0261	1.0409±0.0469	1.9758±0.0736
17 classes	C6: 1.30%	GWENN-SS	75.95	75.25	0.7275	0.8430	4.1389
31×31	C4: 60.45% C5: 25.26%	ss-Kmeans++	67.16±1.57	66.42±1.88	0.6342±0.0166	1.0304±0.0624	2.1046±0.1508
16 classes	C6: 5.12% C7: 0.68%	GWENN-SS	72.48	70.96	0.6879	0.8825	13.412
35×35	C4: 60.13% C5: 31.75%	ss-Kmeans++	66.16±2.07	66.20±2.28	0.6241±0.0213	1.0396±0.0539	2.1458±0.1447
16 classes	C6: 10.81% C7: 2.65%	GWENN-SS	72.69	71.26	0.6902	0.9305	13.423
39×39	C4: 60.72% C5: 37.60%	ss-Kmeans++	66.16±1.51	64.74±1.77	0.6221±0.0168	1.0343±0.0500	2.3188±0.1647
15 classes	C6: 17.11% C7: 5.52%	GWENN-SS	71.24	66.13	0.6736	0.9168	16.846

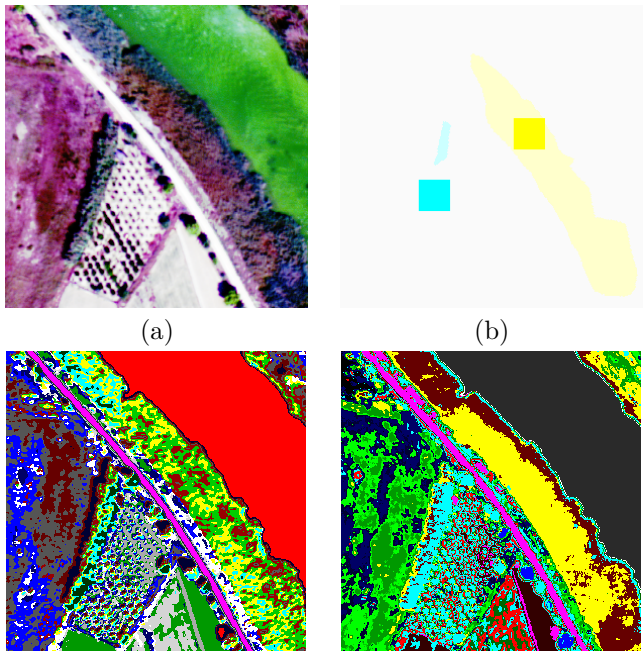

Figure 4. Semi-supervised pixel classification for the *Guardamar* HSI. (a): Color composite (bands 33, 19, 5); (b) Superimposition of square LS patches over the GT; (c) best classification result over 10 runs with ss-Kmeans++; (d): classification result with GWENN-SS.