Active Learning with SVM for Land Cover Classification - What Can Go Wrong?

S. Wuttke [1,2,*], W. Middelmann [1], U. Stilla [2]
sebastian.wuttke@iosb.fraunhofer.de
wolfgang.middelmann@iosb.fraunhofer.de
stilla@tum.de

[1] Fraunhofer IOSB, Gutleuthausstr. 1, 76275 Ettlingen, Germany
[2] Technische Universitaet Muenchen, Arcisstr. 21, 80333 Muenchen, Germany
[*] Corresponding author

Abstract

Training machine learning algorithms for land cover classification is labour intensive. Applying active learning strategies tries to alleviate this, but can lead to unexpected results. We demonstrate what can go wrong when uncertainty sampling with an SVM is applied to real world remote sensing data. Possible causes and solutions are suggested.

1 Introduction

The United Nations define: "Land cover is the observed (bio)physical cover on the earth's surface." [DJ00]. It is important to know which land cover class is found in different areas of the earth to make informed political, economical, and social decisions [And76]. In urban planing for example it is important to differentiate between closed and open soil to predict the effects of rainfall. Achieving this at a large scale and high level of detail is impossible without the help of machine learning algorithms. However, these need laborious training which generates high costs in human annotation time and money. Especially in the field of remote sensing, since acquiring ground truth information often involves expensive ground surveys. Therefore only a limited number of training samples can be produced. How to chose which samples should be labelled out of the large amount of unlabelled data samples is the topic of active learning.

There are many approaches for active learning in remote sensing. In general [TRP$^+$09], [TVC$^+$11] as well as with support vector machines (SVMs) [FM04], [BP09]. This paper investigates if the conventional methods can be easily applied to land cover classification on airborne acquired images. Therefore we use a readily available implementation of an SVM and the intuitive uncertainty sampling and query by committee strategies; and apply them to four publicly available real world datasets: Indian Pines, Pavia Centre, Pavia University, and Vaihingen.

The main contributions of this paper are:

- Apply an SVM with uncertainty sampling and query by committee on five real world datasets
- Present the results and discuss possible causes for the underperforming of active learning
- Suggest future actions to alleviate the observed problems

In: G. Krempl, V. Lemaire, E. Lughofer, and D. Kottke (eds.): Proceedings of the Workshop Active Learning: Applications, Foundations and Emerging Trends, AL@iKNOW 2016, Graz, Austria, 18-OCT-2016, published at http://ceur-ws.org

2 Method

The presented method is deliberately kept simple to reduce possible error sources. The results are still expected to demonstrate the advantages of active learning compared to passive learning.

2.1 Pre-Processing

For noise reduction and lowering the amount of data to be processed, segmentation is applied. Here we use the Multi-Resolution-Segmentation algorithm of the eCognition Software [Tri14] with its default parameters. All pixels of a segment are then combined to an average value, which is the new feature. The reasoning behind this is the smoothness assumption [Sch12]. This assumption states that, because of increasing sensor resolution, the probability that two neighbouring pixels belong to the same class, increases. As a result each training sample represents the average spectrum of the materials present in its segment. This step was not applied to the Indian Pines dataset because of its low resolution. Therefore this dataset has an order of magnitude more samples than the others. No further feature extraction was done to keep possible error sources to a minimum. Classes with less then 15 samples were removed to reduce outlier effects.

2.2 Classification Algorithm

The used classification algorithm is the Mathworks MATLAB [Mat15] implementation of a support vector machine. The pre-set "fine Gaussian SVM" was chosen and all kernel parameters set to their default values. As multi-class method the One-vs-All strategy was selected. The chosen SVM uses Error Correcting Output Codes (ECOC) to transform the multi-class problem into multiple two-class problems resulting in the training of multiple SVMs instead of a single one.

2.3 Selection Strategies

Four different training scenarios were implemented. The first is used as a reference the other three are the comparison between active and passive learning:

All at Once uses all available training samples to get the best possible performance. This value can be seen as a reference to which the other strategies are compared.

