Institut für Geographische Wissenschaften
der
Freien Universität Berlin

Applied One-Class Classification of Remote Sensing Data

Inaugural-Dissertation zur
Erlangung des akademischen Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
des Fachbereichs Geowissenschaften
der Freien Universität Berlin

vorgelegt von

Benjamin Mack

Berlin, 2017

Erstgutachter: Prof. Dr. Björn Waske

Zweitgutachter: Prof. Dr. Sebastian Schmidtlein

Tag der Disputation: 16.06.2017

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende arbeit selbsständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe.

Die aus fremden Quellen direkt oder indirekt übernommenen Inhalte sind als solche kenntlich gemacht. Die Arbeit hat in dieser oder ähnlicher Form bisher noch keiner Prüfungsbehörde vorgelegen.

Benjamin Mack

München, den 10.3.2017

Abstract

Land use and land cover maps are crucial information products required for various purposes in scientific, administrative and commercial domains. Such maps can be efficiently derived by supervised classification of remote sensing data. In many such mapping projects the collection of reference data, which is required for building the classification model, is one of the largest items of expenditure. When only one or a few classes need to be mapped, e.g. an invasive species, one-class classification (OCC) is an attractive pattern recognition approach. It allows for the learning of a classification model from labeled reference data for the class of interest only. There is no need for a representative dataset for the counter-class which consists of all other classes and is therefore often much more difficult to generate. However, in real-world applications it can be very challenging to handle flexible state-of-the-art OCC algorithms.

There is a large body of scientific literature addressing OCC which can be grouped in methodological and applied research. Likewise, the scientists generating this research can be grouped in two communities: the developers and users of OCC algorithms. This thesis reflects on the differences between the prevalent methods, objectives and datasets in the two communities. It identifies and closes knowledge, methodological and technological gaps that are particularly relevant from a user's perspective.

In particular, this thesis provides an in-depth comparative study including three base classifiers and several parameter and model selection approaches. The study is innovative since, in contrast to other comparative studies, it incorporates the potential performance of the base classifiers and analyses the performance loss due to the model selection approaches. It shows that in many cases a high performance loss has to be accepted when relying on fully automatic approaches. Furthermore it shows that the potential performance of MaxEnt, one of the most frequently used algorithms in the user community, is poorer than the biased SVM, a less frequently used algorithm that is perceived as more difficult to tune. The results directly motivate the development of strategies and analytical tools which support the user during model selection and improves the handling of flexible but complicated algorithms such as the biased SVM. Finally, an objective was to study a specific type of OCC problem in which the class of interest is very rare in the area to be mapped and where the number of positive labeled training samples is small. While such data characteristics are frequently met by users they are rather unusual in the benchmark datasets used by developers. A novel OCC approach is developed which is designed for handling such problems.

Zusammenfassung

Landbedeckungs- und Landnutzungskarten sind wesentliche Informationsprodukte für zahlreiche wissenschaftliche, administrative und kommerzielle Aufgaben. Solche Karten können mit Hilfe der überwachten Klassifikation effizient aus Fernerkundungsdaten gewonnen werden. In vielen Kartierungsprojekten ist die Generierung von Referenzdaten für das Training des Klassifikators ein wesentlicher Kostenpunkte. Wenn nur wenige Klassen kartiert werden müssen, z.B. eine invasive Art, stellt die 1-Klassen Klassifikation (1KK) einen attraktiven Ansatz dar. Sie ermöglicht das Lernen eines Klassifikationsmodells aus gelabelten Refernzdaten der Zielklasse. Ein representativer Datensatz für die Gegenklasse ist nicht nötig. Diese Gegenklasse besteht aus allen anderen Klassen des Untersuchungsgebietes und ist daher schwer zu charakterisieren. In realen Anwendungen können flexible 1-Klassen-Klassifikatoren allerdings schwer zu handhaben sein.

Die existierende wissenschaftliche Literatur zu 1KK kann in methodische und angewandte Forschung gruppiert werden. Ebenso lassen sich die Wissenschaftler in zwei Gruppen einteilen, die Entwickler und die Nutzer von 1KK-Algorithmen. In dieser Doktorarbeit werden die Unterschiede zwischen den vorherrschenden Methoden, Zielen und Datensätzen in den beiden Gruppen kritisch reflektiert. Forschungsbedarf, der insbesondere aus der Perspektive des Nutzers von Relevanz ist, wird identifiziert.

Im Rahmen dieser Arbeit wurde eine vergleichende Studie durchgeführt, in welcher drei Basisklassifikatoren, verschiedene Modellselektionsverfahren untersucht wurden. Die Studie ist innovativ, da sie die potentielle Performanz der Basisklassifikatioren und den Performanzverlust aufgrund der Modellselektionsverfahren offenlegt. Das zeigt, dass mit voll-automatischen Modellselektionsverfahren in vielen Fällen ein hoher Performanzverlust in Kauf genommen werden muss. Außerdem wird deutlich, dass die potentielle Performanz von *MaxEnt*, einer der am häufigsten verwendenten Klassifikatoren in der Gruppe der Anwender, schlechter abschneidet als die *biased SVM*, ein weniger häufig verwendeter Algorithmus, der als schwieriger in der Handhabung gilt. Die Ergebnisse motivieren die Entwicklung einer Strategie und analytischer Werkzeuge, welche den Nutzer während der Modellselektion unterstüzen und die Handhabung flexibler und komplizierter Algorithmen wie der *biased SVM* vereinfachen. Schließlich ist ein Ziel dieser Arbeit, die Untersuchung der 1KK besonders schwieriger Datensätze mit ungleicher Klassenverteilung und einer kleinen Menge an gelabelten Trainingsdaten. Während Datensätze mit solchen Charakteristika unter Nutzern häufig auftreten, sind diese eher selten unter den Benchmark-Datensätzen der Entwickler anzutreffen. Ein neuartiger 1KK-Ansatz wurde entwickelt, welcher besonders für die Lösung solcher Probleme geeignet ist.

Contents

List of Figures

List of Tables

Chapter I

Introduction

1 Motivation

Human activities dominate and change the earth's ecosystem in an unprecedented way (Vitousek, 1997). We are faced with a variety of social-ecological challenges from the local to the global scale (Steffen et al., 2004). Social-ecological issues are complex due to increasing interconnectedness and interdependency. Human action is driven by and has consequences on various components of the social-ecological system, including different spatial, temporal and institutional scales (Biermann et al., 2016). Sustainable and equitable management of resources requires understanding and monitoring of social-ecological systems. Amongst other spatial variables land-use and land-cover (LULC) are essential pieces of information for planning and resource management and for modeling a variety of environmental variables. They are an essential input to models in meteorology (Schicker, Arias, and Seibert, 2015), hydrology (Wagner and Waske, 2016), biodiversity (Roy and Tomar, 2000), ecosystem services (Andrew, Wulder, and Nelson, 2014), and many more. They are also used for impact assessments in environmental (Banse et al., 2008) and economic (Hertel, Rose, and Tol, 2009) analyses.

In these areas of application LULC information is required in a variety of spatial and temporal scales and thematic detail. Remote sensing data provides a unique source of data for repeatedly mapping LULC over extensive areas in a cost-efficient way. The image data consists of pixels which capture the radiation of different wavelength reflected or emitted from discrete spatial units on the earth's surface. Categorical LULC classes can be extracted from the data by means of visual interpretation or automated image processing techniques. Accuracy is considered the principle advantage of visual interpretation. However, visual interpretation is expensive since it requires a large amount of working hours of trained experts (Ozdogan, 2015). More cost-efficient is the information extraction from remote sensing data by using classification techniques (Mather and Tso, 2016) based on methods from the fields of statistical learning (Trevor Hastie, 2009), pattern recognition and machine learning (Bishop, 2006).

Traditionally, classification methods are grouped into unsupervised and supervised methods. Unsupervised classification algorithms group pixels based on the similarity of their values into clusters or spectral classes. Clusters do not necessarily correspond to the information classes to be mapped, e.g. LULC classes. Thus, human-guided pre-processing, e.g. stratification and masking, and post-processing, e.g. merging and splitting clusters, are important for the success of cluster algorithms and extracting the desired information classes (Gómez, J. C. White, and Wulder, 2016). Supervised classification algorithms learn from labeled examples (training data) and return the information classes as results. Often, the acquisition of training data needs to be done carefully since the quality

is crucial for the classification outcome (Foody, Pal, et al., 2016). It often requires considerable in-situ campaigns or on-screen collection by trained experts.

In the last decades the cost for deriving remote sensing based information in general and LULC information in particular has decreased tremendously. First, more and more earth observation data is available (Toth and Jóźków, 2016), often free of cost such as Landsat (Woodcock et al., 2008) and Sentinel (Berger et al., 2012) data. Second, the price for hardware required for data processing is relatively low due to technological advancements. Nowadays, it is possible to conduct remote sensing analysis for the local scale on a common computer. Even without purchasing a high performance computing environment large scale analysis is possible using cloud services, such as the Google Earth Engine (Google Earth Engine Team, 2015) or Amazon Elastic Compute Cloud (Amazon, 2016). Third, more and more powerful open source software for remote sensing analysis has become available and is constantly being further developed. This includes user-friendly open source software with graphical user interfaces, such as the EnMap-Box (Linden et al., 2015) or QGIS plugins such as the Orfeo-Toolbox (Inglada and Christophe, 2009) and STEM (Nex, 2015). Additionally, there are a variety of remote sensing specific and general purpose data analysis packages/application programming interfaces (API) in high-level programming languages such as R and Python. Fourth, the cost of training data acquisition can be reduced by different approaches. If available, training data can be extracted automatically from existing but outdated LULC maps and, eventually, auxiliary information (Radoux et al., 2014; Balzter et al., 2015). Furthermore, the required amount of training data can be reduced by using methods from advanced learning paradigms such as semi-supervised learning (Zhu, 2005), active learning (Settles, 2010; Tuia et al., 2009) and one-class classification (Minter, 1975), to name a few.

This thesis addresses one-class classification (OCC) which is an emerging advanced method for LULC classification with remote sensing data. OCC is an option when only one specific class of interest needs to be mapped and when distinguishing between all the other classes is not necessary. In this case all other classes can be conceptualized as one counter-class, i.e. the complement of the class of interest. Such a task might be solved with any binary classifier using training data of both classes. However, the acquisition of representative training samples for the counter-class can be very expensive since it consists of a variety of sub-classes. It is also possible to use a non-representative sample for the counter-class, e.g. by using only informative training cases or a sample of the sub-classes facing the class of interest in the feature space (Foody, Mathur, et al., 2006). But in practice this is difficult to achieve since it is usually unknown which of the cases or sub-classes are most relevant. One-class classifiers are able to solve the binary

classification problem only with labeled training samples of the class of interest, eventually by additionally using unlabeled data which is available at no cost. The eventual high acquisition costs for the counter-class training samples can thus be avoided. It is important to stress that in this thesis the term one-class classification is used in a very strict way and is used only if the full model building process, including parameter and threshold selection, is performed without using negative data. This also excludes the case when negative data is used for selecting parameters and/or a threshold for a one-class classifier, which here would be denoted as supervised classification with a one-class classifier.

Unfortunately, the convenience of OCC with respect to the training data collection may come at the cost of difficulties when building the classification model. The successful implementation of an accurate one-class classifier may require more analytical work than pressing one button or calling one function. However, reading the scientific literature on one-class classification as an optimistic beginner might convey the impression that OCC can be easily solved with many different approaches. There is a large number of methodological scientific papers in which a new fully automatic OCC method is proposed and proved to perform well on a variety of OCC problems. There is also a large number of applied scientific papers in which fully automatic OCC approaches are shown to work on specific applied classification problems. It is more rare to find negative OCC results which is likely due to the positive publication bias. Thus, reading the scientific literature on one-class classification as an enthusiastic and optimistic beginner might convey the impression that OCC is easy. Starting as a new practitioner might then be a frustrating experience when the one-shot methods fail and there is no guidance on how to proceed and little information on what might cause a problem. This might particularly occur to users of these methods who have a strong background in fields such as ecology, biology, geology, etc., but a weaker background in machine learning and pattern recognition.

The overarching goal of this dissertation is to advance the understanding and usability of applied one-class classification in real world remote sensing applications for practitioners who are neither pattern recognition/machine learning experts nor highly skilled programmers. To achieve this, this thesis focuses on three main objectives: First, the performance of the three base classifiers, MaxEnt, biased SVM and one-class SVM is investigated. Each of the base classifiers is tuned with different parameter and threshold selection approaches. Particularly, this study focuses on the comparison between MaxEnt and biased SVM. Biased SVM is more often used in comparative analyses in methodological research studies but perceived as more complicated to use in the applied sciences (Skowronek, Asner, and Feilhauer, 2017). MaxEnt is one of the most frequently used one-class classifier in applied studies since it is easier to use but it is rarely compared to other algorithms. Furthermore, this study reports the performance loss due

to different fully automatic parameter and threshold selection approaches with respect to the potential performance of the base classifiers. Such deeper insights are usually not reported in comparative studies despite being highly informative. When selecting a base classifier this information is helpful, e.g., for a practitioner who feels confident to manually proof and – if necessary – adjust the parameter(s) and threshold, or for a researcher to select a suitable base classifier, e.g. for conducting research on improved model selection approaches.

The second main objective is to develop a strategy and tools to support the practitioner in the evaluation, comparison and improvement of OCC results in the absence of representative and complete test data. Due to the low reliability of fully automatic OCC approaches, such developments are maybe one of the most urgent needs in applied one-class classification. However, so far it is not discussed explicitly in the remote sensing community. As mentioned previously, the methodological research focuses on new methods that are always proved to perform well on a variety of classification problems. However, it is not discussed what to do when such an approach fails on a specific real world task – which might often occur when trying to solve a more challenging OCC problem.

The third main objective is to develop an approach for solving a specific type of OCC problem, i.e. a dataset where the class of interest only occurs rarely in the area to be mapped and the available labeled dataset is small. It is worth noting that such problems are rarely found in the benchmark datasets that are commonly used in methodological research papers. The method has been developed to specifically map raised bogs but can be transferred to map other classes of interest.

The remainder of this thesis is structured as follows: This section provides an illustrative introduction to one-class classification. It can be read as a starting point for (potential) users of one-class classification methods having a less profound background in machine learning, pattern recognition and statistical learning. It focuses on concepts and challenges a user is faced with when using any OCC algorithm to solve a real-world problem. A simple two-dimensional artificial dataset is used such that it is possible to visualize the data, distributions and models in the feature space. The introduced matter is of practical relevance for the user of OCC algorithms. It focuses on aspects which can and should be controlled by the user for successfully solving real-world OCC problems with high-dimensional datasets, complex class distributions and limited training data. Section 2 reveals some practical difficulties users of OCC are confronted with. Based on this background Section 3 describes the prototype characters of OCC algorithm developers and users. It then raises the question about the relevance of the developers'

advancements for improving the users' capabilities to handle OCC algorithms in practice. Some gaps between the two communities that have already been addressed briefly above are carved out in more detail and motivate the contributions of this thesis. A more detailed overview over the main objectives is provided in Section 4. The main contributions are then presented in the Chapters II–,IV and a synthesis of this thesis is provided in Chapter V.

2 Background

2.1 One-Class and Binary Classification

The goal of OCC is to separate a class of interest, or positive class $\mathcal{C}_+$, from all other classes, or negative class $\mathcal{C}_-$. Thus, the goal of OCC and binary classification is the same and it makes sense to first introduce binary classification together with the artificial dataset.

The artificial remote sensing image used for illustration in this chapter has a size of 500×500 pixels and consists of two grayscale images (Figure I.1 B&C). Each pixel (also sample or example) of the image can be represented by a feature vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ which lives in a two-dimensional feature space. Figure I.1 A shows such a feature space with 100 samples selected randomly from the image. The set of samples are colored according to their real class label. In the real world it is often expensive to assign the class label to an adequate amount of samples since it has to be done by a trained expert. Such a reference set with a limited number of labeled samples is enough to train a classification model. With the trained model all image pixels can be classified, or predicted, in order to create a classification image, or more specifically a land use/land cover map (Figure I.1 D).

