A Study on the Label Noise Impact on the Hyperspectral Image Classification

Meizhu Li[1], Shaoguang Huang[1], Aleksandra Pižurica[1]
[1]*Department of Telecommunications and Information Processing, TELIN - GAIM, Ghent, Belgium*
{meizhu.li, shaoguang.huang, aleksandra.pizurica}@ugent.be

***Abstract*—Supervised classification methods rely heavily on labeled training data. However, errors in the manually labeled data arise inevitably in practice, especially in applications where data labeling is a complex and expensive process, as is often the case in remote sensing. Erroneous labels affect the learning models, deteriorate the classification performances and hinder thereby subsequent image analysis and scene interpretation. In this paper, we analyze the effect of erroneous labels on spectral signatures of landcover classes in remotely sensed hyperspectral images (HSIs). We analyze also statistical distributions of the principal components of HSIs under label noise in order to interpret the deterioration of the classification performance. We compare the behaviour of different types of classifiers: spectral only and spectral-spatial classifiers based on different learning models. Our analysis reveals which levels of label noise are acceptable for a given tolerance in the classification accuracy and how robust are different learning models in this respect.**

***Index Terms*—Robust classification, hyperspectral images, remote sensing, label noise.**

I. Introduction

Hyperspectral images (HSIs) are being extensively used in numerous applications in various domains, including geosciences [1], [2], agriculture [3], defense and security [4] and environment monitoring [5]. Image classification, which assigns a class label to each image pixel, plays an essential role in the automatic analysis and interpretation of HSIs.

Supervised classification models are typically being developed and tested under the assumption that ideally correct labeled data are available for training. However, in practice imprecise labels are inevitable as labeling is often labor intensive and involves a lot of manual work [6], [7]. The erroneous labels falsely increase the feature variability within class and decrease the discrimination of features across classes. This affects the training and leads to incorrect recognition and classification performance of the new samples. We shall refer to the erroneous data labels as label noise. Classification methods built on diverse principles, ranging from traditional k-nearest neighbours to deep learning are likely to be influenced by the label noise differently. Thus, it is of great interest to investigate which levels of label noise can be tolerated in practice, for a given (user-defined or application-dependent) allowed drop in the classification accuracy and how does this depend on the particular classifier type.

Observing that research on this problem is very scarce, in this paper we study thoroughly the behaviour of several representative supervised classification approaches in the scenarios where different levels of label noise are present in the training data. We assume that the label noise is uniformly distributed in the training data of different classes. We characterise statistically its effect on the spectral signatures of landcover classes and the statistical distributions of features. Our empirical results explain from this perspective clearly the reason for the excellent robustness of Bayesian classifiers (and in particular the simple naive Bayesian classifier) compared to some more complex approaches. At the same time, the empirical results show how erroneous labels affect the model, resulting in a deteriorated classification performance. In addition, the comparison between spectral-based and spectral-spatial based methods demonstrates the benefit of using spatial information to improve the robustness to label noise. We also analyze the classifiers' tolerance to label noise given an acceptable OA degradation.

The rest of the paper is organized as follows. The representative classification methods for HSI that are used for analysis in this paper are briefly introduced in Section II. In Section III, we explain our simulation approach and we analyze the influence of label noise on different aspects. Experimental results and analysis are given in Section IV. We conclude the paper in Section V.

II. Representative Classification Methods for HSI

Here, we review briefly the classifiers that we use for the analysis in this paper. We denote by $\mathbf{x} = (x_1, \cdots, x_m)$ a training sample and $\mathbf{y} = (y_1, \cdots, y_m)$ a test sample, where x_i and y_i are the corresponding i-th features. Let C denote the class variable that is assigned to these samples and that takes values c in a finite set $\mathcal{C}$.

A. Naive Bayes Classifiers (NBCs)

NBCs are simple Bayesian classifiers. For any given feature vector $\mathbf{x}$, an NBC returns the Maximum a Posteriori (MAP) estimate of the class variable C, assuming the conditional independence $P(\mathbf{x}|c) = \prod_{i=1}^{m} P(x_i|c)$. The estimated class is thus:

$$\hat{c} = \arg\max_{c \in \mathcal{C}} P(c|\mathbf{x}) = \arg\max_{c \in \mathcal{C}} P(c) \prod_{i=1}^{m} P(x_i|c). \quad (1)$$

B. K-nearest-neighbor classifier (k-NN)

In k-NN algorithm, the test sample $\mathbf{y}$ is classified by the majority voting of its k nearest neighbors, which are often

measured by the Euclidean distance as follows:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum\nolimits_m (x_m - y_m)^2}. \quad (2)$$

Let $\mathcal{N}_y$ be the set of k nearest neighbors of $\mathbf{y}$ according to Equation (2). The test sample $\mathbf{y}$ is assigned to the class that is most common among $\mathcal{N}_y$.