Random sampling was implemented by choosing the next training samples at random and represents the passive learning approach.

Uncertainty sampling represents an active learning approach and employs the strategy of the same name [LC94]. The certainty measure used is the estimated posterior probability which is included in the MATLAB default implementation.

Query by committee is a different active learning approach originally introduced in [SOS92]. We used a committee size of 5 and vote entropy as disagreement measure. It selects the next query sample x as

$$\underset{x}{\operatorname{argmax}} - \sum_{y} \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|} \log \frac{\text{vote}_{\mathcal{C}}(y,x)}{|\mathcal{C}|},$$

where $\text{vote}_{\mathcal{C}}(y,x) = \sum_{\theta \in \mathcal{C}} \mathbf{1}_{\{h_\theta(x)=y\}}$ is the number of "votes" that the committee $\mathcal{C}$ assigns to label y for sample x.

The latter three scenarios start their first training iteration with three samples per class. This is a requirement by the SVM implementation to estimate the posterior probability. The batch size was chosen such that after 30 iterations all training samples were exhausted. This resulted in the following batch sizes: Abenberg: 6, Indian Pines: 134, Pavia Centre: 7, Pavia Uni: 4, Vaihingen: 8. For reasons of computational costs the scenarios "uncertainty sampling"and "query by committee" for the Indian Pines dataset were aborted after 4,000 training samples were selected.

2.4 Accuracy Calculation

The first step of evaluating one method is to split the samples randomly into a training (75%) and a test set (25%). During the following training the evaluated selection strategy is allowed access to the feature data of the training set. Only after the samples for the next iteration are selected the label information is provided to the classification algorithm and the next iteration begins. The test set is never connected to the training process and only used for calculating the performance after each iteration. The performance measure used in this paper is the classification accuracy. This is the ratio of correct classified samples to total samples in the test set. Multiple runs of the whole process are done to get statistically robust results.

2.5 Area Under the Curve

To test if the difference between the three iterative scenarios is statistical significant, the performance of each execution was condensed into a single value. To achieve this the area under the learning curve was chosen (learning curve: accuracy vs. number of training samples). The learning curves were matched to span the same range of training samples. Then the trapezoid method was used to calculate the area.

3 Data

This work uses one internal and four publicly available real world datasets. This section gives a short description of them. For a visual impression of the data see Figure 1. It displays the data with overlayed ground truth. An overview of the datasets after the preprocessing step is given in Table 1.

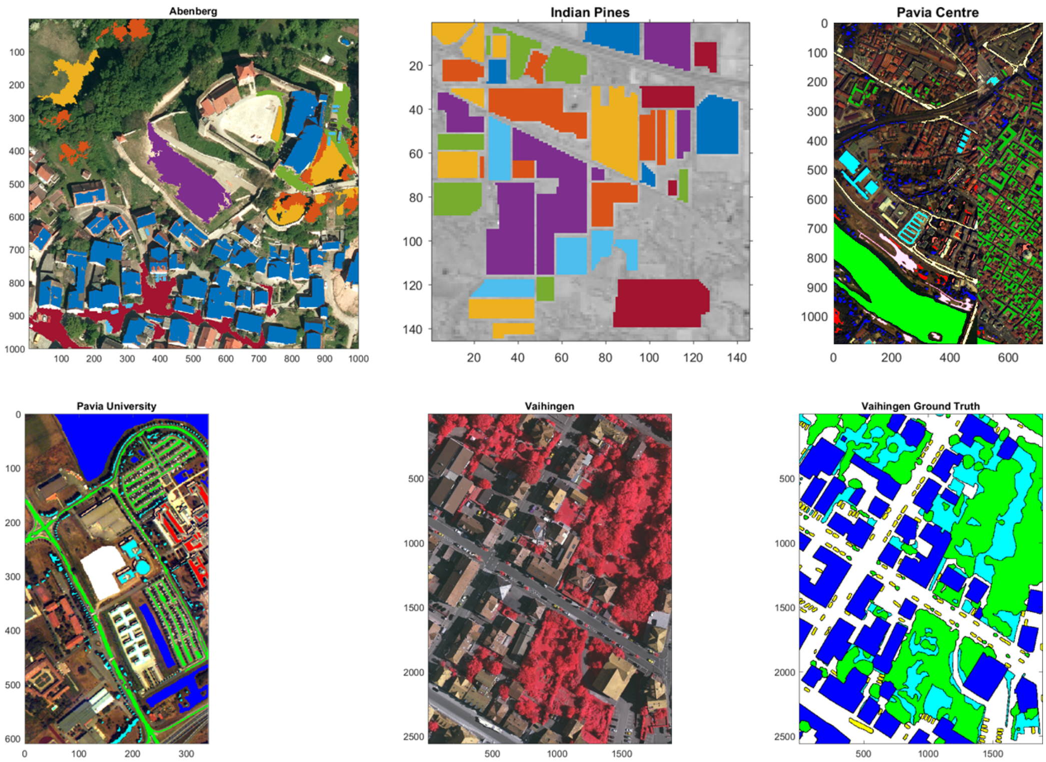
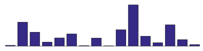
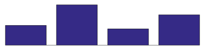
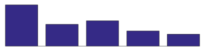