Most learning algorithms are designed such that the output is the predicted class. However, internally the binary classifier generates a real-valued continuous output which is then converted into a class label by applying a threshold. Figure I.1 H shows such a continuous output together with the threshold, or decision boundary, which separates the feature space in the positive and negative class regions. Specifically, the figure shows the probability of class membership for the positive class derived from Bayes' theorem using the full knowledge of the parameters which generated the data. The Bayes' theorem allows for easily deriving a classifier which maximizes the classification accuracy. This is usually the goal during model building. Understanding Bayes' theorem is useful for two reasons. First, it is intuitive and shows the components required for deriving the most accurate classification model. Second, it can be used to show how good we

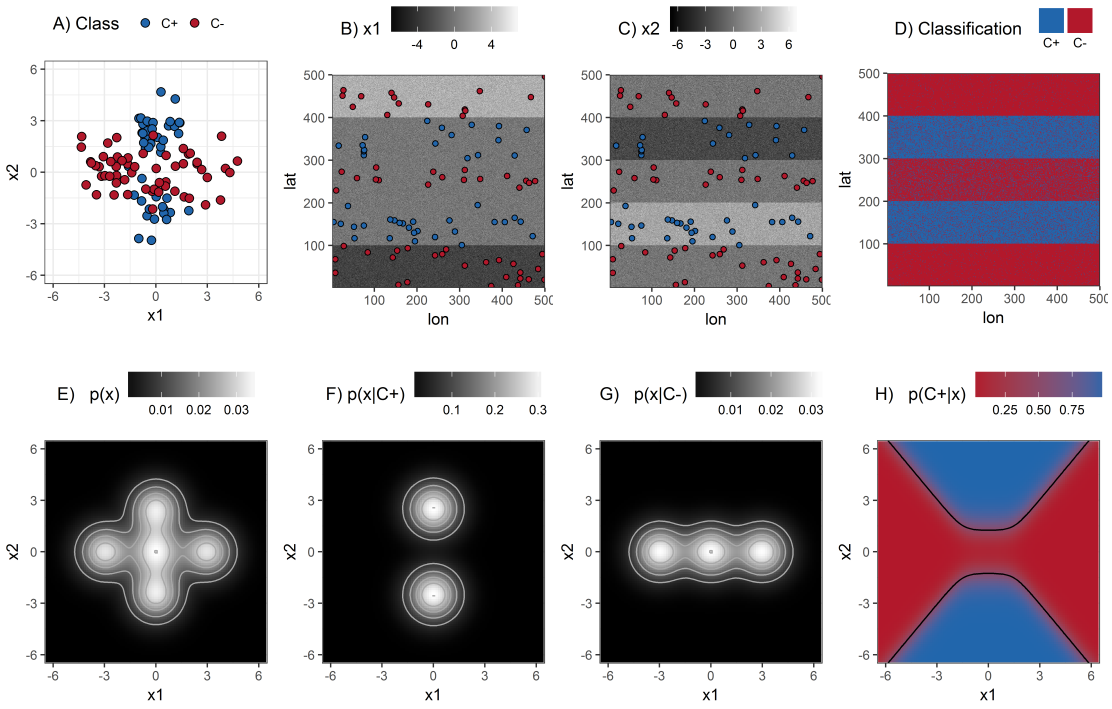

FIGURE I.1: In remote sensing, a labeled dataset (A) usually consists of a subsets of image pixels with known class labels (B&C). Such a dataset can be used to estimate the joint density (E) and the class-conditional density of the positive (F) and negative (G) class which are useful to build a classification model (H). The model can be applied to the whole image data to derive a classification image land cover over the whole imaged area (D).

can get with one-class classification, which will be discussed below. Bayes' theorem is a straightforward formula:

$$p(\mathcal{C}_+|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_+)P(\mathcal{C}_+)}{p(\mathbf{x})},$$

where $p(\mathcal{C}_+|\mathbf{x})$ is the probability of class membership, or posterior probability, for the positive class (Figure I.1 H), $p(\mathbf{x}|\mathcal{C}_+)$ is the conditional density of the positive class (Figure I.1 F), $p(\mathbf{x})$ is the evidence or joint density (Figure I.1 E) and $P(\mathcal{C}_+)$ is the prior probability of the positive class. In order to convert the continuous output $p(\mathcal{C}_+|\mathbf{x})$ to a class estimate a threshold is applied. Since the continuous output is the probability of class membership a threshold of 0.5 is optimal in terms of the classification accuracy.

In the case of binary classification the following equations apply:

$$P(\mathcal{C}_-) = 1 - P(\mathcal{C}_+),$$

$$p(\mathbf{x}|\mathcal{C}_-)P(\mathcal{C}_-) = p(\mathbf{x}) - p(\mathbf{x}|\mathcal{C}_+)P(\mathcal{C}_+),$$

Thus, the conditional density of the negative class (Figure I.1 G) is also hidden in the Bayes' theorem above.

Based on the supervised training set shown in Figure I.1 A it is possible to approximate all these quantities. The class conditional densities can be estimated from the respective reference samples. Since the reference set is a random sample from the whole dataset, the prior probabilities can be estimated from the respective class fractions in the training set.

A great plenty of classification approaches exist to solve a classification problem. Some explicitly model the class-conditional and/or joint densities and the class probabilities and then derive the posterior probabilities based on which classification is performed. Some directly determine the posterior class probabilities based on which classification is performed. Others directly derive the class estimates and do not provide class probabilities (Bishop, 2006). However, most algorithms, including the ones of the last group, allow access to a continuous output which together with a particular default threshold is converted into a class estimates.

Particularly in the case of classification approaches where a model has to be learned from a difficult training dataset, such as in the case of one-class classification it can be crucial to explicitly analyze this continuous output and not rely on any default binary class estimates. This is according to what (Provost, 2000) stresses for the problem of learning from imbalanced data: "The bottom line is that when studying problems with imbalanced data, using the classifiers produced by standard machine learning algorithms without adjusting the output threshold may well be a critical mistake (depending on your research question)."

2.2 P- and PU-Learning

According to (Khan and Madden, 2014) one-class classification has been studied extensively under three broad learning paradigms:

- Learning from positive examples only (P-Learning),
- Learning from positive examples and some amount of poorly sampled negative examples (PN^{poor}-Learning)[1] and
- Learning from positive and unlabeled examples (PU-Learning).

[1] PN^{poor}-learning is not discussed in detail in this thesis. However, it is important to note that the strategies developed for user assisted model and threshold selection Chapter I Section 2.3 can also be used to improve the handling of such learning approaches.

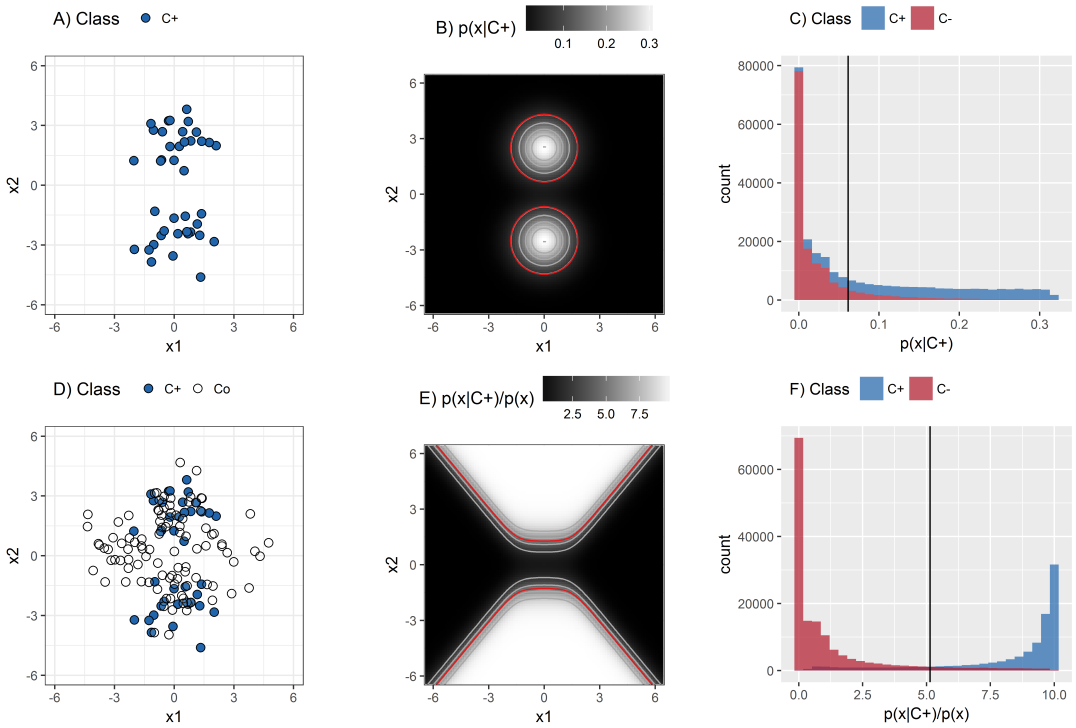

FIGURE I.2: The continuous outputs (B&E) and best achievable decision boundaries (red lines in B&E) that can be approximated with the P- (A) and PU- (D) training dataset differ significantly due to non-uniform class-overlaps. The histograms of the continuous output of all image pixels (C&D) colored by their true class label shows that the separability of the P-based model (C) is lower than the one of the PU-based model (F). Note that the color in B&E corresponds to the x-axes of C&F as well as the red decision boundaries correspond to the vertical black thresholds.

P-learning only allows for estimating the class-conditional probability of the positive class or similar continuous outputs (Tax, 2001). As a consequence, the decision boundary of a P-learner has a structure similar to any density level of $p(\mathbf{x}|\mathcal{C}_+)$. The best achievable decision boundary in terms of the maximum classification accuracy of a P-learner is shown in Figure I.2 B. It results in a producer's and user's accuracy of 81 % and 80 % respectively.

With PU-learning it is possible to derive continuous outputs that implicitly take the density of the negative class into account by exploiting the unlabeled data (Figure I.2 E). It is possible to estimate $\frac{p(\mathbf{x}|\mathcal{C}_+)}{p(\mathbf{x})}$ or similar continuous outputs that are proportional to the posterior probability. Since the posterior probability contains the optimal decision boundary in terms of the minimum classification error it can thus also be derived from the continuous output of a PU-learner. This is possible without additional labeling cost since $p(\mathbf{x})$ can be derived from unlabeled image data. With the best decision boundary based on $\frac{p(\mathbf{x}|\mathcal{C}_+)}{p(\mathbf{x})}$ it is possible to achieve a producer's and user's accuracy of 86 % and 89 % respectively. It has already been shown in the early days of remote sensing that it

is possible to solve the one-class classification problem with Bayes' theorem and PU-data given that the a priori probability is also known (Minter, 1975; Lin and Minter, 1976).

Considering the class-conditional densities of the artificial dataset (Figure I.1 F&G) it is intuitive to understand why the P-learner cannot be optimal. Setting the threshold on $p(\mathbf{x}|\mathcal{C}_+)$ to a high value avoids false positive classifications (positive predictions on pixels belonging to the negative class) in the regions of the feature space where the positive and negative class distributions overlap. However, unnecessary false negative classifications (negative predictions on pixels belonging to the positive class) have to be accepted in regions without class overlap. Instead, a low threshold avoids false negative classifications but at the cost of a high amount of false positive classifications in the overlapping area. More generally, a decision boundary derived from a P-learner cannot be optimal unless the negative class distribution $p(\mathbf{x}|\mathcal{C}_-)$ is uniform on the support of the positive class distribution $p(\mathbf{x}|\mathcal{C}_-)$ or if there is no class overlap (Blanchard, G. Lee, and Scott, 2010).

Given the strong impact of unlabeled data on the theoretical properties of the decision boundary it is important to be aware of the difference between P- and PU-learning when selecting a suitable one-class classifier for a specific application. Particularly, if it is assumed that there is some class overlap between the positive and the negative class it is not recommendable to use a P-learning approach, such as the OCSVM or SVDD[2].

A variety of PU-learners have been proposed in the last decades. In practical applications however, only a few are frequently used, such as MaxEnt (Phillips and Dudík, 2008) and the so called PUL-algorithm (Li, Guo, and Elkan, 2011).

2.3 Model selection

So far, two potential continuous outputs have been analyzed $p(\mathbf{x}|\mathcal{C}_+)$ and $\frac{p(\mathbf{x}|\mathcal{C}_+)}{p(\mathbf{x})}$ which can be approximated by learning from P- or PU-data respectively. However, given the dataset type (P or PU) it has been assumed that the exact data generating distributions can be found. Furthermore, for identifying the classification accuracy, the optimal thresholds have been found by using the one leading to the best achievable accuracies over all possible thresholds.

In real-life OCC applications the challenge is to find a classification model based on P- and PU-data. A variety of methods have been proposed in the scientific literature for solving the problem of OCC. In early approaches it has been assumed that $p(\mathbf{x}|\mathcal{C}_+)$ can be modeled as a multivariate Normal distribution from the positive training samples

[2]This statement is valid when using SVDD as a P-learner. However, it can also be used as a PN^{poor}-learner and it might return similar results as a PU-learner if the negative samples are suitable.

and $p(\mathbf{x})$ as a mixture of multivariate Normal distributions for the unlabeled data (Lin and Minter, 1976). Such assumptions and approaches might be valid for simple classes and datasets, such as the artificial dataset in this chapter or a single-date optical image. However, such assumptions and approaches are not adequate for modern classification problems. The features of modern datasets come from different bands, acquisition dates, sensors, additional features derived from the spatial neighborhood, auxiliary data such as digital elevation models or soil categories, etc. In such high-dimensional feature spaces the structure of the class distributions is unknown and complex and therefore difficult to approximate. This is particularly true with limited (labeled) training data. The difficulty and performance loss with increasing dimensionality is well known as the curse of dimensionality or the Hughes phenomenon (Hughes, 1968).

State-of-the-art machine learning approaches, such as Support Vector Machines or Neural Networks, can be converted to one-class classifiers in different ways (B. Liu et al., 2003; Elkan and Noto, 2008). They are able to learn complex distributions and are less susceptible to the curse of dimensionality. However, since the complexity of such data-fitting models can be arbitrarily high they must be tuned carefully in order to avoid over-fitting on the training data and assure good generalization on new unseen data.

Choosing the model complexity controlling parameters and the threshold converting a continuous output to binary predictions are crucial elements of most, if not all, OCC algorithms. Often, the algorithms are designed in a way that the user is not confronted with model selection, i.e. the algorithm is fed with the training data and returns a model that internally adjusted parameters and uses a default threshold to directly return class estimates when applied to new unseen data. However, particularly in the case of OCC it is challenging to build fully automatic parameter and threshold selection methods which reliably provide good results for a wide range of different OCC problems.

In supervised learning a common approach is to use an independent validation set or to do cross-validation in order to empirically estimate the performance for all candidate models and select the best one. The performance metric is usually an accuracy metric, such as the overall accuracy, F-score, Kappa coefficient, etc. These measures are derived from the number of correctly and erroneously classified samples that can be summarized in a confusion matrix (Table I.1 A). Obviously, the difficulty of model selection in OCC arises from the lack of representative negative examples. In OCC it is not possible to get an estimate of the false positives (FP) and true negatives (TN) since no negative labeled samples are available (Table I.1 B).

Several approaches have been proposed to solve the problem of model selection with P- or PU-data. For example, (Tax and Duin, 2004) proposed a consistency-based model selection for P-classifiers where the complexity of the model is optimized given a user

		PN-Reference			PU-Reference	
		P	N		P	U
Prediction	P	True P	False P	P	True P	Predicted P
	N	False N	True N	N	False N	Predicted N

TABLE I.1: Confusion matrix derived from a PN- (left) and PU-dataset. P: positive, N: negative, U: unlabeled.

defined false negative rate. The algorithm starts with a model having low complexity and thus a low false positive rate. With increasing complexity the false positive rate also increases. The last consistent model is then selected, i.e. the last model in which the empirical false negative rate is still consistent with the pre-defined target false negative rate. This is a reasonable model selection approach, however the necessity of defining the target false negative rate is difficult since the optimal value is usually not known, e.g. if the overall accuracy is to be optimized and not the false negative rate.

Other approaches based on PU-data define PU-performance metrics similar to the ones derived from PN-data. For example (W. S. Lee and B. Liu, 2003) proposed to use the PU-metric similar to the F-score which is the harmonic mean of the precision (user's accuracy of the positive class) and recall (producer's accuracy of the positive class). Their PU-based metric behaves similar as the F-score since it is "large when both precision and recall are large and is small if either precision or recall is small" (W. S. Lee and B. Liu, 2003). A very similar PU-based metric has been proposed by (Li and Guo, 2014).

It is also possible to do unsupervised model selection, i.e. without relying on performance metrics based on P- or PU-data. However, it is then required to make assumptions such as the clustering assumption. The clustering assumption means that the two classes form clusters that are separated by low density regions. Based on a low density criterion it is possible to select the final decision boundary from a variety of candidates (Morsier et al., 2013).

While P- and PU-based metrics are valuable, they are not guaranteed to be optimal for all learning problems. Their performance are more or less dependent on various factors such as the class distribution of the positive class, the number of validation samples used to estimate the metric, the class separability and the distribution of false positive and false negative errors (C. Liu, Newell, and M. White, 2015). Likewise, the assumptions underlying unsupervised model selection approaches are not valid for all learning problems. The difficulty of finding adequate models in a real-world OCC application is further complicated when faced with an imbalanced dataset (Section 2.4) and/or a small positive labeled training set.

