DOI 10.1109/IGARSS.2017.8127795 – http://ieeexplore.ieee.org/document/8127795/

CORRECT AND STILL WRONG: THE RELATIONSHIP BETWEEN SAMPLING STRATEGIES AND THE ESTIMATION OF THE GENERALIZATION ERROR

R. Hänsch, A. Ley, O. Hellwich

Computer Vision & Remote Sensing, Technische Universität Berlin.

ABSTRACT

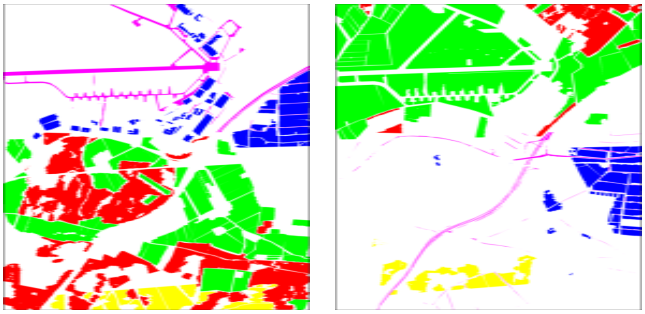
The automatic generation of semantic maps from remotely sensed imagery by supervised classifiers has seen much effort in the last decades. The major focus has been on the improvement of the interplay between feature operators and classifiers, while experimental design and test data generation has been mostly neglected. This paper shows that sampling strategies applied to partition the available reference data into train and test sets have a large influence on the quality and reliability of the estimated generalization error. It illustrates and discusses problems of common choices for sampling schemes, i.e. the violation of the independence assumption and the illusion of the availability of global knowledge in the training data. Furthermore, a novel sampling strategy is proposed which circumvents these problems and achieves a less biased estimate of the classification error.

Index Terms— Supervised classification, Sampling, Error estimation

1 Introduction

Remote Sensing (RS) is one of the major tools for the observation, analysis, and interpretation of natural and manmade processes on the surface of the earth. Corresponding sensors are typically mounted on airplanes or satellites and are able to acquire measurements over relatively large areas in a time and cost efficient way. Especially sensors, that produce images of the earth's surface, have gained importance, such as optical cameras, hyperspectral sensors (HS), or synthetic aperture radar (SAR).

The semantic interpretation of RS images provides the basis to many high-level interpretation processes, such as risk and damage assessment, monitoring of urban growth, land cover mapping, road network extraction, and object detection. The task to assign a class label to each pixel in the image is commonly addressed by supervised learning approaches. The parameters of a sufficiently complex model are adjusted during a training phase with the aim that a given input (i.e. a data sample) produces a given output (i.e. a class label). The supervision is thus provided by the availability of training data, i.e. images alongside with the corresponding reference data (e.g. manually labeled semantic maps). During application the trained method is applied to unlabeled data and estimates the most probable class label. The usability of the final product largely depends on the accuracy and reliability of the produced semantic maps. The training error (i.e. the error on the training data) mainly depends on the model capacity. Given a sufficiently complex method it can often be easily pushed close to 0%. This corresponds most likely to a mere memorizing of the training data, which is why the performance of the classifier on the training data is of little interest. More important is its performance on previously unseen data, i.e. its generalization capabilities. The estimation of the generalization error is achieved by dividing the available labeled data into two (disjoint) sets, where one is used during training and the other during testing.



(a) Train data (b) Test data

Fig. 1: Cluster Sampling applied to Oberpfaffenhofen (OPH) dataset.

The production of reference data nearly always involves manual processing and is thus cost and time consuming. In particular, because many RS data (e.g. SAR or HS images) are more difficult to interpret by human operators than close-range optical images. This leads to the common practice to train and validate methods on one single image only. In this case, the sampling strategy to produce train and test data has a large influence on the estimation accuracy of the generalization error, which might explain the gap between the scientific progress in the automatic interpretation of RS images and their applicability in practice [1].

For an unbiased estimate of the generalization error, train and test data must be independent while test and real world application data should be identically distributed. The common procedure for unstructured data is to randomly sample training and test data while ensuring that both sets are disjoint.

Since early classification approaches of RS images ignore the spatial structure of image data and are based on single pixels alone, this method was thought to be sufficient. However, the independence assumption is violated since both, the value and label of adjacent pixels in an image, are spatially highly correlated. This problem is only increased by the joint usage of spectral-spatial information as exploited by all modern classification pipelines, such as data preprocessing (e.g. noise reduction by local averaging), feature extraction (e.g. textural features), classification (e.g. patch based classification), or post processing (e.g. random field approaches). The implicit or explicit usage of spatial information leads to locally stronger correlated samples. Thus, disjoint train and test sets are not sufficient, but a pixel that contributed in any way to calculate the values of a training sample must not be used for testing and test samples should not be in close proximity to train samples to avoid correlation.