Figure 1: Visual impression of the five datasets overlayed with ground truth information. Vaihingen ground truth is displayed separately for better visualization.

3.1 Abenberg

This dataset is not publicly available. It is an aerial image produced by a survey of Technische Universitaet Muenchen over the Bavarian town of Abenberg in Germany. Each pixel has intensity information for 4 spectral bands: infrared, red, green, and blue. For this work a subset of 1,000 by 1,000 pixel was chosen such that it contains buildings, roads, woodlands, and open soil areas. The ground truth was manually created by this author (8 classes: roof, tree, grass, soil, gravel, car, asphalt, water).

Table 1: Overview of the different datasets after the pre-processing step. All datasets are aerial images of urban and vegetational areas. Each feature is one spectral band.

Dataset	Features	Classes[1]	Samples	SVM Accuracy	Class Distribution [2]
Abenberg	4	8	250	0.86	
Indian Pines	200	16	10,249	0.81	
Pavia (Centre)	102	7	355	0.99	
Pavia (University)	103	8	164	0.86	
Vaihingen (area 30)	3	6	320	0.71	

[1] Contents of original dataset.
[2] Displayed is the final distribution after classes with fewer than 15 samples were removed.

3.2 Indian Pines

This publicly available dataset is an aerial image of the Purdue University Agronomy farm north west of West Lafayette, USA [BBL] and covers different vegetation types. Each pixel is a spectrum containing 200 channels in the 400 to 2,500 nm range of the electromagnetic spectrum. Ground truth is available and contains 16 classes (Alfalfa, Corn-notill, Corn-mintill, Corn, Grass-pasture, Grass-trees, Grass-pasture-mowed, Hay-windrowed, Oats, Soybeans-notill, Soybeans-mintill, Soybeans-clean, Wheat, Woods, Building-Grass-Tree-Drives, Stone-Steel-Towers).

3.3 Pavia Centre & University

These two datasets are also publicly available [Bas11]. They consist of two aerial images of the city Pavia in northern Italy. Contained are urban and vegetation areas with a geometric ground resolution of 1.3 meters. The provided ground truth consists of 9 classes (Water, Trees, Meadows, Self-Blocking Bricks, Bare Soil, Asphalt, Bitumen, Tiles, Shadows), but is mislabelled in the cited datafile (as of submission of this paper). The correct labelling can be found in [Che06], page 494.

3.4 Vaihingen

The Vaihingen dataset stems from the ISPRS Benchmark Test on Urban Object Detection and Reconstruction[1]. It is publicly available and contains multiple aerial images of the town of Vaihingen in Baden-Württemberg, Germany. For each pixel there are intensity values for three channels: infrared, green, and blue. Height information acquired by a LiDAR scanner is also available, but not used in this work. The provided ground truth has six classes (Car, Tree, Low vegetation, Building, Impervious surfaces).