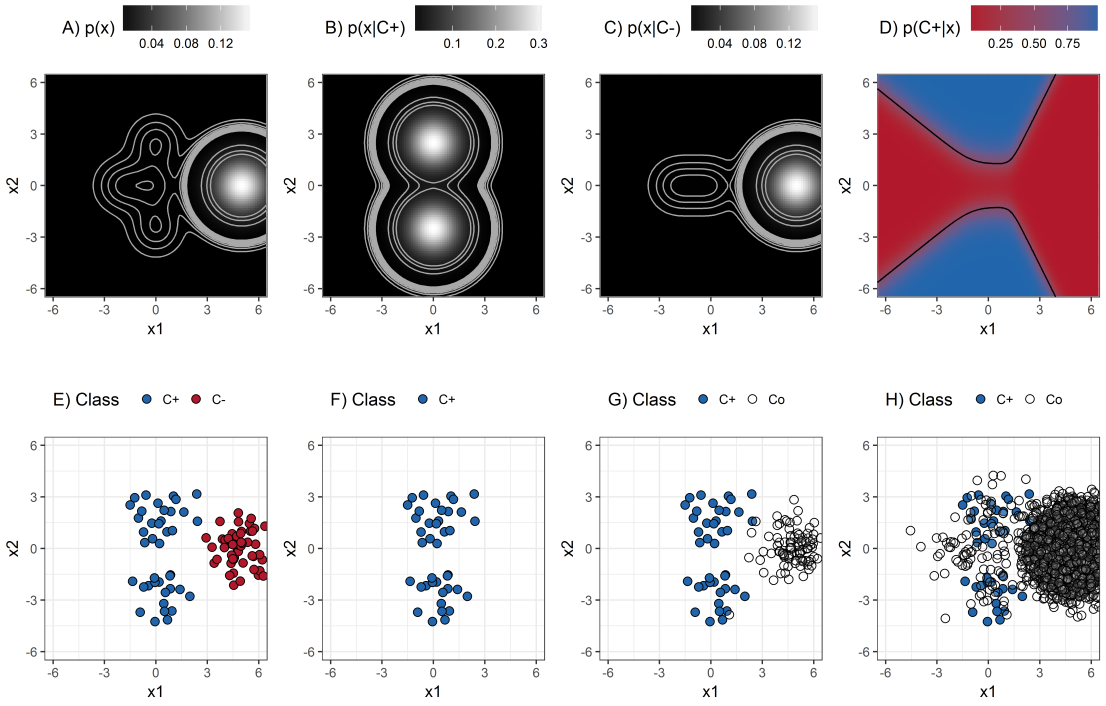

FIGURE I.3: Illustration of imbalanced dataset and small disjuncts. In the imbalanced dataset there are still two positive clusters (positive-conditional density in B) and three negative clusters (negative-conditional density in C). However, the negative data is much more numerous as can be seen by the joint density in A. Furthermore, two of the negative clusters are small disjuncts (C), i.e. they only account for a small fraction of the negative data. In D the posterior probability is shown with the best achievable decision boundary. With a P-dataset (F) or a PU-dataset with too few unlabeled samples it is impossible to get close to the optimal decision boundary (D). However, with sufficient unlabeled samples (H) this is possible. Also a random supervised labeled training set must be large or it does not contain sufficient information (E).

2.4 Imbalanced datasets and small disjuncts

In Section 2.2 the value of unlabeled data for model building has been discussed. Exploiting unlabeled data, e.g. for the estimation of $p(\mathbf{x})$, is often expensive in terms of computational cost. Thus, a random sample of the dataset is usually used instead of exploiting the whole dataset, i.e. all image pixels. Some studies might suggest the usage of training sets with a similar number of positive and unlabeled samples, i.e. balanced PU-training sets (Chen et al., 2016). It is important to understand in which situations it might be a critical mistake to follow such a recommendation and when it is worth paying particular attention to the unlabeled samples used for model building.

The problem of a suitable set of unlabeled samples is closely related to the problems of class imbalance and small disjuncts. The main property of an imbalanced dataset is that the number of examples differs strongly between the classes. In one-class classification

the negative class samples, which is usually an aggregation of many classes, often outnumber the positive class samples. Some of these sub-classes may be represented within small clusters also known as small disjuncts. Selection of unlabeled data becomes particularly challenging when such small disjuncts are similar or even overlapping with the positive class.

Figure I.3 shows an imbalanced dataset. The percentage of positive, and negative, samples is 0.5 %, and 99.5 % respectively. Furthermore, two of the three negative clusters are small disjuncts and only add up to 0.5% (Figure I.3 C). While the large clusters of the negative class is easily separable from the positive class, the small disjuncts are more similar and partially overlapping (Figure I.3 A). The unlabeled data of the PU-training set is again randomly sampled from the whole data, but the subset does not represent the two small disjuncts (Figure I.3 G). From such a data set it is not unlikely to build a model that erroneously classifies the small disjuncts of the negative class as positives. It is important to be aware that an overall accuracy of 99 % could be achieved with a poor unlabeled training set. Still, the classification result might over-predict the positive class with a factor of two which might not be acceptable in a given real-world application.

3 The User's Needs and the Developer's Focus

There are some crucial differences between the developer of OCC algorithms and the user who test them for specific scientific applications or apply them in real-world applications. The characters described here are of course extremes and many remote sensing scientist are somewhere in between. The developers' expertise is in fields such as statistical learning (Trevor Hastie, 2009), pattern recognition and machine learning (Bishop, 2006) and good skills in programming and data analysis are usual. The goal is to design enhanced and fully automatic learning algorithms such that more reliable and more accurate results can be achieved. The algorithms and analysis of the developer are self-made and implemented in programming environments such as MatLab which is widely used in engineering communities. Being aware of the No-Free-Lunch theorem for machine learning (Forster, 1999; Wolpert, 2002) it is acceptable from the developer's view if a novel algorithm fails on some problems as long as it shows significant advantages on others. Simplified, the theorem says that "in the lack of prior knowledge [...], any learning algorithm may fail on some learnable task" (Ben-David, Srebro, and Urner, 2011). Therefore the goal is typically to do better in as many learning tasks as possible. In order to proof the usefulness of a new algorithm it is usually compared to other state-of-the-art algorithms based on a set of learning problems. In the training stage of such comparative exercises negative data is not used but representative and complete

test data is available for a reliable determination of the algorithm performance. As a consequence, for the developer it is straightforward to distinguish a successful from a poor classification result.

Instead, the user has *one specific* mapping problem and seeks to resolve it as good as possible by any means. The problem is that a representative and complete PN-testset might not be available at the moment of model building. Thus, it is not easy to assess the outcome of a classification model. However, since there is often only one or very few classification problems to be resolved at a time the method must not necessarily be fully automatic. If necessary, there is time to invest in manually refined parameter and threshold selection and analyze different classification outcomes in order to find an optimal solution. However, the required tools and skills for such an analysis can be a challenge for the user of an OCC algorithm since the user's background lies in domains such as ecology, biology, geology, environmental planning, resource management, etc. and often in a specific geographical region, and might not be strong in pattern recognition. Furthermore, the users' do not necessarily have the skills in programming and data analysis required to implement one-class classification algorithms and properly analyze their outcomes. As a consequence, only OCC algorithms that are implemented in accessible and familiar software can be used and are usually employed with default settings.

Another crucial difference is the nature of the datasets the two communities are faced with. In many cases the developers benchmark datasets that are relatively small, the number of labeled samples (for the positive class) is large, the classes are relatively well separable, the imbalance ratio between the positive and negative class is relatively small and/or there are no critical small disjuncts of the negative class. A user's real-world dataset might be more challenging and one or more of these dataset attributes might not be favorable (Stenzel et al., 2017). For example, the labeled reference data is often collected by costly fieldwork conducted under limited time and financial budget and thus consists of small sample sizes. As a consequence, an OCC algorithm that is successful on a variety of typical developer datasets does not mean that it also successfully solves a given real-world classification problem.

With these considerations in mind, it is surprising that there is no active scientific discussion about how to support manually-guided model and threshold selection in the absence of representative and complete reference data. Particularly, this is surprising because it is likely that such decision support tools and strategies are more likely to advance the usage of OCC algorithms in the applied sciences and in real-world applications than an additional OCC algorithm that will eventually never be used in these domains. It is worth stressing that the importance and value of novel algorithms shall

by no means be questioned here since they are important to advance the methodological remote sensing science. However, the issues raised here might have more impact on the usage and usability of state-of-the-art OCC algorithms by users in practical and real-world applications.

The MaxEnt algorithm is an interesting case showing in part how separated the two communities work. MaxEnt has been developed for species distribution modeling (SDM), a common task in biogeography and ecology. The problem is similar to OCC for LULC classification with remote sensing data since the unavailability of negative training samples (or absences in SDM terminology) is frequent. In fact, the specific implementation of the maximum entropy principle estimates $\frac{p(\mathbf{x}|\mathcal{C}_+)}{p(\mathbf{x})}$ based on the positive (or presences) and unlabeled (or background) data. Furthermore, the features or predictors for modeling are usually also present in form of raster datasets. The developers of MaxEnt provide a software which directly handles raster data and which has a user-friendly graphical user interface and an extensive tutorial. There is therefore no need to have skills in any programming language. Additional important advantages of MaxEnt from the user perspective is that in the SDM community it has a reputation for working with very small sample sizes and performs well without any user-driven model selection.

These are probably the most important reasons why MaxEnt is the most widely used OCC approach in applied studies using or investigating OCC for LULC classification with remote sensing data. In many of these studies MaxEnt has been shown to perform well by using the default settings. Unfortunately, it is difficult to say how well MaxEnt performs compared to other machine learning OCC approaches since there is a lack of comparative studies. The algorithm is mostly ignored in the developer community as a benchmark approach. This is questionable given its prominence in applied studies. On the other hand, in the applied studies MaxEnt is often the only investigated algorithm. Due to the low likelihood that negative results are published (publication bias) it is however not clear how often MaxEnt performs poorly. Thus, even though MaxEnt sounds like ”free lunch” it is impossible to rate its performance for LULC classification applications based on the scientific literature.

4 Objectives and Organization of this Thesis

In the previous sections the background of one-class classification with a focus on the user's challenges have been presented (Section 2). Furthermore, a gap between the developer's focus and the user's needs has been described (Section 3). These considerations motivate the objectives and research questions of this thesis. The main objective is to advance the usage of OCC for users from geo-scientific domains with limited background

in statistical learning, pattern recognition and machine learning. For this purpose, the following research topics and research questions have been identified.

First, an in-depth comparative study has been conducted including the three base algorithms MaxEnt, biased SVM and ocSVM implemented with various parameter and threshold selection approaches. The comparison is based on a variety of classification problems including different classes (8) and image data sets (3). This study was motivated by two main observations: First, comparing MaxEnt to other state-of-the-art one-class classifiers was overdue considering the fact that it is rarely part of such studies despite its popularity in applied research papers. Second, in most if not all comparative research papers, different base classifiers, e.g. biased SVM or MaxEnt, are implemented with a single parameter and/or threshold selection approach and compared to each other. The informativeness of such comparisons is limited first for users when selecting candidate classifiers for their purpose and second for developers when prioritizing research. In the in-depth comparison the potential performance was also reported, i.e., the best achievable accuracy over all investigated parameter and threshold settings. This information is not usually, if at all, reported in comparative studies. However, this is helpful information about the base classifiers, particularly if it is assumed that the user is eventually able to fix automatically selected parameters and/or thresholds. The following research questions guided the setup of the research presented in Chapter II.

- How do the OCC approaches perform compared with a fully supervised binary SVM?
- Which OCC approach is most accurate? Particularly,
  - which base algorithm has the highest discriminating potential, independent of parameter and threshold selection,
  - how good is the default parameterization of MaxEnt?

The second objective was to develop a user-oriented one-class classification strategy for supporting the user during the one-class classification process in the absence of a complete and representative validation set. This contribution has been motivated by the uncertainty inherent in any fully-automatic OCC approach. There is No-Free-Lunch and particularly not in OCC and related machine learning domains (such as learning from imbalanced data) where the training set is incomplete and/or not representative. As a result, fully-automatic OCC algorithms are more likely to fail when learning a model from the training data compared to classification algorithms based one complete and representative training data. While the developers mainly focus on the difficulty of designing robust and optimal fully-automatic OCC algorithms, the user has to be aware

that any of them might still fail on the specific classification problem to be solved. Thus, the research questions motivating the research presented in Chapter III were:

- How can the outcome of any one-class classification model be visualized and interpreted in order to understand
  - the rough degree of class separability,
  - the suitability of the threshold,
  - the appropriateness of the unlabeled training samples.
- How can an optimal model be identified by the user in an efficient way from a large amount of candidate models.

Finally, the third objective was to design an OCC approach that is particularly suited for datasets which i) have an imbalanced class distribution, i.e. where only a small fraction of the image pixels belong to the positive class, where ii) the positive class potentially overlaps with small disjunct clusters of the negative class (see Section 2.4) and iii) where the number of positive training samples is very low. In the datasets commonly used by developers, such classification problems are rarely addressed. Instead, the class distributions are usually more balanced and the number of labeled training samples is large. The algorithm to be developed was motivated by the following requirements:

- How can the imbalancedness of a dataset be reduced?
- How can the joint parameter and threshold selection problem be solved more robustly, particularly when the set of positive labeled training samples is small?

These topics are treated in the main chapters (II-IV) of this thesis each of which is a self-contained manuscript which has been published in an international peer-reviewed journal:

II Benjamin Mack and Björn Waske (2017). In-depth comparisons of one-class SVM, MaxEnt and biased SVM for one-class classification of remote sensing data. *Remote Sensing Letters*, 8 (3), 290-299.

III Benjamin Mack, Ribana Roscher and Björn Waske (2014). Can I Trust My One-Class Classification? *Remote Sensing*, 6 (9), 8779-8802.

IV Benjamin Mack, Ribana Roscher, Stefanie Stenzel, Hannes Feilhauer, Sebastian Schmidtlein and Björn Waske (2016). Mapping raised bogs with an iterative one-class classification approach, *ISPRS Journal of Photogrammetry and Remote Sensing*, in press.

It is also worth mentioning the open-source R package *oneClass* that has been developed while working on this thesis. The package is a collection of user-friendly functions for training and analyzing one-class classification models and results in the absence of test data (Mack, 2017). It comprises an extensive tutorial which illustrates how to use the package and find a suitable one-class classification model in a user-guided and controlled way (https://github.com/benmack/oneClass/blob/master/notebooks/oneClassIntro.ipynb).

References

Amazon (2016). *Amazon Elastic Compute Cloud*. [ONLINE] Available at: https://aws.amazon.com/documentation/ec2/. [Accessed 09 September 2016].

Andrew, M. E., M. A. Wulder, and T. A. Nelson (2014). "Potential contributions of remote sensing to ecosystem service assessments". In: *Progress in Physical Geography* 38.3, pp. 328–353.

Balzter, H., B. Cole, C. Thiel, and C. Schmullius (2015). "Mapping CORINE Land Cover from Sentinel-1A SAR and SRTM Digital Elevation Model Data using Random Forests". In: *Remote Sensing* 7.11, pp. 14876–14898.

Banse, M., H. van Meijl, A. Tabeau, and G. Woltjer (2008). "Will EU biofuel policies affect global agricultural markets?" In: *European Review of Agricultural Economics* 35.2, pp. 117–141.

Ben-David, S., N. Srebro, and R. Urner (2011). "Universal learning vs. no free lunch results". In: *Philosophy and Machine Learning - Workshop at NIPS*, pp. 1–3.

Berger, M., J. Moreno, J. A. Johannessen, P. F. Levelt, and R. F. Hanssen (2012). "ESA's sentinel missions in support of Earth system science". In: *Remote Sensing of Environment* 120, pp. 84–90.

Biermann, F., X. Bai, N. Bondre, W. Broadgate, C.-T. A. Chen, O. P. Dube, J. W. Erisman, M. Glaser, S. van der Hel, M. C. Lemos, S. Seitzinger, and K. C. Seto (2016). "Down to Earth: Contextualizing the Anthropocene". In: *Global Environmental Change* 39, pp. 341–350.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York Inc. 758 pp.

Blanchard, G., G. Lee, and C. Scott (2010). "Semi-Supervised Novelty Detection". In: *Journal of Machine Learning Research* 11, pp. 2973–3009.

Chen, X., D. Yin, J. Chen, and X. Cao (2016). "Effect of training strategy for positive and unlabelled learning classification: test on Landsat imagery". In: *Remote Sensing Letters* 7.11, pp. 1063–1072.

Elkan, C. and K. Noto (2008). "Learning classifiers from only positive and unlabeled data". In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. Association for Computing Machinery (ACM).

Foody, G. M., A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd (2006). "Training set size requirements for the classification of a specific class". In: *Remote Sensing of Environment* 104.1, pp. 1–14.