To train and test on the same image does not only increase the risk of violating the independence assumption, but has also implications with respect to the assumption of identically distributed data. Before application, the method will be trained on all available training data and will then be used to predict the labels of new data. Categories like "Forest" or "City" show a strong intra-class variation even if restricted to a certain geographic location avoiding differences caused by different cultures (e.g. cities in Europe vs. North America) or habitats (e.g. European forests vs. tropical rain forests). In many application scenarios it is infeasible to collect training data that covers the whole data variation. Restricting the training data to one single image limits the data distribution to a small fraction of the true distribution and thus underestimates the true variance of the data. A successful classification method thus needs to be able to generalize from a low-variance training set to a high(er) variance test set. If train and test data are sampled globally from the same image, the variance difference between train and application phase is highly underestimated, which leads to a severe underestimation of the generalization error.

The majority of research is focussed on improving the classification performance, while above discussed problems have been mostly neglected. That spatially correlated train and test samples lead to a biased estimate of the generalization error was already noted in [2]. The increase of spatially correlated data by spectral-spatial features and its influence on the quality of the estimate of the generalization error is discussed in [3] for hyperspectral images. A recent extension [4] of this work proposes a sampling scheme that minimizes the spatial overlap between train and test data. However, their method aims to capture the full spectral variation of the image by globally sampling compact regions.

Our work evaluates different sampling strategies as described in Sec. 2 by solving a typical land use classification task for different sensor types including optical, HS, as well as SAR images (Sec. 3). Furthermore, we propose a new sampling strategy (illustrated in Fig. 1) that selects training data for each class locally as well as compactly and thus does not violate the independence assumption. It also simulates a realistic gap of data variation between train and application phase. The conducted experiments confirm previous findings regarding random sampling techniques and show that the proposed sampling scheme leads to less biased error estimates.

Finally it should be noted that above discussed problems can only truly be solved by creating a sufficiently large database along with reference data. The creation and usage of realistic and standardized benchmark data (as e.g. in the DASE [5] project) is a first step in that direction and should be of utmost importance for the RS community.

2 Sampling Strategies

The simplest sampling scheme, that uses the whole dataset for training as well as for testing, is an obvious violation of the independence assumption, leads to an severe underestimation of the generalization error, and is thus (hopefully) never done in practice. The most often used scheme is denoted as (stratified) **Random (R) Sampling**. Randomly selecting (labeled) pixels from the whole image would lead to strongly imbalanced training data including empty classes as the worst case. Thus, the R sampling scheme randomly selects the desired amount of data points within the area of each class. Despite this spatial constraint the training samples are randomly and uniformly (as much as the spatial class distribution allows) distributed over the image.

Despite the widespread use of R sampling it has a major disadvantage: A certain amount of test pixels are very close to pixels selected for training. The spatial context of images renders adjacent pixels as correlated and thus violates the independence assumption. To avoid this problem a minimum distance between train and test samples is sometimes introduced, which needs to be at least as large as the window size used to integrate spatial information. As this leads to a significant decrease of the number of test samples, in particular for small and strongly localized classes, this distance is often selected as too small. Another strategy to ease this problem is denoted as **Patch (P) Sampling**: The image is divided into chessboard like blocks, training samples are selected by applying R sampling only in non-adjacent blocks, test samples are from blocks that haven't been used for training. Both of the above mentioned sampling strategies have a common disadvantage: They underestimate the true variability of the data by sampling data globally distributed over the available image data. This is impossible in real application scenarios where training images can only be acquired over a very restricted portion of the earth but the method is expected to generalize to other areas.

This paper proposes a sampling technique denoted as **Cluster (C) Sampling** that aims to mitigate all of the above mentioned problems by producing a balanced dataset with minimal proximity (and thus minimal correlation) of train

and test samples as well as a non-global distribution of the training data. For each class the spatial coordinates of all samples are clustered into two clusters. Training samples of a class are randomly drawn from one of the clusters, the other cluster is used as test data. If two adjacent clusters (of any classes) contribute to train and test data, a spatial border around the corresponding training samples ensures non-overlapping train and test areas. In this way train and test samples of one class are maximally distinct from each other (ensuring maximal independence between test and train samples) as well as being locally compact simulating the application case where train and application areas are not in proximity.