4 Results

Figure 2 shows the learning curves of the methods for the five datasets. They were generated by plotting the classification accuracy (see 2.4) over the number of used training samples. Four of the five datasets show similar performance between active and passive learning with a slight advantage towards passive learning. On the Indian Pines dataset uncertainty sampling greatly underperformes in comparison to random sampling. Each learning curve converges towards the performance of the "All at Once" method which is to be expected, because when all samples are used, the order doesn't matter.

Histograms of the area under the curve for each execution are shown in Figure 3. Those values were used to determine if there are statistically significant differences between the selection strategies. Table 2 lists the p-values of the used statistical test (here: two-sample t-test).

[1] The Vaihingen data set was provided by the German Society for Photogrammetry, Remote Sensing and Geoinformation (DGPF) [Cra10]: `http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html`.

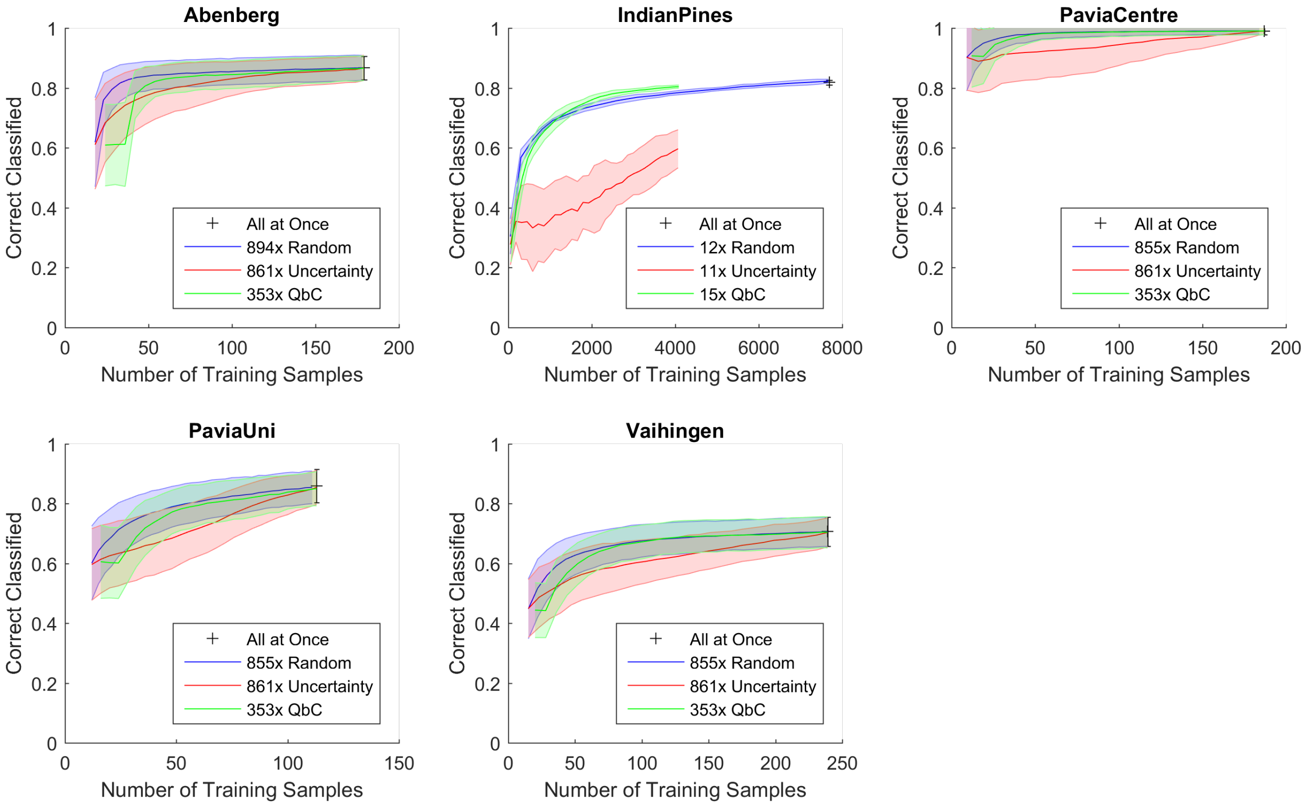


Figure 2: Learning curves for passive ("Random") and active ("Uncertainty", "Query by Committee") learning. "All at Once" is the reference for maximal achievable accuracy. All methods were executed multiple times to reduce the influence of the random splitting into training and test set. The centre line of the graphs is the mean accuracy and the shaded area is the standard deviation.