Foody, G. M., M. Pal, D. Rocchini, C. Garzon-Lopez, and L. Bastin (2016). "The Sensitivity of Mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data". In: *ISPRS International Journal of Geo-Information* 5.11, p. 199.

Forster, M. R. (1999). *Notice: No-Free-Lunches for Anyone, Bayesians Included*. [ONLINE] Available at: http://www.no-free-lunch.org/Fors99.pdf. [Accessed 09 September 2016].

Gómez, C., J. C. White, and M. A. Wulder (2016). "Optical remotely sensed time series data for land cover classification: A review". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 116, pp. 55–72.

Google Earth Engine Team (2015). *Google Earth Engine: A planetary-scale geo-spatial analysis platform*. [ONLINE] Available at: https://earthengine.google.com. [Accessed 09 September 2016].

Hertel, T. W., S. K. Rose, and R. S. J. Tol, eds. (2009). *Economic Analysis of Land Use in Global Climate Change Policy*. Taylor and Francis. 368 pp.

Hughes, G. (1968). "On the mean accuracy of statistical pattern recognizers". In: *IEEE Transactions on Information Theory* 14.1, pp. 55–63.

Inglada, J. and E. Christophe (2009). "The Orfeo Toolbox remote sensing image processing software". In: *2009 IEEE International Geoscience and Remote Sensing Symposium*. Institute of Electrical and Electronics Engineers (IEEE).

Khan, S. S. and M. G. Madden (2014). "One-class classification: taxonomy of study and review of techniques". In: *The Knowledge Engineering Review* 29.03, pp. 345–374.

Lee, W. S. and B. Liu (2003). "Learning with Positive and Unlabeled Examples Using Weighted Logistic Regression". In: *Proceedings of the Twentieth International Conference on Machine Learning*. Ed. by T. Fawcett and N. Mishra. Washington DC: The AAAI Press, pp. 448–455.

Li, W. and Q. Guo (2014). "A New Accuracy Assessment Method for One-Class Remote Sensing Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 52.8, pp. 4621–4632.

Li, W., Q. Guo, and C. Elkan (2011). "A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data". In: *IEEE Transactions on Geoscience and Remote Sensing* 49.2, pp. 717–725.

Lin, G. C. and T. C. Minter (1976). "Bayes estimation on parameters of the single-class classifier". In: *Proceedings of Symposium on Machine Processing of Remotely Sensed Data*. 3A–22–3A–27. West Lafayette, IN, USA.

Linden, S. van der, A. Rabe, M. Held, B. Jakimow, P. Leitão, A. Okujeni, M. Schwieder, S. Suess, and P. Hostert (2015). "The EnMAP-Box—A Toolbox and Application Programming Interface for EnMAP Data Processing". In: *Remote Sensing* 7.9, pp. 11249–11266.

Liu, B., Y. Dai, X. Li, W. L. Lee, and P. S. Yu (2003). "Building text classifiers using positive and unlabeled examples". In: *Third IEEE International Conference on Data Mining*. IEEE Computer Society, pp. 179–186.

Liu, C., G. Newell, and M. White (2015). "On the selection of thresholds for predicting species occurrence with presence-only data". In: *Ecology and Evolution* 6.1, pp. 337–348.

Mack, B. (2017). *oneClass: one-class classification in the absence of test data*. https://github.com/benmack/oneClass.

Mather, P. and B. Tso (2016). *Classification Methods for Remotely Sensed Data, Second Edition*. CRC Press. 376 pp.

Minter, T. C. (1975). "Single-Class Classification". In: *Proceedings of Symposium on Machine Processing of Remotely Sensed Data*. 2A-12–2A-15. West Lafayette, IN.

Morsier, F. de, D. Tuia, M. Borgeaud, V. Gass, and J.-P. Thiran (2013). "Semi-Supervised Novelty Detection Using SVM Entire Solution Path". In: *IEEE Transactions on Geoscience and Remote Sensing* 51.4, pp. 1939–1950.

Nex, F. (2015). "Land Cover Classification and Monitoring: the STEM Open Source Solution". In: *EuJRS*, p. 811.

Ozdogan, M. (2015). "Image Classification Methods in Land Cover and Land Use". In: *Remotely Sensed Data Characterization, Classification, and Accuracies*. Ed. by P. S. Thenkabail. CRC Press. Chap. 11, pp. 231–258.

Phillips, S. J. and M. Dudík (2008). "Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation". In: *Ecography* 31.2, pp. 161–175.

Provost, F. (2000). "Machine learning from imbalanced data sets 101". In: *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, pp. 1–3.

Radoux, J., C. Lamarche, E. V. Bogaert, S. Bontemps, C. Brockmann, and P. Defourny (2014). "Automated Training Sample Extraction for Global Land Cover Mapping". In: *Remote Sensing* 6.5, pp. 3965–3987.

Roy, P. and S. Tomar (2000). "Biodiversity characterization at landscape level using geospatial modelling technique". In: *Biological Conservation* 95.1, pp. 95–109.

Schicker, I., D. A. Arias, and P. Seibert (2015). "Influences of updated land-use datasets on WRF simulations for two Austrian regions". In: *Meteorol Atmos Phys* 128.3, pp. 279–301.

Settles, B. (2010). *Active learning literature survey.* Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Skowronek, S., G. P. Asner, and H. Feilhauer (2017). "Performance of one-class classifiers for invasive species mapping using airborne imaging spectroscopy". In: *Ecological Informatics* 37, pp. 66–76.

Steffen, W., B. L. Turner, R. J. Wasson, R. A. Sanderson, and P. D. Tyson (2004). *Global Change and the Earth System.* Springer Berlin Heidelberg.

Stenzel, S., F. E. Fassnacht, B. Mack, and S. Schmidtlein (2017). "Identification of high nature value grassland with remote sensing and minimal field data". In: *Ecological Indicators* 74, pp. 28–38.

Tax, D. M. J. (2001). "One-class classification - Concept-learning in the absence of counter-examples". PhD thesis. Delft University of Technology.

Tax, D. M. J. and R. P. Duin (2004). "Support Vector Data Description". In: *Machine Learning* 54.1, pp. 45–66.

Toth, C. and G. Jóźków (2016). "Remote sensing platforms and sensors: A survey". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 115, pp. 22–36.

Trevor Hastie Robert Tibshirani, J. F. (2009). *The Elements of Statistical Learning.* Springer-Verlag New York Inc. 767 pp.

Tuia, D., F. Ratle, F. Pacifici, M. Kanevski, and W. Emery (2009). "Active Learning Methods for Remote Sensing Image Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 47.7, pp. 2218–2232.

Vitousek, P. M. (1997). "Human Domination of Earth's Ecosystems". In: *Science* 277.5325, pp. 494–499.

Wagner, P. D. and B. Waske (2016). "Importance of spatially distributed hydrologic variables for land use change modeling". In: *Environmental Modelling & Software* 83, pp. 245–254.

Wolpert, D. H. (2002). "The Supervised Learning No-Free-Lunch Theorems". In: *Soft Computing and Industry.* Springer Nature, pp. 25–42.

Woodcock, C. E., R. Allen, M. Anderson, A. Belward, R. Bindschadler, W. Cohen, F. Gao, S. N. Goward, D. Helder, E. Helmer, R. Nemani, L. Oreopoulos, J. Schott, P. S. Thenkabail, E. F. Vermote, J. Vogelmann, M. A. Wulder, and R. Wynne (2008). "Free Access to Landsat Imagery". In: *Science* 320.5879, 1011a–1011a.

Zhu, X. (2005). *Semi-supervised learning literature survey.* Computer Sciences Technical Report 1530, University of Wisconsin–Madison.

Chapter II

In-depth comparisons of one-class SVM, MaxEnt and biased SVM for one-class classification of remote sensing data

Remote Sensing Letters, vol. 8 (3), pp. 290-299
Benjamin Mack and Björn Waske
http://dx.doi.org/10.1080/2150704X.2016.1265689

Chapter III

Can I Trust My One-Class Classification?

Remote Sensing, vol. 6 (9), pp.8779-8802
Benjamin Mack, Ribana Roscher and Björn Waske
https://doi.org/10.3390/rs6098779

Abstract

Contrary to binary and multi-class classifiers, the purpose of a one-class classifier for remote sensing applications is to map only one specific land use/land cover class of interest. Training these classifiers exclusively requires reference data for the class of interest, while training data for other classes is not required. Thus, the acquisition of reference data can be significantly reduced. However, one-class classification is fraught with uncertainty and full automatization is difficult, due to the limited reference information that is available for classifier training. Thus, a user-oriented one-class classification strategy is proposed, which is based among others on the visualization and interpretation of the one-class classifier outcomes during the data processing. Careful interpretation of the diagnostic plots fosters the understanding of the classification outcome, e.g., the class separability and suitability of a particular threshold. In the absence of complete and representative validation data, which is the fact in the context of a real one-class classification application, such information is valuable for evaluation and improving the classification. The potential of the proposed strategy is demonstrated by classifying different crop types with hyperspectral data from Hyperion.

1 Introduction

In the last decades, remote sensing sensor technology and data quality (in terms of radiometric, spectral, geometric, and/or temporal resolutions) improved vigorously (Richards, 2005). The availability of such high quality data will probably increase further due to new data policies (European Union, 2013). For example, with the recent and planned Landsat 8 (Roy et al., 2014), EnMAP (Stuffler et al., 2009), and Sentinel (Malenovský et al., 2012) missions the future availability of high-quality data is secured. Moreover, the availability of powerful and free/low cost image processing software for the analysis of remote sensing data, such as R (R Development Core Team, 2013), the EnMAP-Box (Rabe et al., 2014), and the Orfeo Toolbox (Inglada and Christophe, 2009; Christophe and Inglada, 2009), fosters the operational use of earth observation (EO) data. In context of decision-making and surveying compliance of environmental treaties, land use land cover (LULC) classifications of remote sensing data are the most commonly used EO products. However, continuously increasing performance requirements demand for the development of adequate classification techniques. It is likely that future development in LULC classification of remote sensing images will be driven among others by: (i) the demand for more detailed as well as accurate LULC classifications; (ii) the interest

in the distribution of only one or very few classes, e.g., invasive species; and (iii) limited financial resources and time constraints.

Regarding (ii)–(iii), supervised binary or multi-class classifiers such as the maximum likelihood classifier or support vector machine (SVM) are not necessarily appropriate approaches. These classifiers assign each pixel to one of the known classes defined in the training set. Thus, an accurate supervised classifier requires an exhaustive and mutually exclusive training set (Russell G. Congalton, 2008). This means that ideally *all* the classes in the area of interest have to be defined in the training set. If this condition is not fulfilled, *i.e.*, if the training set is incomplete, significant classification errors can occur because all pixels of the unknown classes will be mapped to one of the known classes. Thus, the larger the area of the unknown classes the higher the commission errors. Obviously but notably, these errors do not even appear in an accuracy assessment if the test set does not include the unknown classes (Foody et al., 2006).

Several LULC classification approaches were introduced which can handle incomplete training sets. In the scientific literature these approaches can be found under the terms "classification with reject option" (Dubuisson and Masson, 1993; Muzzolini, Yang, and Pierson, 1998; Fumera, Roli, and Giacinto, 2000), "partially supervised classification" (Jeon and Landgrebe, 1999), and "one-class classification" (OCC) (Minter, 1975; Tax, 2001). While a common supervised classifier maps each pixel to one of the known classes, these classifiers reject the classification of a pixel if it does not sufficiently match one of the known classes. With such algorithms the cost for map production can be significantly reduced, particularly, if the cost for reference data acquisition is high and the user is interested in only one or few classes.

Although the lack of need for training samples from the classes of no interest can be a great facilitation in the training data acquisition step, it turns out to be a burden during the classification. Independent from the approach, an accurate classification requires adequate training data and parameter settings. When using supervised methods, estimation of accuracy measures from complete validation data or the training data itself by cross-validation is commonly used for the selection of an adequate classifier and parameter setting (Trevor Hastie, 2009). In contrast, in the case of OCC the full confusion matrix cannot be derived from the reference data available during the training stage because labeled samples are only available for the class of interest, *i.e.*, the positive class, but not for the other classes, *i.e.*, the negative class (Table III.1). This is a serious problem for the user, because for an accurate classification the user's and producer's accuracies (UA and PA) need to be high.

	y_+	y_-	UA
$\hat{y}_+$	✓	✗	✗
$\hat{y}_-$	✓	✗	✗
PA	✓	✗	✗

TABLE III.1: Confusion matrix with the reference information, $y_{(\cdot)}$ with $(\cdot)$ being the positive $(+)$ or negative $(-)$ class, in the columns and the classified class $\hat{y}$ in the rows. Only y_+ samples are available during OCC, which complicates the selection and training of a suitable model.

Existing one-class classifiers can be separated into several categories, e.g., depending on the type of the training data and the classifier function. Two main categories, P-classifiers and PU-classifiers, are distinguished based on whether the training data set includes positive samples only (P-classifiers) or positive and unlabeled samples (PU-classifiers). PU-classifiers are computationally much more expensive, due to the fact that additional information is extracted extracted from an often very large number of unlabeled samples. However, PU-classifiers can be much more accurate, particularly in the case of significant spectral ambiguities between the positive and the negative class. In such cases a P-classifier cannot perform as accurate as a PU-classifier (Jeon and Landgrebe, 1999; W. Li and Guo, 2010). P-classifiers usually consist of two elements (Tax, 2001): The first element is a similarity measure such as the distance between the positive training samples and the pixel to be classified. The second element is a threshold that is applied on the similarity measure to determine the final class membership. Different approaches to this problem are treated comprehensively in (Tax, 2001).

In the remote sensing community, the one-class SVM (OCSVM) (Schölkopf et al., 2001; P. Li and Xu, 2010; Muñoz-Marí et al., 2010; Sánchez-Azofeifa et al., 2011) and the Support Vector Data Description (SVDD) (Tax, 2001; Foody et al., 2006; Munoz-Mari et al., 2007; Sanchez-Hernandez, Boyd, and Foody, 2007; Bovolo, Camps-Valls, and Bruzzone, 2010) are state-of-the-art P-classifier. As in the case of a supervised SVM two parameters have to be determined, a kernel parameter and a regularization parameter. In practice, the regularization parameter is defined via the omission/false negative rate on (positive only) validation data. This means that the user has to specify the percentage of the positive training data to be rejected by the model. This parameter has to be chosen carefully in order to ensure a good classification result. While values such as 1% or 5% can be suitable when the positive class is well separable (P. Li and Xu, 2010), these parameter settings will result in a high commission/false positive rate when a significant class overlap exists.

The SVDD has been applied in a one-class classifier ensemble where the single classifiers differed in the input features (Munoz-Mari et al., 2007). It has been shown that the ensemble outperformed feature fusion approach, *i.e.*, the classification with the stacked

features, which can possibly attributed to the higher dimensionality. It is worth noting that classifier ensembles have also been applied successfully in the field of species distribution modeling (Drake, 2014; Stohlgren et al., 2010). Furthermore they are a focus of intense research in pattern recognition and machine learning (Désir et al., 2013; Krawczyk, Woźniak, and Cyganek, 2014). These are important developments because multiple classifier systems have been shown to be successful supervised classification of remote sensing data (Briem, Benediktsson, and Sveinsson, 2002; Du et al., 2012; Waske and Braun, 2009) and should be further investigated for one-class classification.

The aforementioned approaches can lead to optimal classification results if (i) there is insignificant class overlap or (ii) if the negative class is uniformly distributed in the part of the feature space where the positive class lives. In the case of significant classes overlap, the second condition is usually not true and any P-classifier will lead to relatively poor results. It is important to note that one-class classifier ensembles based on P-classifiers are also not suitable for such classification problems.

PU-classifiers try to overcome the problems by exploiting unlabeled data. Usually, it is not feasible to use all the unlabeled pixels of an image and a random selected subset is used. This should be as small as possible (such that the algorithm is computational efficient) but large enough to contain the relevant information. The adequate number of samples depends on the classification problem, particularly, the complexity of the optimal decision boundary and the occurrence probabilities of the positive and the overlapping negative classes. There are also support vector machine approaches which allow the exploitation of unlabeled samples such as the semi-supervised OCSVM (S^2OCSVM) (Muñoz-Marí et al., 2010) and the biased SVM (BSVM, see also Section 3.2) (B. Liu et al., 2003; Muñoz-Marí et al., 2010).

Another possibility, which is also addressed in this paper, is the usage of Bayes' rule for the one-class classification with positive and unlabeled data (Minter, 1975):

$$p(y_+|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|y_+)P(y_+)}{p(\mathbf{x}_i)} \tag{III.1}$$

where $p(y_+|\mathbf{x}_i)$ is the a posteriori probability of the positive class given a pixel $\mathbf{x}_i$, $p(\mathbf{x}_i|y_+)$ the conditional probability of the positive class, $P(y_+)$ the a priori probability of the positive class, and $p(\mathbf{x}_i)$ the unconditional probability (see also Section 2 for more details).