3 Experiments

3.1 Data and Methods

The above sampling strategies are applied to three different data sets: a) Indian Pines (IP) obtained by the AVIRIS sensor, a commonly used benchmark dataset for hyperspectral image analysis with 145×145px, 220 bands, and 16 classes; b) Oberpfaffenhofen (OPH) obtained by the ESAR sensor (DLR), a fully-Polarimetric Synthetic Aperture Radar (PolSAR) image with 1390×6640px, 3 channels, and 5 classes; and c) Dorsten (D) obtained by an aerial camera[1], an optical color image with 1000×1000px, 3 channels, and 2 classes.

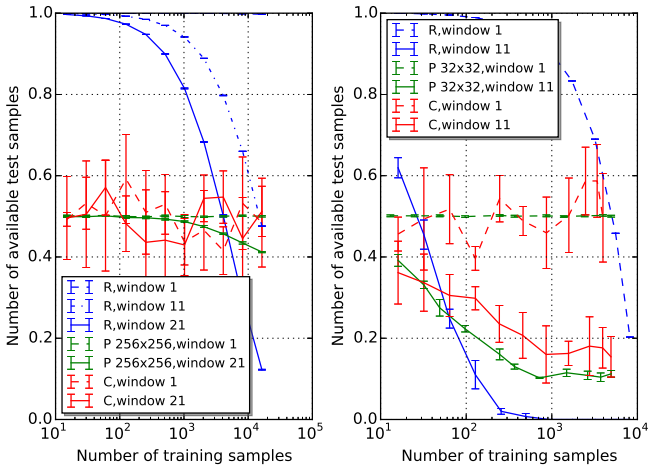
Feature extraction and classification methods are kept simple for an easier interpretation of the obtained results and consist of a) the 10 first spectral principal components as well as local statistics of the average band intensity of the hyperspectral image (denoted as HS); b) standard polarimetric features (log-intensities, span, entropy, anisotropy, α-angle) of the PolSAR image (denoted as SAR); and c) Hue and saturation as well as local statistics of the grayscale image of the optical image (denoted as COL). For the three data based features, a Random Forest classifier is used.

To illustrate the influence of the sampling strategy on the classification rate even if an apparently useless feature is used, we also employ a toy-feature comprising only the spatial position of a sample (LOC) which is classified with simple k-Nearest-Neighbors ($k = 1$) to achieve maximal overfitting.

3.2 Results and Discussion

Fig. 2 shows mean and standard-deviation (based on 5 runs) of the number of independent test samples in the OPH and IP datasets for different sampling strategies, window sizes, and number of training samples. For a small number of training samples there are significantly less independent test samples available for P- and C- as for R sampling. However, as the number of training samples increases, the number of possible test samples decreases exponentially for R sampling, while it stays nearly constant for the schemes P and C.

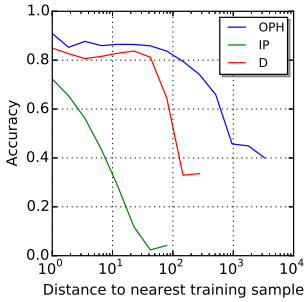
[1]Geobasisdaten: Land NRW, Bonn, 2111/2009



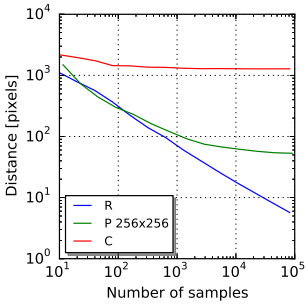
(a) Oberpfaffenhofen. (b) Indian Pines

Fig. 2: Number of independent samples for different sampling strategies and window sizes.

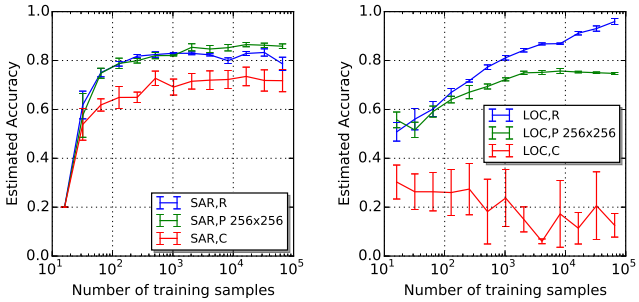
Pixel values do not only correlate due to spatial integration processes within the classification pipeline, but of course also due to the underlying physical process causing the received image signal. The variation of this process is locally much smaller than globally, e.g. the trees within one forest are much more alike among each other as to trees in a different forest further away. This is illustrated in Fig. 3a, which shows the influence of the spatial distance to the nearest training sample on the probability of correct classification (for R sampling). Test samples in close proximity to a training sample are significantly more likely to be classified correctly. The distance for which the probability of correct classification stays high does strongly correlate with the average size of the objects in the image (which is larger for OPH as for IP). Another effect contributing to the decrease of correct classification probability is the fact that class borders tend to get misclassified, and class borders tend to be distinct to training samples. Fig. 3b shows the relationship between the number of used training samples and the average minimal distance between test and train samples. While it decreases almost linearly for R sampling, it saturates in the case of P-, and stays nearly unaffected for C sampling. The left part of Figures 3c-3d illustrates the estimated generalization error for the different sampling methods. R- and P sampling achieve much higher accuracy rates on the sampled test data as C sampling. However, since the average minimal distance to the closest trainings sample is much smaller in these cases (see Fig. 3b) and has a tremendous influence on the classification performance (Fig. 3a), it is likely that this is a biased estimate and the true generalization error is higher. The estimated accuracy also increases for C sampling, but not as strong as for the R- and P scheme. Since the minimal distance between train and test data is only marginally influenced by the number of samples (see Fig. 3b), this increase can be fully attributed to