Table 2: Resulting p-values of the two-sample t-test between the three selection strategies. All combinations except query by committee vs. random sampling on Indian Pines and query by committee vs. uncertainty sampling on Abenberg show p-values nearly equal to zero which indicates a strong statistically significant difference between their performance.

Dataset	Random vs. Uncertainty	Random vs. Query by Committee	Uncertainty vs. Query by Committee
Abenberg	10^{-54}	10^{-32}	0.26
Indian Pines	10^{-9}	0.09	10^{-11}
Pavia (Centre)	10^{-106}	10^{-7}	10^{-39}
Pavia (University)	10^{-109}	10^{-18}	10^{-22}
Vaihingen (area 30)	10^{-104}	10^{-6}	10^{-36}

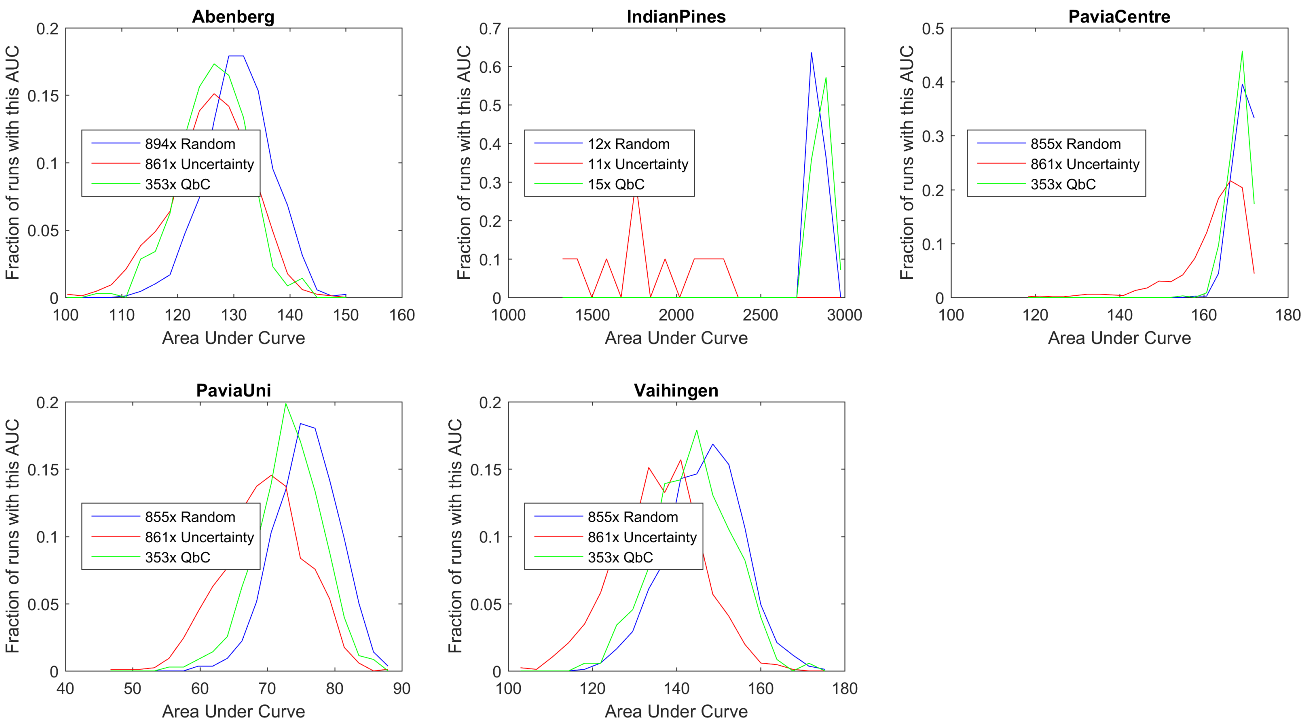


Figure 3: Histograms of the areas under the curve for the different selection strategies and data sets. Each histogram was normalized so that all bins sum to 1.