There are different ways of solving the OCC problem based on Bayes' rule. A probabilistic discriminative approach can be implemented to solve the classification problem

(Elkan and Noto, 2008; W. Li, Guo, and Elkan, 2011). Also different generative approaches have been proposed. They differ in the way that the probability density functions, $p(\mathbf{x}_i|y_+)$ and $p(\mathbf{x}_i)$, and the a priori probability are estimated (Lin and Minter, 1976; Jeon and Landgrebe, 1990; Fernández-Prieto, 2002; Mantero, Moser, and S. Serpico, 2005; Fernandez-Prieto and Marconcini, 2011; Marconcini, Fernandez-Prieto, and Buchholz, 2014). The Maxent approach (Elith et al., 2010; Phillips, Anderson, and Schapire, 2006; Phillips and Dudík, 2008), developed in the field of species distribution modeling, also estimates the density ratio $\frac{p(\mathbf{x}_i|y_+)}{p(\mathbf{x}_i)}$. In contrast to the aforementioned approaches, Maxent has been used more frequently for one-class land cover classification in applied studies (Amici, 2011; Evangelista et al., 2009; W. Li and Guo, 2010; Morán-Ordóñez et al., 2012; Ortiz, Breidenbach, and Kändler, 2013).

It is important to note that the probabilistic discriminative and the generative approaches return a posteriori probabilities which offers the user an intuitive possibility to solve the thresholding problem. Thresholding these probabilities at 0.5 corresponds to the maximum a posteriori rule and leads to an optimal classification result in terms of the minimum error rate. This requires accurate estimates of the terms of Bayes' rule (see Equation (III.1)). In (Amici, 2011; Evangelista et al., 2009; W. Li and Guo, 2010; Morán-Ordóñez et al., 2012; Ortiz, Breidenbach, and Kändler, 2013) $P(y_+)$ has not been available neither has it been estimated from the data. Thus, the derived continuous output is not an a posteriori probability with an intuitive meaning. Instead, the user has to find a different way to solve the threshold problem, *i.e.*, the conversion of the continuous Maxent output, often called suitabilities, to a binary classification output. In (Ortiz, Breidenbach, and Kändler, 2013) the value of 0.5 is applied on the logistic Maxent output, even though the authors are aware of the fact that they are not dealing with "true" probabilities. In (W. Li and Guo, 2010) the 5 % omission rate estimated on a (positive only) validation set is used. A detailed theoretical and empirical comparison of threshold approaches used in the field of species distribution modeling is provided in (C. Liu, White, and Newell, 2013). However, it is important to underline that all these techniques do not generally provide the optimal classification result. The usefulness of such thresholds in terms of the minimum error rate depends on the specific classification problem. Therefore, the result must be evaluated by the user based on the limited reference data.

Besides the threshold selection, the solution by Bayes' rule seems further interesting. The derived posteriori probabilities can be used as input in advanced spatial smoothing techniques (Roscher, Waske, and Forstner, 2012; Moser and S. B. Serpico, 2013) or for combining OCC outputs of several classes in one map (Munoz-Mari et al., 2007; Guo et al., 2012). With a posteriori probabilities it is also straightforward to consider different mis-classification costs for false positive and false negative classifications (Bruzzone,

2000). Finally, error probabilities (both the probability of omission and commission, *i.e.*, false negative and false positive) can be estimated by integrating over the probability densities (Minter, 1975). Unfortunately, it is very challenging to accurately estimate the required quantities $p(\mathbf{x}_i|y_+)$, $p(\mathbf{x}_i)$, and $P(y_+)$, particularly, if the positive labeled training data is scarse and the dimensionality of the image is large. This is well known under the terms Hughes phenomena or course of dimensionality (Hughes, 1968; Shahshahani and Landgrebe, 1994).

In this paper we propose a user-oriented strategy to support the user in handling one-class classifiers for a particular classification problem. Thus, the complicated handling of one-class classifiers can be overcome, the application of a state-of-the-art methodology is advanced and the increased requirements for effective analysis of remote sensing imagery may be easier fulfilled. In a nutshell, the user first performs any OCC, e.g., the BSVM as in this study. To evaluate the classification result, the continuous output of the one-class classifier is further analyzed, e.g., the distance to the separating hyperplane in case of the BSVM. The distributions of the classifier output and the positive and unlabeled training data are visualized. If interpreted carefully, this diagnostic plot is very informative and helps to understand (i) the discriminative power, or separability, of the classifier; and (ii) the suitability of a given threshold applied to convert the continuous output to class estimates. In addition, a posteriori probabilities are estimated by solving Bayes' rule in the one-dimensional classifier output space. Therefore, the thresholding problem is objectively solved.

It is important to note that no new one-class classification algorithm is introduced. However, to the best of our knowledge the combination of a modern or state-of-the-art one-class classifiers, e.g., the BSVM, with subsequent analysis of the one-dimensional one-class classifier output space with Bayes' rule has not been proposed before. Note, that one of the most important advantage of this strategy is the ease of visualization in one-dimensional feature space. In the absence of representative validation data, as in the case of OCC applications, this is useful to evaluate the quality of particular model outcomes, e.g., the continuous output, threshold, or a posteriori probabilitites. The presented strategy should support the user in better understanding a particular one-class classification outcome in the absence of complete reference data. This is an important component for successfully apply one-class classification in real-world applications and has not been addressed in previous studies. These studies propose particular solutions for the problems of model and threshold selection and prove the functioning of the selected approach by means of representative test sets. Testing new solutions by means of a representative test set is an essential element in a scientific research papers. However, it does not guaranteed that they perform well when applied on different data sets in new real-world applications. This is the case in general but particularly critical in one-class

classification where reference data is extremely limited. We want to stress that the results of this strategy do not necessarily provide improved accuracies compared to other well working approaches. However, they provide the user with easy to interpret information in order to assess the quality of a selected threshold (see the synthetic example in Section 2), estimated a posteriori probabilities (see the example in Section 5.1), and/or the selected one-class classification model (parameterization) (see the experiment in Section 5.2). Therefore, poor solutions might be detected even without a representative reference set which we believe to be of utmost value in real world applications.

This paper is structured as follows: In the the next section we present the proposed strategy and illustrate it with a two-dimensional synthetic data set . The specific methods for the implementation of the strategy are described in Section 3. The data and experiments conducted to demonstrate the strategy are presented in Section 4. The results are presented and discussed in Section 5. The conclusions close the paper in Section 7.

2 A User-Oriented Strategy for One-Class Classification

In this section the strategy is illustrated by means of a two dimensional synthetic data set. In two dimensions we can visualization the data and BSVM model (see Figure III.1 a.1, a.2) and should facilitate the understanding of this section and the strategy. In practice, visualization of the original input feature space is usually not possible because high-dimensional data sets are used for classification. Therefore, we recommend the analysis of the classification problem in the one-dimensional output space of a given one-class classifier, which can be visualized in practice (see Figure III.1 b.2).

The synthetic example is generated from three normal distributions (Figure III.1 a.1). Two of the normal distributions belong to negative class, one with an a priori probability of 0.96 and the other one with 0.02. The third normal distribution is assumed to generate the data of the positive class with an a priori probability of 0.02. The positive class overlaps with the "small negative distribution" but is well separable from the "large negative distribution" (see Figure III.1 a.1). Additionally a test set $\mathcal{X}^{\text{te}}$ consisting of 100,000 samples is generated from the three normal distributions according to their a priori probabilities. First, a one-class classifier $g(\cdot)$ is trained with the training data $\mathbf{x}_i \in \mathcal{X}^{\text{tr,PU}}$ with $i \in \{1, \ldots, I\}$, consisting of 10 positive and 250 unlabeled samples (Figure III.1 a.1). In this paper the BSVM is used to implement $g(\cdot)$ (see also Section 3.2). The example training set, the mixture of normal distributions $p(\mathbf{x}_i)$, the output of the trained classifier $z_i = g(\mathbf{x}_i)$, and the default and optimal decision boundaries are shown in Figure III.1 a.1, a.2. The default decision boundary of the BSVM, *i.e.*, the separating hyperplane or $z = 0$, and the optimal decision boundary are also shown in

Figure III.1 a.2). The latter is derived by applying the maximum a posteriori rule on the a posteriori probabilities derived by the known data generating distributions and a priori probabilites. For explanation and visualisation purposes the synthetic dataset is chosen to be two-dimensional and the optimal decision boundary is known because we defined the data generating processes. However, for the proposed user-oriented strategy for handling OCC, higher dimensional data can be used and the optimal trained classifier model need not to be known.

Second, the continuous classifier outputs are predicted using the trained classifier with $\mathcal{Z} = g(\mathcal{X})$. Figure III.1 b.2 shows the so-called diagnostic plot. It comprises the density histogram of the predictions $\mathcal{Z}$ shown in gray and the distributions of the training data $\mathcal{Z}^{\mathrm{PU}} = \mathcal{Z}^{\mathrm{P}} \cup \mathcal{Z}^{\mathrm{U}}$, where $\mathcal{Z}^{\mathrm{P}}$ (shown as blue boxplot) and $\mathcal{Z}^{\mathrm{U}}$ (shown as grey boxplot) are the cross-validated predictions of the training set $\mathcal{X}^{\mathrm{tr,P}}$ and $\mathcal{X}^{\mathrm{tr,U}}$. In order to ensure that the predictions $\mathcal{Z}^{\mathrm{PU}}$ are not biased, the held-out predictions of a ten-fold cross validation are used.

Third, a posteriori probabilities for the training sample set $p(y_+|z_i)$ (see Figure III.1 b.1) are derived with Bayes' rule

$$p(y_+|z_i) = \frac{p(z_i|y_+)P(y_+)}{p(z_i)} \tag{III.2}$$

where $z_i \in \mathcal{Z}$ is the predicted value for sample $\mathbf{x}_i$ (see also Equation (III.1)). In the same way, the a posteriori probabilities for the test set $p^{\mathrm{te}}(y_+|z_i)$ can be obtained. Thus the estimation of the conditional probabilities $p(y_+|z_i)$ and the a priori probabilities $P(y_+)$ are conducted in one-dimensional feature space. In this study a standard kernel density estimation method is used for the estimation of the probability density functions (see Section 3.2), but also other suitable density estimation techniques can be applied. The estimation of the a priori probability is done using the approach of (Guerrero-Curieses et al., 2002) and explained in detail in Section 3.3.

The diagnostic plot provides evidence on the plausibility of the Bayes' analysis, *i.e.*, the estimated quantities $\hat{p}(z_i|y_+)$, $\hat{P}(y_+)$, $\hat{p}(z_i)$, $\hat{p}(y_+|z_i)$, and of given binarization threshold, such θ^{MAP} derived from $\hat{p}(y_+|z_i)$ or the default threshold θ^0 of the BSVM. It may thus reveal if inadequate models are used for estimating these quantities and/or if critical assumptions are violated. For example, a certain degree of class separability is usually assumed for estimating $P(y_+)$ (see Section 3.3). The visualized quantities in the diagnostic plot constitutes an informative source for interpretation and evaluation of the classification result, which is especially valuable if no complete and representative test set is available. Therefore, if implausible estimates are diagnosed, the user can go back to one of the previous steps in order to improve the results.

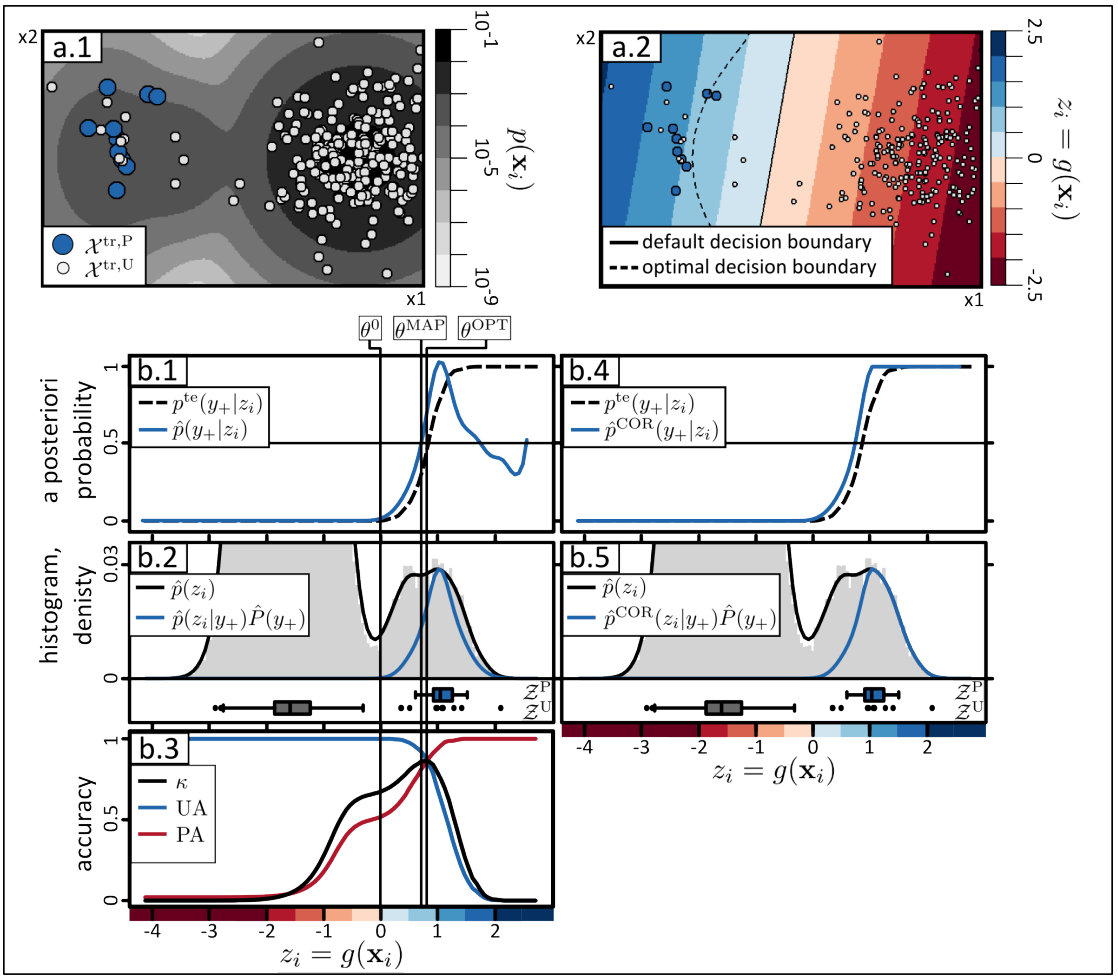

FIGURE III.1: Illustration of the strategy with the two-dimensional synthetic data set. The training data (**a.1**) and the thereof derived BSVM model $g(\cdot)$ (**a.2**) are shown. Compared to the default threshold of the BSVM θ^0 the threshold derived from the a posteriori probability θ^{MAP} is closer to the optimal threshold θ^{OPT} (**b.1**). The diagnostic plot (**b.2**) is useful to gain a rough idea of the accuracy of the one-class classification output and the plausibility of the estimated terms of the Bayes' rule used to derive the a posteriori probability. It shows the histogram of the predicted image, the distribution of $\mathcal{X}^{\mathrm{tr,PU}}$ in the output space of $g(\cdot)$, *i.e.*, $\mathcal{Z}^{\mathrm{PU}}$ (boxplots), and the thereof derived densities. In this example, the diagnostic plot gives evidence to rather trust θ^{MAP} than θ^{OPT} (see Section 2 for a detailed explanation). This is confirmed by the threshold dependent accuracy assessment (**b.3**), which cannot be estimated in a OCC application. Also implausible estimations of the required terms of the Bayes' rule, *i.e.*, $\hat{p}(z_i)$, $\hat{p}(z_i|y_+)$, and $\hat{P}(y_+)$ can be detected and sometimes improved by simple approaches (see (**b.4**) and (**b.5**), and Equation (III.5)). After the improvement, the estimated, $p(y_+|z_i)^{\mathrm{COR}}$, and test, $p(y_+|z_i)^{\mathrm{te}}$, a posteriori probabilities are similar over the whole output range (**b.4**). Please refer to the text for detailed explanations.

Let us first evaluate the two thresholds θ^0 and θ^{MAP} based on the diagnostic plot (Figure III.1 b.2). It can be observed that θ^0 and θ^{MAP} differ significantly. Apriori, we should not prefer one of the two thresholds because if any of the estimates $\hat{p}(z_i|y_+)$, $\hat{P}(y_+)$, $\hat{p}(z_i)$ are not plausible θ^{MAP} can lead to poorer binary classification result than θ^0. Therefore, careful interpretation is required in order to decide which threshold is more plausible.