C. Support vector machine (SVM)

SVM learns a separating hyperplane from a given set of training data with an optimal decision boundary to each class [8], and categorizes new data points by the learned hyperplane. Let $K(\mathbf{x}_i, \mathbf{x}_j)$ be a kernel function which defines an inner product in the feature space. The decision function implemented by SVM can be written as:

$$f(\mathbf{y}) = sgn(\sum\nolimits_{i=1}^{N} c_i \alpha_i \centerdot K(\mathbf{y}, \mathbf{x}_i) + b), \quad (3)$$

where c_i is the corresponding label of sample $\mathbf{x}_i$, b is a real number and the coefficients α_i are obtained by solving the convex Quadratic Programming (QP) problem [9].

D. Sparse Representation Classification (SRC)

SRC identifies the label of test data in two steps: sparse representation and classification. Sparse representation represents a test data $\mathbf{y}$ by a linear combination of a few atoms from a dictionary $\mathbf{D} \in \mathbb{R}^{m \times d}$, which in SRC is constructed specially by the training samples $\{\mathbf{x}_i\}_{i=1}^{d}$. We denote by $\mathbf{D}_i \in \mathbb{R}^{m \times d_i}$ the i-th subdictionary in $\mathbf{D} = [\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_c]$ where each column of $\mathbf{D}_i$ is a training sample of i-th class. The resulting sparse coefficients vector $\boldsymbol{\alpha} \in \mathbb{R}^d$ of $\mathbf{y}$ can be obtained by solving the following optimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \qquad s.t. \qquad \|\boldsymbol{\alpha}\|_0 \leq K, \quad (4)$$

where $\|\boldsymbol{\alpha}\|_0$ denotes the number of non-zero elements in $\boldsymbol{\alpha}$ and K is the sparsity level, i.e., the largest number of atoms in dictionary $\mathbf{D}$ needed to represent any input sample $\mathbf{y}$. The optimization problem in Eq. (6) is typically solved with a greedy algorithm, such as Orthogonal Matching Pursuit (OMP) [10]. Then, the class of the test sample is identified by calculating the class-specific residuals r_i [11]:

$$\begin{aligned} class(\mathbf{y}) &= \arg\min_{i=1,2,...,C} r_i(\mathbf{y}) \\ &= \arg\min_{i=1,2,...,C} \|\mathbf{y} - \mathbf{D}_i\boldsymbol{\alpha}_i\|_2, \end{aligned} \quad (5)$$

where $\boldsymbol{\alpha}_i$ are the sparse coefficients associated with class i.

E. SRC-based classifier with spectral-spatial features

We also consider a representative of SRC-based method where spatial information is included, and in particular we will use in our analysis the method of [12], called SJSRC, which employs super-pixel segmentation and encodes jointly all the pixels within one super-pixel. It assumes that similar pixels in local regions, which are defined by super-pixel segmentation, can be represented by a few common atoms in $\mathbf{D}$. This results in a row sparsity pattern on the coefficients matrix of the pixels within the same super-pixel. Let $\mathbf{X} \in \mathbb{R}^{m \times n}$ represent a super-pixel composed of n pixels in m spectral bands and $\mathbf{A} \in \mathbb{R}^{d \times n}$ the corresponding coefficients matrix. SJSRC solves the following problem

$$\arg\min_{\mathbf{A}} \ \|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2 \ \ s.t. \ \|\mathbf{A}\|_{row,0} \leq K_0, \quad (6)$$

where $\|\mathbf{A}\|_{row,0}$ denotes the number of non-zero rows of $\mathbf{A}$ and K_0 is the row-sparsity level. After finding $\mathbf{A}$, the class for the whole super-pixel $\mathbf{X}$ is decided as:

$$class(\mathbf{X}) = \arg\min_{i=1,2,...,C} \|\mathbf{X} - \mathbf{D}_i\mathbf{A}_i\|_F, \quad (7)$$

where $\mathbf{A}_i$ is the sub-matrix of $\mathbf{A}$ corresponding to class i.