5 Discussion

Almost all strategies show significant differences for AUC performance. However, the magnitude of the difference measured in correct classified percentage is less then ten percent points which is well within the observed standard deviation. To summarize: the results show that random sampling has a, though small in magnitude, still statistically significant advantage over active learning. This is in stark contrast to most literature. It needs to be determined if this is a problem of the datasets, implementation, or choice of selection strategy. Following we give a list of possible causes, suggest how to test for them, and offer potential solutions.

5.1 Wrong Quality Metric

Cause [RLSKB16] have shown that using a single performance measure can be misleading in an active learning setting.

Test To test this, other measurements such as F_1-Score or area under the receiver operating characteristic curve (AUROC) should be evaluated.

Solution There is no metric that fits every problem. Instead the metric must be chosen to accommodate the domain specific needs. In remote sensing the costs of acquiring more training samples is often higher than the cost of false negatives. However some instances can be weighted opposite for example in the area of Counter-IED (improvised explosive device) detection.

5.2 Uneven Class Distribution

Cause Related to the problem of the wrong quality metric is the problem of uneven class distributions. This is the case if one class is much more common or rarer than others. Random sampling replicates this distribution so that the classification algorithm is trained on the same distribution as it is tested on. In the case of active learning the distribution changes and doesn't match the one from the test data. However it should be noted that it is sometimes argued this bias is the advantage of active learning since it avoids querying redundant samples from overrepresented classes [Mit90].

Test This can be tested by noting which samples are selected during the training process and observing their change of class distribution directly. Also artificially reducing the presence of one class could lead to new insights.

Solution This problem can be alleviated by avoiding use of a classification algorithm that relies on the sample distribution like a Maximum-Likelihood classifier [WMS14]. Instead a non-statistical classifier should be chosen.

5.3 Separability

Cause In remote sensing data the individual pixels often don't contain a single material, but rather a mixture of materials. This leads to overlapping representations in the feature space. If the samples are not separable based on the given features the classifier can't generalize very well.

Test In case of an SVM this could be observed by analysing how many support vectors are used. If the number doesn't increase with more training samples the generalization of the SVM is good. The effect of overlapping classes can be investigated in detail by using specifically generated synthetic datasets or comparing two easily separable classes versus two difficult to separate classes in a two-class setting.

Solution To increase the separability a pre-processing step with feature extraction needs to be introduced. However it remains to be seen if this is an advantage for active learning or just an increase in overall accuracy for both active and passive learning.

5.4 Too Many Samples Per Iteration

Cause The used uncertainty sampling method is based on the estimated posterior probability. To get a good estimate at least three samples per class are needed. Because of this large initial training size the SVM has very good performance from the beginning and shows only very small improvements for the rest of the training so that it is hard to improve by active learning methods. Furthermore for batch sizes with multiple samples the redundant information contained in one batch increases.

Test Observe the classification accuracy of the SVM when initially trained with fewer samples per class. Using smaller batch sizes to reduce the amount of redundant information that is selected in each iteration, should increase the performance.

Solution Use the distance to the hyperplane instead of the estimated posterior probability for the uncertainty sampling method. This alleviates the need for multiple initial training samples per class. The redundant information in one batch can be reduced by adding a second decision criterion like maximising the distance between selected samples in feature space (e.g. density weighted active learning or a diversity criterion [BP09]).

5.5 Using Only Label Information

Cause The presented variant of uncertainty sampling selects samples only based on the state of the learning algorithm. Therefore only information based on the labels of the data is used and information gained from the unlabelled data itself is not utilized.

Solution Applying methods from semi- and unsupervised learning can be beneficial and lead to strategies such as cluster based and hierarchical active learning [LC04], [SOM+12].