The histogram of $\mathcal{Z}$ shows two main clusters of data which are separated by a low density region at $z_i \approx 0$ (Figure III.1 b.2). The default threshold of the BSVM θ^0 is located in this low density region. It is tempting to believe that the data right of θ^0 belong to the positive class and left to the θ^0 to the negative class. However, the distribution of the positive data $\mathcal{Z}^{\mathrm{P}}$ (the blue boxplot in Figure III.1 b.2) does not support such believes. It rather provides evidence that only a part of the data right to the low density area belongs to the positive class. Under the assumptions that (i) the positive training data $\mathcal{X}^{\mathrm{tr,P}}$ contains representative samples of the positive class; and (ii) the cross-validated values $\mathcal{Z}^{\mathrm{P}}$ of $\mathcal{X}^{\mathrm{tr,P}}$ are not strongly biased, θ^{MAP} can be approved to be more suitable. More precisely, if the Bayes' analysis is valid, we have to expect that the threshold θ^0 leads to a very high producer's accuracy (*i.e.*, true positive rate) but also a very low user's accuracy (*i.e.*, a high false positive rate). Instead, with θ^{MAP} we can expect that the producer's and user's accuracies for the positive class are rather balanced. It is proved by the threshold dependent accuracies in Figure III.1 b.3 that this interpretation is correct. Please note that if we would belief that (i) $\theta^0 = 0$ is a suitable threshold and (ii) over-predictions of the hold-out predictions $\mathcal{Z}^{\mathrm{P}}$ are unlikely, than this implies that the positive training set is not representative and does not cover an important part of the positive class exhibiting differing spectral characteristics. In order to draw the right conclusion, the user should recall all knowledge, expectations and believes to judge the derived estimations.

This example also shows that the diagnostic plot is useful for understanding if the size of unlabeled training data $|\mathcal{X}^{\mathrm{tr,U}}|$ is suitable. Remember $\mathcal{Z}^{\mathrm{U}}$ are the cross-validated predictions of the unlabeled training data $\mathcal{X}^{\mathrm{tr,U}}$ and are visulized by the grey boxplot in the diagnostic plot (Figure III.1 b.2). Here, the large part of the samples are located at very low z-values and only seven samples, *i.e.*, 3 % of the unlabeled samples, exhibit $z \geq 0$. This means that the most relevant region of the feature space, *i.e.*, where the optimal decision boundary should be located, is not sampled very well (see Figure III.1 a.2). This also explains why the default BSVM threshold θ^0 is very low. Therefore, in a practical application we would rather re-train the BSVM with a more suitable, e.g., larger, set of unlabeled training samples. Eventually, this could improve the discriminative power of the model.

Let us now evaluate the a posteriori probabilities. Figure III.1 b.1 shows that the a posteriori probabilities derived from the training and test sets are similar over a large part of the output range. However, at high z-values $\hat{p}(y_+|z_i)$ is obviously implausible. We reasonably assume that the a posteriori probability is monotonically increasing in z, which is not the case in Figure III.1 b.1. Here, the drop of the a posteriori probabilities are not plausible but rather an artifact of the non-matching densities in Figure III.1 b.2.

Thanks to the simple structure of the one-dimensional feature space it is easy to correct for such implausible effects as is shown in Figure III.1 b.4, b.5 (see Section 3.4).

It has already been argued in Section **??** that there is no OCC approach which is likely to perform optimally in all classification problem. The same is true for the density and a priori estimation approaches. Thus, it is not the objective of this paper to promote any particular approach for $g(\cdot)$ or to derive $\hat{p}(z_i|y_+)$, $\hat{P}(y_+)$, $\hat{p}(z_i)$. Instead, it is recommended to start with simple approaches for all the steps, analyze the outcome and improve or change the approximations where necessary.

3 Implementation of the Framework

In this section we shortly describe the methods used for the (i) one-class classification; (ii) density estimation; (iii) estimation of the prior probability; and (iv) optimization of the density estimation, *i.e.*, $g(\cdot)$, $p(z_i|y_+)$, $p(z_i)$, and $P(y_+)$. To keep the paper concise only one method is considered for each of the estimation problems. However, the user can chose among different methods to find an optimal solution.

3.1 Biased Support Vector Machine

For the experiments in this paper the biased SVM (BSVM) (B. Liu et al., 2003) is used to implement the one-class classifier $g(\cdot)$. The BSVM is a special formulation of the binary SVM which is adapted to solve the OCC problem with a positive and unlabeled training set $\mathcal{X}^{\mathrm{tr,PU}}$.

Two mis-classification cost terms C_+ and C_0 are used for the positive and unlabeled training samples. If the unlabeled training set is large enough it contains a significant amount of positive samples. On the other hand, the positive training set is labeled and therefore no or only few negative samples are contained in it. Thus, it is reasonable to penalize the mis-classifications on the unlabeled training samples less strong. As in the case of the binary SVM the kernel trick can be applied to create a non-linear classifier by fitting the separating hyperplane in a transformed feature space. The Gaussian radial basis function is maybe the most commonly applied kernel and is also used here. Thus, the inverse kernel width σ needs to be tuned additionally to C_+ and C_0.

Tuning these three parameters is done by performing a grid search over the combinations of pre-specified parameter values. To select the optimal parameter combination a performance criteria is required which is estimated from the positive and unlabeled training data. Given the nature of the data, a reasonable goal is to correctly classify

most of the positive labeled samples while minimizing the number of unlabeled samples to be classified as positives. This goal can be achieved by the performance criteria $\mathrm{PC}^{\mathrm{PU}}$ (X. Li and B. Liu, 2003)

$$\mathrm{PC}^{\mathrm{PU}} = \frac{P(\hat{y}_+|y_+)^2}{P(\hat{y}_+)} \tag{III.3}$$

where $P(\hat{y}_+|y_+)$ is the true positive rate and $P(\hat{y}_+)$ is the probability that a unlabeled sample is classified as positive. $\mathrm{PC}^{\mathrm{PU}}$ is estimated by cross-validation from $\mathcal{X}^{\mathrm{tr,PU}}$.

The BSVM has been implemented in R (R Development Core Team, 2013) via the package kernlab (Karatzoglou et al., 2004).

3.2 Density Estimation

For the estimation of $p(z_i|y_+)$ an univariate kernel density estimation with adaptive kernel bandwidth is used as implemented in the package pdfCluster (Azzalini and Menardi, 2014). An adaptive kernel density estimation has been selected due to the fact that the size of $\mathcal{Z}^{\mathrm{P}}$ is relatively small. In contrast, $p(z_i)$ can be estimated from the large data set $\mathcal{Z}$ and thus, it is estimated by a univariate kernel density estimation with fixed bandwidth. This is computationally feasible even with a large data set such as $\mathcal{Z}$. Here the implementation of the R base environment (R Development Core Team, 2013) is used.

3.3 Estimation of the a priori Probability

The a priori probability $P(y_+)$ is estimated following the approach in (Guerrero-Curieses et al., 2002), which is straightforward once the estimates $\hat{p}(z_i|y_+)$ and $\hat{p}(z_i)$ are available. Accurate estimation of $p(z_i|y_+)$ and $p(z_i)$ are thus a prerequisite for an accurate estimation of $P(y_+)$. The approach assumes that the positive and the negative class distributions do not overlap at the point $\tilde{z}$, *i.e.*, $p(\tilde{z}|y_-) = 0$. If this is true $P(y_+)$ can be derived with the following equation

$$P(y_+) = \frac{p(\tilde{z}|y_+)}{p(\tilde{z})} \tag{III.4}$$

In the experiments, the median of the cross-validated positive training samples $\mathcal{Z}^{\mathrm{P}}$ (the blue boxplots in the diagnostic plot) is used to determine $\tilde{z}$. The visualization of the estimations $\hat{p}(z_i)$ and $\hat{p}(z_i|y_+)\hat{P}(y_+)$ allows to examine their plausibility and gives evidence if the separability assumption is reasonable.

3.4 Optimizing the Density Estimation

If $p(z_i|y_+)$ and $p(z_i)$ are estimated independently, $\hat{p}(z_i|y_+)$ can be adjusted to match with $\hat{p}(z_i)$ at high z-values. For this region it is usually justifiable to assume that $p(z_i|y_-)$ equals zero, or equivalently, to assume that only the positive class contributes to $p(z_i)$. Based on this assumption, $\hat{p}(z_i|y_+)$ can be adjusted by applying the following rule:

$$\hat{p}^{\mathrm{COR}}(z_i|y_+) = \begin{cases} \hat{p}(z_i|y_+) & \text{if } z < z^{\mathrm{COR}} \\ \frac{\hat{p}(z_i)}{\hat{P}^{\mathrm{COR}}(y_+)} & \text{otherwise} \end{cases} \tag{III.5}$$

where z^{COR} is the z-value where $\hat{p}(y_+|z_i)$ first reaches one. This means that we force $\hat{P}(y_+)\hat{P}^{\mathrm{COR}}(y_+)$ to accurately correspond to $\hat{p}(z_i)$ for high z-values (compare Figure III.1 b.4, b.5).

4 Data and Experiments

4.1 Data

In the experiments of this paper, a Hyperion spaceborne imaging spectroscopy dataset (Figure III.2) is used to demonstrate the strategy. The data was acquired at 24 May 2012 over an agricultural landscape located in Saxony Anhalt, Germany (image center latitude/longitude: 51°23′01.62″N/11°44′39.12″E). The Level 1 Terrain Corrected (L1T) product of the image has been used. In order to further increase the geometric accuracy the image was shifted with a linear transformation according to eight ground control points selected uniformly over the image. The nominal size of a pixel at ground is 30 m.

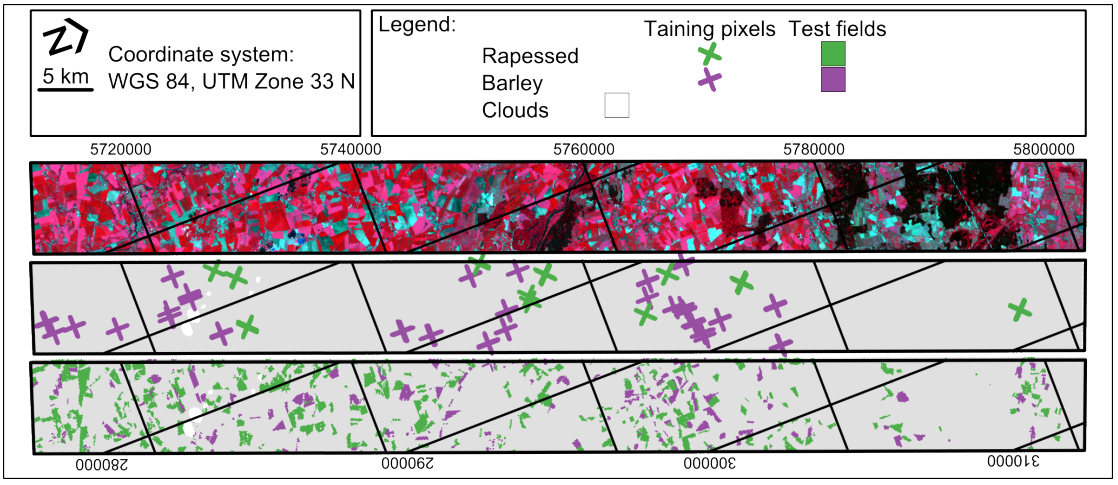

FIGURE III.2: Image data and reference information used in the experiments.

From the 242 spectral bands 87 bands with low signal to noise ratio have been removed. The remaining 155 bands are located in the spectral ranges 426 nm–1336 nm (88 bands),

1477 nm–1790 nm (32 bands), and 1982 nm–2355 nm (35 bands). The pixel values of each spectral band were independently linearly scaled between 0 and 1.

The reference data used in this study was provided by the Ministry of Agriculture and Environment, Saxony Anhalt, Germany. The information was gathered in the framework of the Integrated Administration and Control System of the European Union. In order to receive financial support from the European Union the farmers need to declare the outlines of the agricultural parcels and the land use/land cover. It is assumed that all parcels of the classes of interest analyzed in this study (rapeseed and barley) have been declared and that irregularities can be neglected.

To evaluate the proposed strategy, the specific objective in our study is the classification of rapeseed (Example A, Section 5.1) and barley (Example B, Section 5.2). While we expect that the classification of rapeseed is relatively simple, the classification of barley is more challenging due to parcel size and spectral ambiguities between different cereal crop types.

For each class of interest, the following steps are a carried out to create the training and test sets. First, a fully labeled reference image $\mathcal{Y}$ corresponding to the Hyperion image data $\mathcal{X}$ was created. The pixels within a parcel of the positive class were labeled positive and all other pixels were labeled negative (Figure III.2). Additionally to $\mathcal{Y}$ we created a reference set $\mathcal{Y}^{\text{INT}}$ without pixels at class borders, such that $\mathcal{Y}^{\text{INT}} \in \mathcal{Y}$, in order to prevent dealing with mixed pixels (Table III.2). This was done by excluding the pixels with positive and negative class occurrences in the spatial 3×3 neighborhood.

	$\boldsymbol{\mathcal{X}}^{\text{tr,PU}}$		$\boldsymbol{\mathcal{X}}^{\text{te}}$			$\boldsymbol{\mathcal{X}}^{\text{te,INT}}$		
Class	**P**	**U**	**P**	**N**	$\boldsymbol{P(y_+)}$	**P**	**U**	$\boldsymbol{P(y_+)}$
Rapeseed	30	5000	96,787	775,317	0.11	63,924	732,540	0.08
Barley	75	5000	38,638	836,456	0.04	24,507	809,008	0.03

TABLE III.2: Overview over the number of positive(P), unlabeled (U) and negative (N) training ($\mathcal{X}^{\text{tr,PU}}$) and test set sizes, where $\mathcal{X}^{\text{te}}$ comprises all pixels of the test fields and $\mathcal{X}^{\text{te,INT}}$ only the interior fields.

In order to generate the training set we randomly selected 50 parcels. The total number of parcels available for the class rapeseed was 626 and for the class barley 315. The positive training pixels $\mathcal{X}^{\text{tr,P}}$ were randomly selected among the non-border pixels of these parcels to minimize the probability of outliers in the set. For the rapeseed experiment we selected 30 and for the barley experiment 75 positive training samples. In both experiments 5000 pixels were selected randomly from the whole image and used as unlabeled training samples for $g(\cdot)$.

It is important to note that for a one-class classification the required number of positive labeled training data might be higher than in the case of supervised classification in

order to yield good classification results. This is particularly true for approaches which estimate a posteriori probabilities in high-dimensional feature space. The number of labeled training samples used in many of these experiments are moderately to very large, *i.e.*, between 100 and 3000 (Jeon and Landgrebe, 1999; Mantero, Moser, and S. Serpico, 2005; W. Li, Guo, and Elkan, 2011; Muñoz-Marí et al., 2010; Marconcini, Fernandez-Prieto, and Buchholz, 2014).

4.2 Experimental Setup

The two experiments presented in this paper are based on the data described in Section 4.1 and the methods described in Section 3. They have been selected in order to demonstrate the usefulness of the diagnostic plots in the context of model selection, derivation of a posteriori probabilities, and threshold selection.

We first selected suitable model parameters for the BSVM based on $\mathrm{PC}^{\mathrm{PU}}$ (Equation (III.3)) by ten-fold cross-validation using the training set $\mathcal{X}^{\mathrm{tr,PU}}$. The cross-validation is also used to generate the sets $\mathcal{Z}^{\mathrm{P}}$ and $\mathcal{Z}^{\mathrm{U}}$ used for constructing the diagnostic plots. The final model is trained with the selected parameters and the complete training data and used to derive the predicted image, *i.e.*, $\mathcal{Z}$.

Next, we estimate $p(z_i|y_+)$ with $\mathcal{Z}^{\mathrm{P}}$ and $p(z_i)$ with $\mathcal{Z}$ (see Section 3.2). With derived density models and $\tilde{z}$ derived from $\mathcal{Z}^{\mathrm{P}}$ we estimate $P(y_+)$ with Equation (III.4). Now the a posteriori probability $p(y_+|z_i)$ can be calculated by applying Bayes' rule (Equation (III.2)) which also gives the θ^{MAP}. Finally, Equation (III.5) is used for correcting $p(z_i|y_+)$ and $p(y_+|z_i)$ at high z-values.

Based on these estimates we construct the diagnostic plots.

With the test set $\mathcal{X}^{\mathrm{te}}$ we perform an accuracy assessment for the binary classification results over the whole range of possible thresholds. Additionally to the confusion matrix we derive the overall accuracy (OA), Cohen's kappa coefficient (κ), the producer's accuracy (PA), and the user's accuracy (UA) for the whole range of possible thresholds. Three thresholds are of particular interest: the "default" threshold θ^0, *i.e.*, 0 and corresponds to the hyperplane of the BSVM, the maximum a posteriori threshold θ^{MAP} *i.e.*, the the z-value where $p(y_+|z_i)$ first exceeds 0.5, and the optimal threshold θ^{OPT} *i.e.*, the threshold which maximizes κ. It is worth to underline that θ^{OPT} cannot be derived in context of a real application, due to the incomplete reference data. However, it is used to analyze the experimental results.