F. Deep learning based spectral-spatial classifier

Deep learning methods have been increasingly used in HSI classification [13]. As a representative of these methods, we select the SSUN algorithm [14], which combines spectral and spatial feature extraction. It incorporates long short-term memory (LSTM) network for band grouping and spectral feature extraction and a convolutional neural network for spatial feature extraction. The loss function is defined as:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}^{joint} + \mathcal{L}^{spectral} + \mathcal{L}^{spatial} \\ &= -\frac{1}{d}\sum\nolimits_{i=1}^{d}[c_i log(\hat{c}^{joint}) + (1-c_i)log(1-\hat{c_i}^{joint})] \\ &- \frac{1}{d}\sum\nolimits_{i=1}^{d}[c_i log(\hat{c}^{spectral}) + (1-c_i)log(1-\hat{c_i}^{spectral})] \\ &- \frac{1}{d}\sum\nolimits_{i=1}^{d}[c_i log(\hat{c}^{spatial}) + (1-c_i)log(1-\hat{c_i}^{spatial})], \end{aligned} \quad (8)$$

where $\mathcal{L}^{joint}$ is the main loss function, $\mathcal{L}^{spectral}$ and $\mathcal{L}^{spatial}$ are two auxiliary loss functions, $\hat{c}^{joint}$, $\hat{c}^{spectral}$ and $\hat{c}^{spatial}$ are the corresponding predicted labels for the ith training sample, c_i is the true label, and d is the size of training set.

III. Model uncertainty caused by label noise

We define the level of label noise ρ as the proportion of training samples that have wrong labels. The erroneous labels are chosen with equal probabilities in $\mathcal{C} \setminus \{c\}$, with $\mathcal{C}$ the set of class values and c the true class.

Fig. 1 illustrates the effect of label noise on spectral signatures in a representative HSI image. Without label noise the spectral signatures of different classes are rather different from each other. In the presence of label noise, they wrongly appear to be more similar to each other. Thus, label noise obviously trends to uniformise all the spectral signatures, which will affect inevitably the classification accuracy.

Fig. 2 shows the effect of label noise on prior probabilities of classes (top left) and on conditional probabilities of the first PC. While the actual prior probabilities of different classes are significantly different from each other, these differences become smaller when label noise increases. The distributions conditioned on the class variable keep a similar shape when increasing ρ from 0 to 0.5, but the peak value decreases and the distribution shape gets more flattened compared to the distributions without label noise.

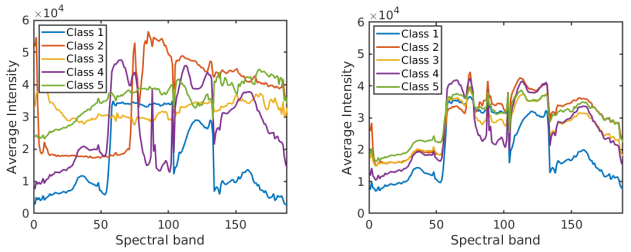

Fig. 1. Average spectral signatures for the *HYDICE Urban* data set with $\rho = 0$ (left) and with $\rho = 0.5$ (right).

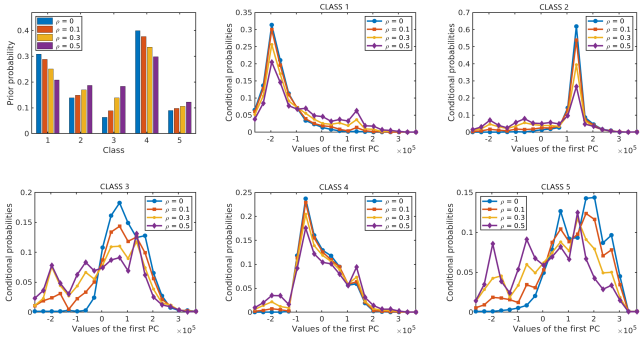

Fig. 2. The prior probabilities of classes (top left) and conditional probabilities of the first PC for different levels of label noise in *HYDICE Urban*.

These results indicate that erroneous labels lead to model uncertainties, which will in their turn affect the classification performance. Baysian models, which are based on conditional probabilities like those in Fig. 2, are likely to be more robust to label noise than some other classifiers that rely more directly on spectral signatures (like those in Fig. 1). Conditional probability distributions do not change significantly until the label noise becomes very large.

In the following section, we will study the performance of the representative classifiers and explore which level of label noise can be tolerated depending on the acceptable drop in the classification accuracy.