References

[And76] J. R. Anderson. *A Land Use and Land Cover Classification System for Use with Remote Sensor Data*. Geological Survey professional paper. U.S. Government Printing Office, 1976. URL: `https://books.google.de/books?id=dE-ToP4UpSIC`.

[Bas11] Basque University. Pavia centre and university: Hyperspectral remote sensing scenes, 2011. URL: `http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes#Pavia_Centre_and_University`.

[BBL] Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 220 band aviris hyperspectral image data set: June 12, 1992 indian pine test site 3. URL: `https://purr.purdue.edu/publications/1947/1`, `doi:10.4231/R7RX991C`.

[BP09] Lorenzo Bruzzone and Claudio Persello. Active learning for classification of remote sensing images. In *International Geoscience and Remote Sensing Symposium*, pages III–693–III–696, Piscataway, NJ, 2009. IEEE. `doi:10.1109/IGARSS.2009.5417857`.

[Che06] C. H. Chen. *Signal and Image Processing for Remote Sensing*. CRC Press, 2006. URL: `https://books.google.de/books?id=9CiW0hgiwKYC`.

[Cra10] Michael Cramer. The dgpf-test on digital airborne camera evaluation overview and test design. *PFG Photogrammetrie, Fernerkundung, Geoinformation*, 2010(2):73–82, 2010. URL: `http://dx.doi.org/10.1127/1432-8364/2010/0041`.

[DJ00] Antonio Di Gregorio and Louisa J. M. Jansen. *Land cover classification systems (LCCS): Classification concepts and user manual*. Food and Agriculture Organization of the United Nations, Rome, 2000.

[FM04] Giles M. Foody and Ajay Mathur. Toward intelligent training of supervised image classifications: directing training data acquisition for svm classification. *Remote Sensing of Environment*, 93(1-2):107–117, 2004. `doi:10.1016/j.rse.2004.06.017`.

[LC94] David D. Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1994.

[LC04] Sanghoon Lee and M. M. Crawford. Hierarchical clustering approach for unsupervised image classification of hyperspectral data. In *International Geoscience and Remote Sensing Symposium: Proceedings*, pages 941–944, 2004. `doi:10.1109/IGARSS.2004.1368563`.

[Mat15] MathWorks. Matlab, 2015.

[Mit90] Tom M. Mitchell. The need for biases in learning generalizations. In ?, editor, *Readings in Machine Learning*. Morgan Kaufmann, 1990.

[RLSKB16] Maria E. Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. Active learning: An empirical study of common baselines. *Data Mining and Knowledge Discovery*, 2016. `doi:10.1007/s10618-016-0469-7`.

[Sch12] Konrad Schindler. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, 50(11):4534–4545, 2012. `doi:10.1109/TGRS.2012.2192741`.

[SOM+12] J. Senthilnath, S. N. Omkar, V. Mani, P. G. Diwakar, and Archana Shenoy B. Hierarchical clustering algorithm for land cover mapping using satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(3):762–768, 2012. `doi:10.1109/JSTARS.2012.2187432`.

[SOS92] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Computational Learning Theory*, pages 287–294. ACM, 1992. `doi:10.1145/130385.130417`.

[Tri14] Trimble Navigation Limited. ecognition developer, 2014.

[TRP+09] Devis Tuia, F. Ratle, F. Pacifici, Mikhail F. Kanevski, and William J. Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009. `doi:10.1109/TGRS.2008.2010404`.

[TVC+11] Devis Tuia, Michele Volpi, Loris Copa, Mikhail F. Kanevski, and Jordi Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011. `doi:10.1109/JSTSP.2011.2139193`.

[WMS14] Sebastian Wuttke, Wolfgang Middelmann, and Uwe Stilla. Bewertung von strategien des aktiven lernens am beispiel der landbedeckungsklassifikation. In *34. Wissenschaftlich-Technische Jahrestagung*, volume 2014, 2014. URL: `http://publica.fraunhofer.de/dokumente/N-283921.html`.