The a posteriori probabilities are evaluated by estimating $p(z_i)$, $p(z_i|y_+)$, $P(y_+)$ and $p(y_+|z_i)$ with the test sets $\mathcal{X}^{\text{te}}$ and $\mathcal{X}^{\text{te,INT}}$. Whereas, $\mathcal{X}^{\text{te,INT}}$ better represents the population from which the positive samples $\mathcal{X}^{\text{tr,P}}$ have been sampled.

For the class barley different diagnostic plots are generated to assess the potential of the plots in context of model selection. This experiment shows that the diagnostic plot can be helpful for manual model selection when the automatic selection process, here based on PC^{PU}, selects an unconvincing model.

The statistical significance of the difference in accuracy has been evaluated a two-sided test based on the kappa coefficient (Foody, 2004) for all compared binary classification results. The widely used 5 percent level of significance has been used for determining if there is a difference. Note, that due to the high amount of test samples also relatively small differences in accuracy are significantly different.

5 Results and Discussion

5.1 Experiment 1: Rapeseed

The class rapeseed can be classified with very high accuracy. Table III.3 show the confusion matrices and additional accuracy measures given the three thresholds θ^0 (at $z = 0$), $\hat{\theta}^{\text{MAP}}$ (at $z = -0.17$), and θ^{OPT} (at $z = -0.42$). The overall accuracy and kappa coefficients exceed 97 % and 0.85 respectively given any of the three thresholds. Although the three thresholds provide comparable kappa coefficients are statistically significant at a 5 % percent level of significance (Foody, 2004).

	θ^0			$\hat{\theta}^{\text{MAP}}$			θ^{OPT}		
	(+)	(−)	**UA**	(+)	(−)	**UA**	(+)	(−)	**UA**
(+)	80,941	4838	94.4%	83,119	5906	93.37%	85,899	8057	91.4%
(−)	15,846	770,479	97.98%	13,668	769,411	98.25%	10,888	767,260	98.6%
PA	83.6%	99.4%	PA	85.9%	99.2%	PA	88.8%	99%	
OA/κ	97.6%/0.87			97.8%/0.88			97.8%/0.89		

TABLE III.3: Confusion matrices and accuracy measures for the class rapeseed given the threshold θ^0 obtained by the BSVM (left), $\hat{\theta}^{\text{MAP}}$ obtained by Bayes' rule (middle), and the optimal threshold θ^{OPT} (right).

These findings are clearly reflected in the diagnostic plot (Figure III.3b). The predictive values of the positive class $\mathcal{Z}^{\text{P}}$ (shown as blue boxplot) correspond well with a distinctive cluster of predicted unlabeled data with high z-values. The wide low density range separating the two clusters corresponds to the wide range of thresholds leading to high classification accuracies. In this experiment, we can be confident to derive a good binary classification result with any threshold in the low density range.

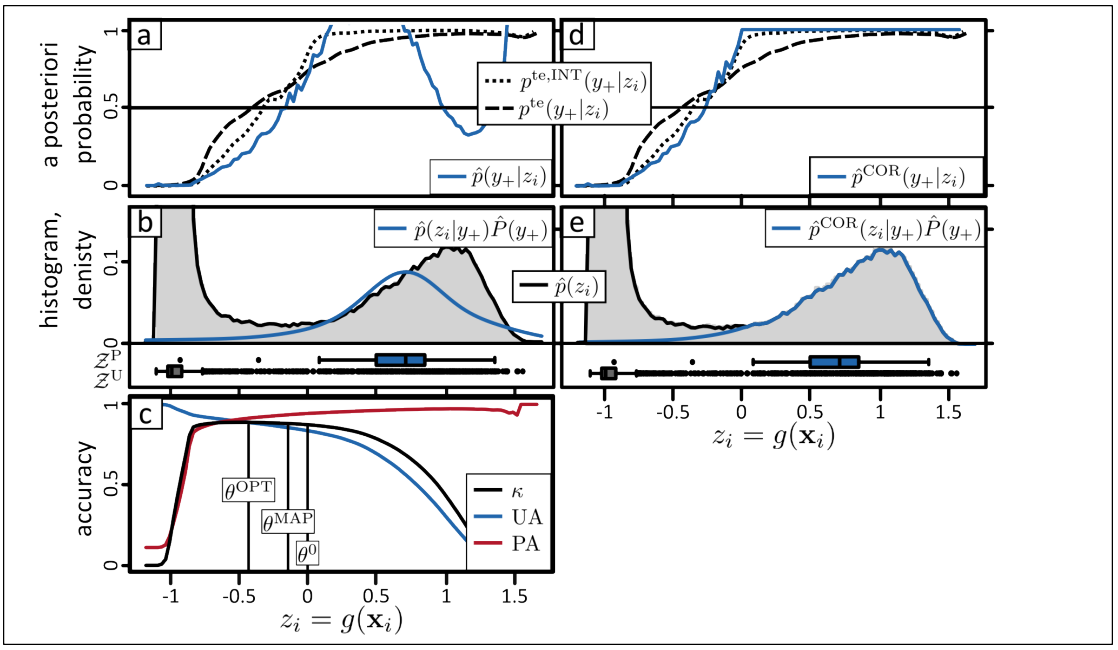

FIGURE III.3: A posteriori probability (**a**), diagnostic plot (**b**) and the threshold dependent accuracy (**c**) for the rapeseed example. Optimizing the conditional density (see Section 3.4) leads to improved a posteriori probabilities at high z-values (**d**,**e**).

The visual assessment of the classification and error maps underlines this findings (Figure III.4). It is well known, that spectral properties of boundary pixels might be a mixture between both classes (e.g., two different crop types). Consequently these mixed pixels do not represent either of the two land cover classes and consequently a mis-classification is more likely to occur.

Deriving accurate a posteriori probabilities is more challenging, particularly with few positive training samples, as in the case here. Under the assumption that the data of the right cluster in Figure III.3b belongs to the positive class, the distributions $\hat{p}(z_i|y_+)\hat{P}(y_+)$ and $\hat{p}(z_i)$ should coincide in this range. However, $\hat{p}(z_i|y_+)\hat{P}(y_+)$ is less skewed towards high z-values than $\hat{p}(z_i)$.

We assume the reason for the discrepancy to be the size of positive training data. It is possible that the small size, *i.e.*, 30, is not sufficient to accurately capture the real distribution of the positive class. Moreover, one may argue that the redundancy is relatively low in a small training data set. When performing cross-validation with such a small set the hold out predictions are more likely to exhibit significantly lower values compared to the predictive values of similar data points predicted with the final model trained with all samples. Furthermore, if $\hat{p}(z_i|y_+)$ cannot be trusted it is unlikely that Equation (III.4) provides an accurate estimate of $\hat{P}(y_+)$.

The visualization of the estimated densities (Figure III.3b) and a posteriori probabilities (Figure III.3a) supports the identification of implausible estimations and helps to

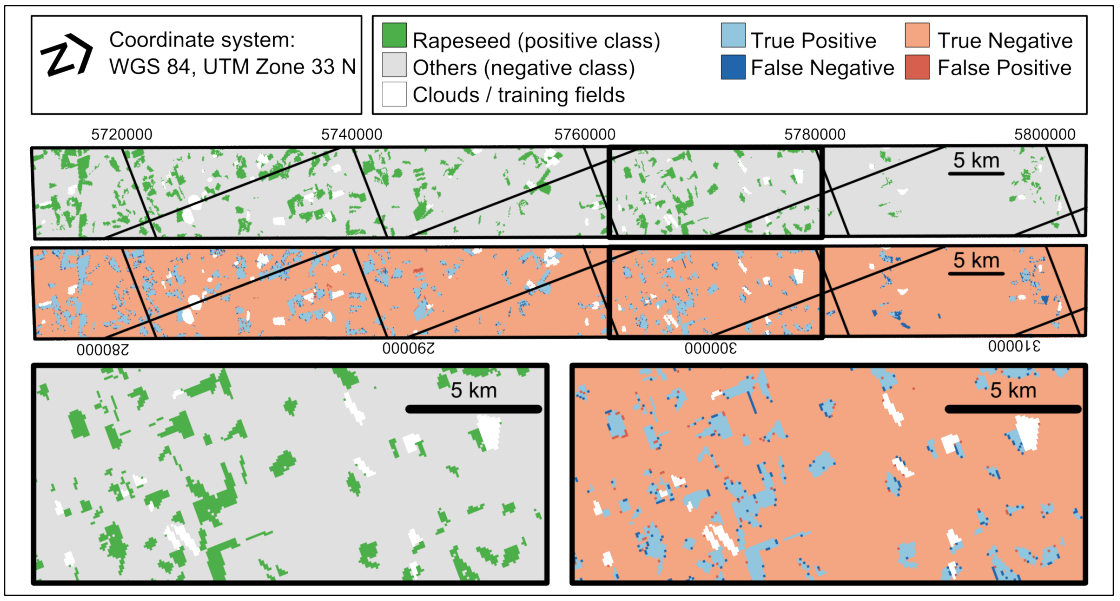

FIGURE III.4: Classification (**upper** image and **bottom left** image) and test errors (**middle** image and **bottom right** image) for the class rapeseed realized with the threshold $\hat{\theta}^{\text{MAP}}$ (see Figure III.3, Table III.3).

find more suitable solutions. To improve the a posteriori probabilities, $\hat{P}(y_+)$ has been re-calculated by the fraction of pixels with $z \geq -0.25$, *i.e.*, in the middle of the low density area. Regarding the visual interpretation of the diagnostic plot and the clear high separability of the classes, it seems adequate to re-calculate $\hat{P}(y_+)$ by this approach. Remember that $\hat{P}(y_+)$ is calculated by $\frac{\hat{p}(\tilde{z}|y_+)}{\hat{p}(\tilde{z})}$, where $\tilde{z}$ is the median of $\mathcal{Z}^{\text{P}}$ (see Section 3.3). Due to the fact that (i) $\hat{p}(z_i|y_+)$ and $\hat{p}(z_i)$ do not match very well at $\tilde{z}$ and (ii) the separability is very high it is likely that the alternative way of estimating $P(y_+)$ is more accurate.

Then the adjusted $\hat{p}^{\text{COR}}(z_i|y_+)\hat{P}(y_+)$ (Equation (III.5)) has been used to estimate the a posteriori probability. Figure III.3d,e show that these solutions substantially improved $\hat{p}(y_+|z_i)$, which remains at a constant value of one for high z-values. Over the complete range of z it is now very close to $p^{\text{te,INT}}(y_+|z_i)$, *i.e.*, the a posteriori probabilities derived with the test set without boundary pixels (Figure III.3d,e). As expected, a stronger discrepancy exists between $\hat{p}(y_+|z_i)$ and $p^{\text{te}}(y_+|z_i)$ due to the influence of mixed pixels and geometric inaccuracies.

5.2 Experiment 2: Barley

As already underlined, in a practical OCC application no complete and representative validation set is available. Therefore, the OA or other accuracy measures based on complete validation sets cannot be estimated and cannot be used for the task of model

selection. Instead, alternative performance measures, such as PC^{PU} (Equation (III.3)), are used which can be derived from PU-data. However, as is the case in this experiment, these measures do not consequently lead to the optimal models in terms of the classification accuracy. In this experiment a positive but noisy relationship exists between PC^{PU} and the OA (see Figure III.5a). The noisiness is typical for PU-performance measures is a problem, as in this experiment, when the highest PC^{PU} value points to a model with relatively low overall accuracy. Assuming the optimal threshold can be found, the selected model (model b in Figure III.5) leads to an overall accuracy of 97.0% ($\kappa = 0.57$) while the optimal model (model g in Figure III.5) to an overall accuracy of 97.9% ($\kappa = 0.73$).

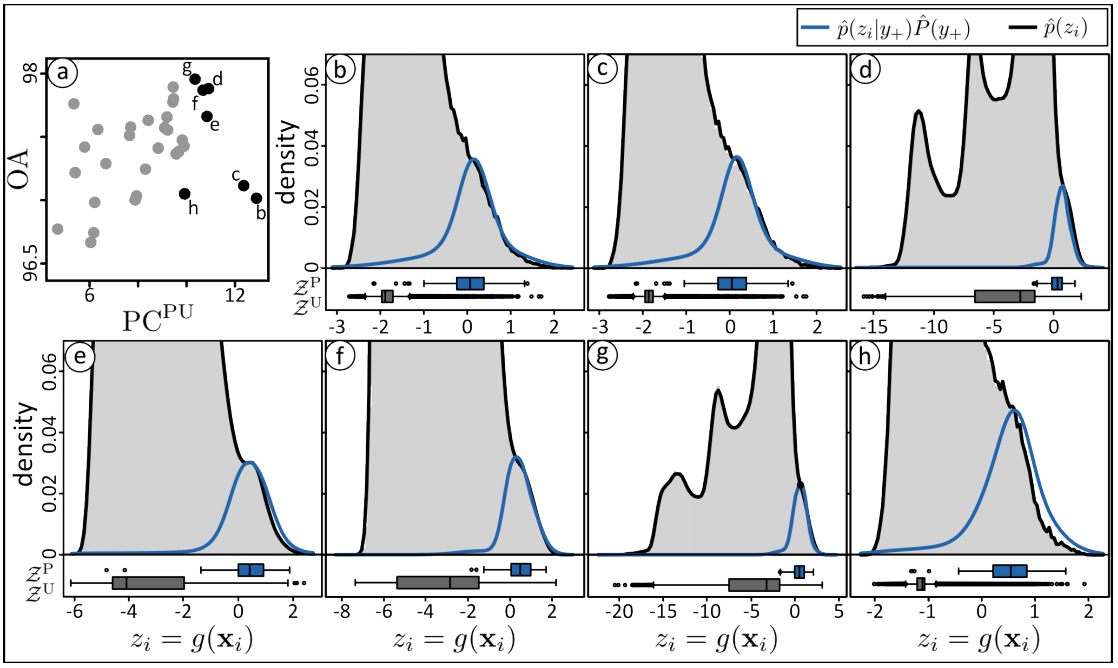

FIGURE III.5: (**a**) Optimization criteria PC^{PU} and maximum overall accuracy OA of BSVM models with different parameterizations. The highest PC^{PU} (**b**) has relatively low OA. The diagnostic plots of the seven models with highest PC^{PU} (black points in (**a**)) are shown in (**b**–**h**). (**e**) is a reasonable choice because the positive data is well clustered at high z-values and it can be best associated with a distinct bunch of data in the histogram and $p(z_i)$.

It is shown in Figure III.5 that comparing the diagnostic plots of different models can support the selection of a more suitable model when the automatic approach fails. In order to select a more accurate model the user can sequentially analyze the diagnostic plot of other models, e.g., in decreasing order of the optimization criteria PC^{PU}. Between different diagnostic plots we would select the one where (i) the positive data $\mathcal{Z}^P$ is most concentrated at high z-values and (ii) where these samples correspond to a distinctive cluster of unlabeled data. Following these rules we would select the model shown in Figure III.5e out of the seven options shown in Figure III.5b–h. Table III.4 shows the

accuracies, given θ^{OPT}, of (i) the model with maximum $\mathrm{PC}^{\mathrm{PU}}$(model b, see also Figure III.5b); (ii) the model selected manually following the argumentation above (model e, see also Figure III.5e); and (iii) the model with the highest overall accuracy (model g, see also Figure III.5g). The overall accuracy/kappa coefficient of the manually selected model (97.7%/0.68) is 0.7%/0.11 higher than the ones of the model with maximum $\mathrm{PC}^{\mathrm{PU}}$ (97.0%/0.57) and only 0.02%/0.04 smaller than the model with highest overall accuracy. Thus, in this experiment the diagnostic plot helps to select a model with significantly higher discriminative power compared to the model selected by maximizing $\mathrm{PC}^{\mathrm{PU}}$ (see Figure III.4). The findings are confirmed by a significance test returning statistically significant differences of the kappa coefficients at a 5% percent level of significance (Foody, 2004).

	$\boldsymbol{\theta}^{\mathrm{OPT}}$ **,b**			$\boldsymbol{\theta}^{\mathrm{OPT}}$**, e**			$\boldsymbol{\theta}^{\mathrm{OPT}}$**, g**		
	(+)	(−)	**UA**	(+)	(−)	**UA**	(+)	(−)	**UA**
(+)	18,530	6022	75.5%	23,158	5039	82.1%	24,904	4821	83.7%
(−)	20,108	830,434	97.6%	15,480	831,417	98.2%	13,734	831,635	98.4%
PA	48.0%	99.3%		60.0%	99.4%		64.5%	99.4%	
OA/κ	97.0%/0.57			97.7%/0.68			97.9%/0.72		

TABLE III.4: Confusion matrices and accuracy measures given θ^{OPT} for the model b selected by maximizing $\mathrm{PC}^{\mathrm{PU}}$ (b), the manually selected model (e), and the optimal model, in terms of the maximum OA (f). See also the corresponding diagnostic plots in Figure III.5b,e,g.