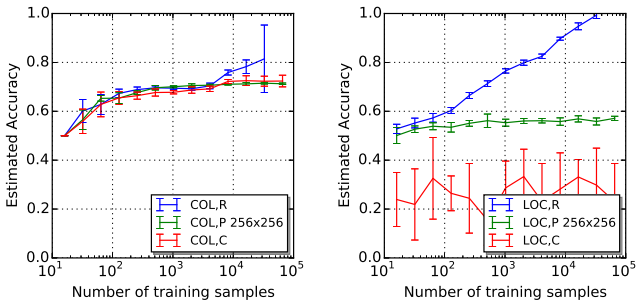
(a) The minimal distance between train and test samples influences the probability of correct classification.



(b) The relationship between the number of samples and minimal distance between test and train samples.



(c) Classification accuracy for dataset OPH based on SAR features (left) and the location feature (right).



(d) Classification accuracy for dataset D based on optical features (left) and the location feature (right).

Fig. 3

the larger amount of training samples. This is supported by the fact that the accuracy based on the location feature (right part of Figures 3c-3d) increases for an increased amount of training samples for R and P sampling, while it stays mostly constant for C sampling.

One could argue that C sampling does not capture the whole variation in the data and thus the accuracy has to be lower. While this is obviously true, it is not a shortcoming but a wanted feature of the proposed approach. In real world scenarios it is very unlikely that samples of all possible object variations of a class are available during training. Relying on the illusion of globally distributed samples during the training of the classifier leads to a severely underestimated generalization error which is avoided by the proposed method.

4 Conclusion

Semantic interpretation of RS images, be it through machine learning approaches, through carefully hand crafted features and rules, or through a combination thereof, aims to produce a black box classifier that can generalize across images to all data of a certain type (e.g. all RS images of a certain sensor, calibration, and use case). Yet, the difficulty of evaluating this key ability is often underestimated. For proper evaluation, test and training data must be independent and the test data must be from the same distribution as the application data later on. Capturing this full distribution can only be ensured by testing on multiple images that at least span the different parameters that might affect the data, such as soil moisture, incidence angle, season, etc.

However, in the absence of a representative, multi-image dataset, the train and test samples have to be drawn from the same image, usually leading to a grave overestimation of the generalization accuracy. We show that the choice of sampling strategy has a big impact on the correlation between training and testing data and that some sampling strategies are less prone to this problem than others. Furthermore, we present a novel sampling strategy, cluster sampling, which seeks to mitigate these problems, at least as far as this can be achieved with a single image.

We believe that large, multi-image datasets will be an important next step in semantic interpretation of RS images, just as it has been, and continues to be, in the field of computer vision. Awareness for the problem of underestimating the generalization error, combined with a more carefully selected sampling strategy, will be necessary to pave the way in this direction.

In future work, we seek to compare cluster sampling with actual multi-image evaluation to better quantify the gap between the estimated and the actual generalization error.

5 References

[1] Deren Li, "Remote sensing in the wenchuan earthquake," *Journal of the American Society for Photogrammetry and Remote Sensing*, vol. 75, no. 5, pp. 506–509, 2010.

[2] M. A. Friedl, C. Woodcock, S. Gopal, D. Muchoney, A. H. Strahler, and C. Barker-Schaaf, "A note on procedures used for accuracy assessment in land cover maps derived from avhrr data," *Int. J. Remote Sens.*, vol. 21, no. 5, pp. 10731077, 2000.

[3] Jun Zhou, Jie Liang, Yuntao Qian, and Yongsheng Gao, "On the sampling strategies for evaluation of joint spectral-spatial information based classifiers," in *Whispers 2015*. IEEE, 2015.

[4] J. Liang, J. Zhou, Y. Qian, L. Wen, X. Bai, and Y. Gao, "On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 862–880, 2017.

[5] "Grss data and algorithm standard evaluation website," `http://dase.ticinumaerospace.com/`.