IV. Experimental Results and Analysis

A. Datasets

We conduct experiments on two benchmark HSI data sets: *Indian Pines* and *HYDICE Urban* data sets. The *Indian Pines* has a data size of $145 \times 145 \times 220$ and consists of 16 classes. We remove the classes where the number of labelled samples is less than 100 and there are 12 classes in our experiments. We use a part of the *HYDICE Urban* image with size 200×200. It contains 5 classes and 188 bands.

B. Experiments setting

The effect of erroneous labels is studied by evaluating the performance of the six representative classification algorithms described in Section II. Four of these (NBC, k-NN, SVM and SRC) are based on spectral features alone, and the remaining two (SJSRC and SSUN) make use of both spectral and spatial features. In the following experiments, 10 percent of samples are randomly selected for training. The reported results are averaged values over 10 runs. We evaluate the classification performance by overall accuracy (OA), which is the ratio between correctly classified testing samples and the total number of testing samples.

C. Experiments on Indian Pines

Fig. 3 (left) shows the overall accuracy of the six algorithms on *Indian Pines* with ρ ranging from 0 to 0.9, to see the behaviour of selected classifiers and to explore at which level of label noise their performance starts to drop. When there is no label noise ($\rho = 0$), the deep learning method SSUN yields the best OA, while the naive Bayesian classifier (NBC) is inferior to all other methods. This can partly be attributed to the fact that this particular NBC makes use of only spectral features while the two best performing methods (SJSRC and SSUN) incorporate spatial next to spectral features. With the increasing levels of label noise, spectral-based algorithms k-NN, SVM and SRC show similar behaviour, but SRC performs worse than the other two and shows approximately linear decrease. The performance of NBC is the most stable, which can be well understood by analyzing the shape of the involved conditional probabilities (see Fig. 2 and the accompanying discussion in Section III). The performance of NBC drops suddenly when ρ exceeds 0.6. At this point, following further the flattening trend from Fig. 2, the conditional probability distributions become too flattened and the classifier can no longer reasonably operate. The overall accuracy of spectral-spatial methods SJSRC and SSUN deteriorates significantly with the increasing label noise and the deep learning method (SSUN) is especially vulnerable in this respect. The sparse coding method SJSRC achieves thus best performance over the whole range where $\rho > 0.1$.

Fig. 3 (right) shows the maximum level of label noise that a classifier can tolerate given a decreasing rate in the OA compared to the case with no label noise ($\rho = 0$). We analyze the tolerance of the six classification models in the cases with OA decreasing in 5%, 10% and 15% compared to the OA of $\rho = 0$. We assume that the OAs between any two successive ρ (in steps of 0.1) decrease linearly as in Fig. 3. NBC shows the highest tolerance to label noise. E.g., if 5% decrease in OA can be tolerated, NBC allows 30% of erroneous labels. The deep learning approach SSUN exhibits very low tolerance to label noise, although it makes use of both spectral and spatial features. The sparse coding approach based on spectral features alone (SRC) also shows low tolerance to label noise, but its version with spatial information (SJSRC) is much more robust, both compared to the basic SRC and to the deep learning method.

D. Experiments on HYDICE Urban

Fig. 4 shows the performance of the six algorithms on *HYDICE Urban*. Spectral-based algorithms k-NN, SVM and SRC show similar behaviour as in the other data set. NBC performs now better and even outperforms other algorithms for very large ρ. Also, NBC is again the most stable method. Its

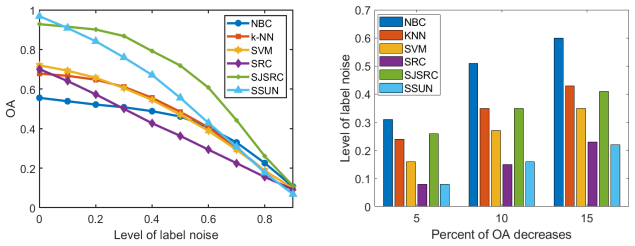

Fig. 3. Influence of label noise on OA (left) and the classifiers' tolerance of label noise at different drops in OA (right). Data set: *Indian Pines*.

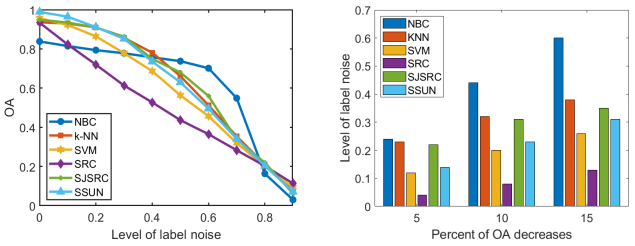

Fig. 4. Influence of label noise on OA (left) and the classifiers' tolerance of label noise at different drops in OA (right). Data set: *HYDICE Urban*.

performance drops suddenly when ρ exceeds 0.6. The spectral-spatial methods SJSRC and SSUN also behave similarly as on the other data set and their overall accuracy deteriorates significantly with the increasing label noise.

Percentages of wrong labels that can be tolerated for a given decrease in OA, shown in the right of Fig. 4, show similar trends as in the first data set. NBC shows again the highest and SRC the lowest tolerance.

V. Conclusion

We analysed the effect of erroneous data labeling on supervised HSI classification from different aspects: the estimated spectral signatures of different classes, the estimated statistical distributions of features and the performance of different types of classification algorithms. The analysis reveals that Bayesian classifiers, even under the simplest naive Bayesian model (NBC) are more robust to label noise than methods based on support vector machines (SVM), sparse coding and deep learning. Deep learning approach exhibited in all our experiments the biggest vulnerability to label noise. This agrees with recent studies that show susceptibility of deep learning to various other perturbations, such as noise in the data and adversarial attacks. We provided explanation for the robustness of the Bayesian approach by analyzing the effect of label noise on the probability distributions of the principal components conditioned on the class variable. These statistical distributions change gently with increasing the label noise (remaining peaked at the same positions and getting gradually flattened). This is the reason why the classification performance of NBC remains very stable until the label noise becomes excessively large. The k-NN method also demonstrated very robust performance, which can be attributed to its majority voting strategy.

Our analysis shows also clearly the importance of using spatial context not only to improve the classification accuracy in ideal settings but also to improve the robustness to label noise. Sparse coding methods that make use of both spectral and spatial information showed excellent performance and can be considered as a good choice of a classifier, which is not only highly accurate but also robust to non-ideal data labeling. It will be also of interest to explore Bayesian classifiers that combine both spectral and spatial features within a unified framework (e.g., as an extension of the NBC that we considered) and to compare those to the sparse coding approach.

Acknowledgment

This work was supported by the China Scholarship Council (CSC), by the Fonds voor Wetenschappelijk Onderzoek (FWO) project under Grant G.OA26.17N and the Flemish Government (AI Research Program).

References

[1] S. J. Buckley, T. H. Kurz, J. A. Howell, and D. Schneider, "Terrestrial lidar and hyperspectral data fusion products for geological outcrop analysis," *Comput. Geosci.*, vol. 54, pp. 249–258, 2013.

[2] S. Huang, H. Zhang, and A. Pižurica, "Semisupervised sparse subspace clustering method with a joint sparsity constraint for hyperspectral remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 989–999, 2019.

[3] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, and J. J. Sousa, "Hyperspectral imaging: A review on UAV-based sensors, data processing and applications for agriculture and forestry," *Remote Sens.*, vol. 9, no. 11, p. 1110, 2017.

[4] M. T. Eismann, A. D. Stocker, and N. M. Nasrabadi, "Automated hyperspectral cueing for civilian search and rescue," *Proc. IEEE*, vol. 97, no. 6, pp. 1031–1055, 2009.

[5] C. Wu, L. Zhang, and B. Du, "Kernel slow feature analysis for scene change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 55, no. 4, pp. 2367–2384, 2017.

[6] J. Jiang, J. Ma, Z. Wang, C. Chen, and X. Liu, "Hyperspectral image classification in the presence of noisy labels," *IEEE Trans. Geosci. Remote Sensing*, vol. 57, no. 2, pp. 851–865, 2018.

[7] M. Li, S. Huang, and A. Pi˘zurica, "Robust dynamic classifier selection for remote sensing image classification," in *Proc. ICSIP*. IEEE, 2019, pp. 101–105.

[8] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, 1999.

[9] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001.

[10] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[11] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, 2008.

[12] S. Huang, H. Zhang, and A. Pižurica, "A robust sparse representation model for hyperspectral image classification," *Sensors*, vol. 17, no. 9, p. 2087, 2017.

[13] X. Li, M. Ding, and A. Pižurica, "Deep feature fusion via two-stream convolutional neural network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 58, no. 4, pp. 2615–2629, 2019.

[14] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 56, no. 10, pp. 5893–5909, 2018.