Based on the diagnostic plot of the manually selected model (Figure III.6) a substantial amount of mis-classifications has to be expected. Contrary to the rapeseed example (Figure III.3) there is no low density region separating the positive and negative class regions. Thus, the distributions of the two classes overlap and lead to significant misclassifications for any given threshold (Table III.5). As in the rapeseed example, the three thresholds provide comparable accuracies but due to the high amount of test samples the differences between the kappa coefficients are statistically significant at a 5% percent level of significance (Foody, 2004).

Also, the classifier performance, which is limited in comparison to the accuracies provided for rapeseed, can be assessed by the diagnostic plot. The analysis of the diagnostic plot (Figure III.6b) underlines among others the threshold dependent trade-off between false positive and false negative classifications. Starting from $\theta^0 = 0$ and moving the threshold to the left apparently increases the false negative classification stronger than it reduces the false negative classifications. This can be concluded by the steep slope of $\hat{p}(z_i)$ in this region.

The higher class overlap in the feature space is also underlined by the visual interpretation of the classification map (Figure III.7). As in the rapeseed example, several boudary

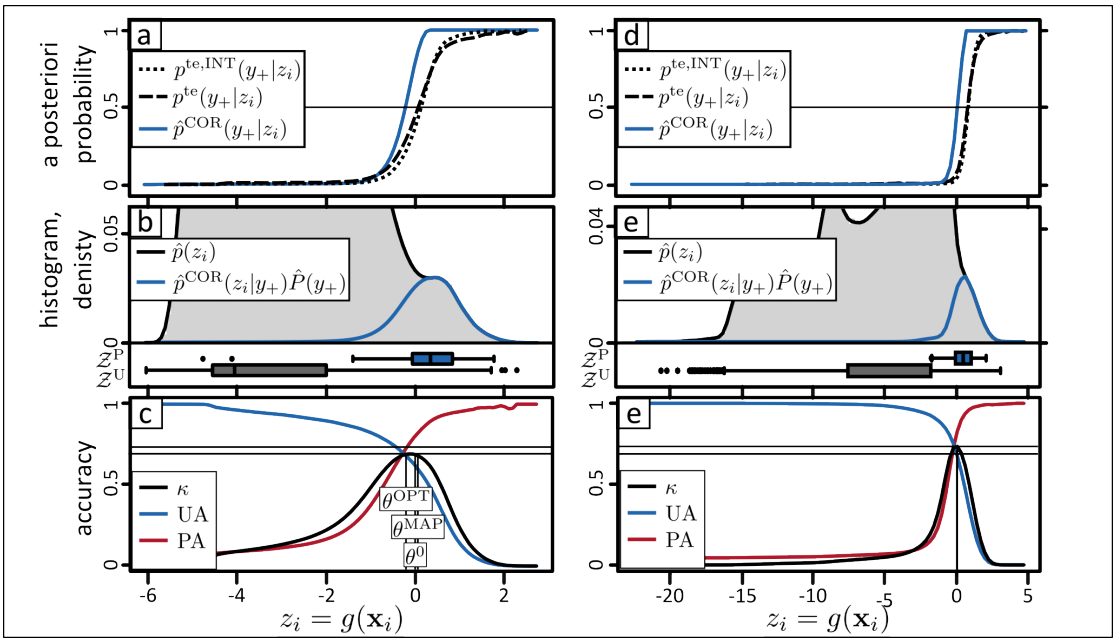

FIGURE III.6: A posteriori probabilities, diagnostic plot, and threshold dependent accuracy for the manually selected model (**a**–**c**) and the optimal model (**d**–**f**) of the barley example (see also Figure III.5).

	θ^0			$\hat{\theta}^{\text{MAP}}$			θ^{OPT}		
	(+)	(−)	**UA**	(+)	(−)	**UA**	(+)	(−)	**UA**
(+)	24,016	5890	80.3%	26,364	9939	72.6%	23,158	5039	82.1%
(−)	14,622	830,566	98.27%	12,274	826,517	98.54%	15,480	831,417	98.2%
PA	62.2%	99.3%		68.2%	98.8%		59.9%	99.4%	
OA/κ	97.7%/0.69			97.5%/0.69			97.7%/0.68		

TABLE III.5: Confusion matrices and accuracy measures for the class barley realized with the manually selected model (see Figure III.5e) given the threshold θ^0 obtained by the BSVM (left), $\hat{\theta}^{\text{MAP}}$ obtained by Bayes' rule (middle), and the optimal threshold θ^{OPT} (right)

pixels are missclassified. The errors at the class border are mainly false negatives, which is in contrast to the rapeseed example where false positives and false negatives occurred in similar amounts. However, the significant amount of false negatives was to be expected, regarding the visual interpretation of the diagnostic plots. As in other studies, these mis-classifications firstly occur at pixels which lie along the boundaries of two objects, e.g., two field plots. Moreover, some complete mis-classified fields are obvious in the north of the study site. However, it is well known that the classification of agricultural areas can be affected by site-internal variations. Therefore we assume the reason for the mis-classifications to be crop growing conditions, which are different in the affected part of the study area.

At this point it is also worth noting that the diagnostic plot extends the interpretability of the classification map alone (Figure III.7). Usually, noisier classification results (*i.e.*,

maps with a strong "salt and pepper" effect), such as the map in Figure III.7, are assumed to contain more errors. Although this assumption might be fulfilled in specific case studies (e.g., (Waske and Braun, 2009)), it is not generally recommendable to base decisions related to model or threshold selection on the appearance of the classification map alone. For example, a lower threshold could lead to a less noisy classification map because the additional false positives possibly occur in clumps, e.g., in the fields of the most similar land cover class. As discussed before, careful analysis of the distributions of $\mathcal{Z}^{\mathrm{P}}$ and $\mathcal{Z}$ reveal such over-predictions.

The example also shows that the derivation of accurate a posteriori probabilities is challenging in the case of strongly overlapping classes (Figure III.6). Here, $\hat{p}(y_+|z_i)$ deviates significantly from both $p^{\mathrm{te}}(y_+|z_i)$ and $p^{\mathrm{te,INT}}(y_+|z_i)$. Nevertheless, this seems expectable following the interpretation of the diagnostic plot and the proposed strategy. Remember that the estimation of $\hat{P}(y_+)$ is based on the assumption that $\hat{P}(y_-)$ is zero at the median of $\mathcal{Z}^{\mathrm{P}}$ (Equation (III.4)). But in this example it is unlikely that the assumption holds because $\hat{p}(z_i)$ rises steeply just to the left of this point. Therefore, it has to be assumed that there is still a significant negative density at $\tilde{z} = 0.36$, resulting in a smaller $\hat{P}(y_+)$, lower $\hat{p}(z_i|y_+)\hat{P}(y_+)$, and a shift of the $\hat{p}(y_+|z_i)$-curve towards higher z-values.

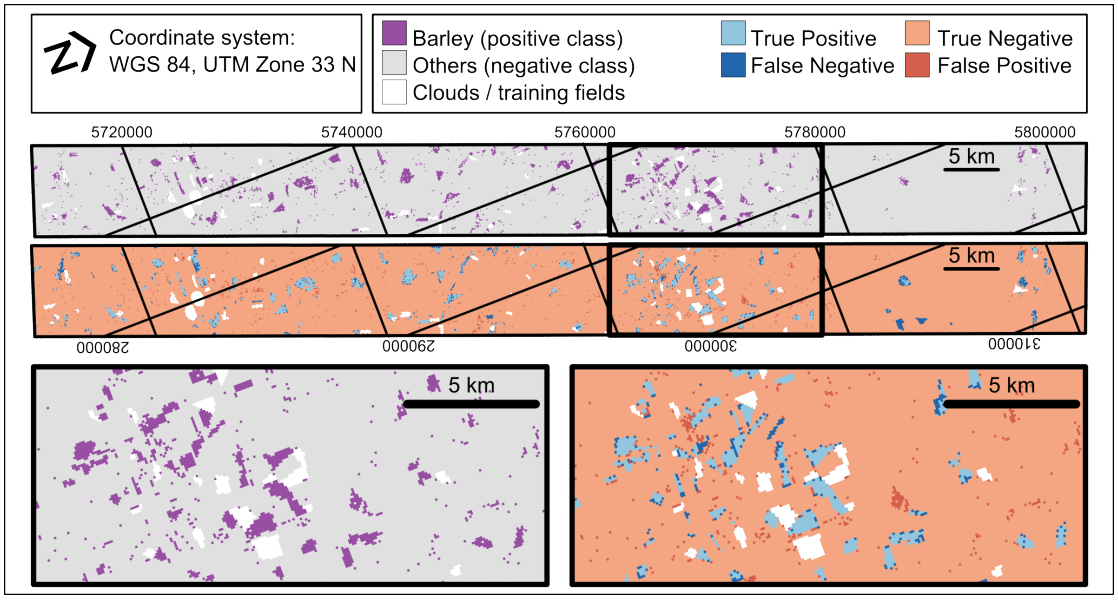

FIGURE III.7: Classification and test errors for the class barley realized with the manually selected model and the threshold $\hat{\theta}^{\mathrm{MAP}}$ (see Figure III.6 and Table III.5).

6 Conclusions

In the presented study, a novel strategy for solving the problem of one-class classification was proposed, tested in experiments, and discussed in the context of classifying hyperspectral data. Although various approaches have been introduced, the generation of accurate maps by one-class classifiers is challenging, due to the incomplete and unrepresentative reference data. As a matter of fact the model and threshold selection, cannot be solved based on traditional accuracy metrics, such as the overall accuracy or the kappa coefficient. Thus, the classification does not necessarily lead to optimal results.

The novelty and potential of the presented strategy lies in the analysis of the one-dimensional output of any one-class classifier. Based on our experiments, it can be assessed that the proposed framework for analyzing and interpreting the classifier outputs can reveal poor model and/or threshold selection results. A proposed diagnostic plot for one-class classification results supports the user in understanding the quality of a given one-class classification result and enables the user to manually select more accurate solutions, whether an automatic procedures failed. Furthermore, it has been shown that reliable a posteriori probabilities with small positive training sets can be derive in the one-dimensional output space of any one-class classifier. Overall, due to the proposed strategy, the use of state-of-the-art OCC can be advanced and the increased requirements for effective remote sensing image analysis of recent data may be easier fulfilled.

Future work should extend the strategy to the more general partially supervised classification problem, *i.e.*, when more than one classes have to be mapped.

The implementations described in this paper have been implemented in the *R* software and are partially available in the package *oneClass*. The package is available via github (Mack, 2014) and can be installed directly from within *R*.

Acknowledgments

The study is realized in the framework of the EnMAP-BMP project funded by German Aeropspace Center (DLR) and Federal Ministry of Economics and Technology (BMWi) (DLR/BMWi: FKZ 50EE 1011). Reference data was made available by the Ministry of Agriculture and Environment, Saxony Anhalt, Germany. We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Karlsruhe Institute of Technology.

References

Amici, V. (2011). "Dealing with vagueness in complex forest landscapes: A soft classification approach through a niche-based distribution model". In: *Ecological Informatics* 6.6, pp. 371–383.

Azzalini, A. and G. Menardi (2014). "Clustering via Nonparametric Density Estimation: The R Package pdfCluster". In: *Journal of Statistical Software* 57.11.

Bovolo, F., G. Camps-Valls, and L. Bruzzone (2010). "A support vector domain method for change detection in multitemporal images". In: *Pattern Recognition Letters* 31.10, pp. 1148–1154.

Briem, G., J. Benediktsson, and J. Sveinsson (2002). "Multiple classifiers applied to multisource remote sensing data". In: *IEEE Trans. Geosci. Remote Sensing* 40.10, pp. 2291–2299.

Bruzzone, L. (2000). "An approach to feature selection and classification of remote sensing images based on the Bayes rule for minimum cost". In: *IEEE Trans. Geosci. Remote Sensing* 38.1, pp. 429–438.

Christophe, E. and J. Inglada (2009). "Open source remote sensing: Increasing the usability of cutting-edge algorithms". In: *IEEE Geoscience and Remote Sensing Newsletter* 35.5, pp. 9–15.

Désir, C., S. Bernard, C. Petitjean, and L. Heutte (2013). "One class random forests". In: *Pattern Recognition* 46.12, pp. 3490–3506.

Drake, J. M. (2014). "Ensemble algorithms for ecological niche modeling from presence-background and presence-only data". In: *Ecosphere* 5.6, art76.

Du, P., J. Xia, W. Zhang, K. Tan, Y. Liu, and S. Liu (2012). "Multiple Classifier System for Remote Sensing Image Classification: A Review". In: *Sensors* 12.12, pp. 4764–4792.

Dubuisson, B. and M. Masson (1993). "A statistical decision rule with incomplete knowledge about classes". In: *Pattern Recognition* 26.1, pp. 155–165.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates (2010). "A statistical explanation of MaxEnt for ecologists". In: *Diversity and Distributions* 17.1, pp. 43–57.

Elkan, C. and K. Noto (2008). "Learning classifiers from only positive and unlabeled data". In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. Association for Computing Machinery (ACM).

European Union (2013). "COMMISSION DELEGATED REGULATION (EU) No 1159/2013 of 12 July 2013". In: *Official Journal of the European Union* 56.L 309, pp. 1–6.

Evangelista, P. H., T. J. Stohlgren, J. T. Morisette, and S. Kumar (2009). "Mapping Invasive Tamarisk (Tamarix): A Comparison of Single-Scene and Time-Series Analyses of Remotely Sensed Data". In: *Remote Sensing* 1.3, pp. 519–533.

Fernández-Prieto, D. (2002). "An iterative approach to partially supervised classification problems". In: *International Journal of Remote Sensing* 23.18, pp. 3887–3892.

Fernandez-Prieto, D. and M. Marconcini (2011). "A Novel Partially Supervised Approach to Targeted Change Detection". In: *IEEE Trans. Geosci. Remote Sensing* 49.12, pp. 5016–5038.

Foody, G. M. (2004). "Thematic Map Comparison". In: *Photogrammetric Engineering & Remote Sensing* 70.5, pp. 627–633.

Foody, G. M., A. Mathur, C. Sanchez-Hernandez, and D. S. Boyd (2006). "Training set size requirements for the classification of a specific class". In: *Remote Sensing of Environment* 104.1, pp. 1–14.

Fumera, G., F. Roli, and G. Giacinto (2000). "Multiple Reject Thresholds for Improving Classification Reliability". In: *Advances in Pattern Recognition: Joint IAPR International Workshops SSPR 2000 and SPR 2000 Alicante, Spain, August 30 – September 1, 2000 Proceedings*. Ed. by F. J. Ferri, J. M. Iñesta, A. Amin, and P. Pudil. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 863–871.

Guerrero-Curieses, A., A. Biasiotto, S. Serpico, and G. Moser (2002). "Supervised classification of remote sensing images with unknown classes". In: *IEEE International Geoscience and Remote Sensing Symposium*. Institute of Electrical & Electronics Engineers (IEEE).

Guo, Q., W. Li, D. Liu, and J. Chen (2012). "A Framework for Supervised Image Classification with Incomplete Training Samples". In: *Photogrammetric Engineering & Remote Sensing* 78.6, pp. 595–604.

Hughes, G. (1968). "On the mean accuracy of statistical pattern recognizers". In: *IEEE Transactions on Information Theory* 14.1, pp. 55–63.

Inglada, J. and E. Christophe (2009). "The Orfeo Toolbox remote sensing image processing software". In: *2009 IEEE International Geoscience and Remote Sensing Symposium*. Institute of Electrical and Electronics Engineers (IEEE).

Jeon, B. and D. Landgrebe (1990). "A New Supervised Absolute Classifier". In: *10th Annual International Symposium on Geoscience and Remote Sensing*. Institute of Electrical & Electronics Engineers (IEEE).

– (1999). "Partially supervised classification using weighted unsupervised clustering". In: *IEEE Trans. Geosci. Remote Sensing* 37.2, pp. 1073–1079.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004). "kernlab - An S4 Package for Kernel Methods in R". In: *Journal of Statistical Software* 11.9.

Krawczyk, B., M. Woźniak, and B. Cyganek (2014). "Clustering-based ensembles for one-class classification". In: *Information Sciences* 264, pp. 182–195.

Li, P. and H. Xu (2010). "Land-Cover Change Detection using One-Class Support Vector Machine". In: *Photogrammetric Engineering & Remote Sensing* 76.3, pp. 255–263.

Li, W. and Q. Guo (2010). "A maximum entropy approach to one-class classification of remote sensing imagery". In: *International Journal of Remote Sensing* 31.8, pp. 2227–2235.

Li, W., Q. Guo, and C. Elkan (2011). "A Positive and Unlabeled Learning Algorithm for One-Class Classification of Remote-Sensing Data". In: *IEEE Transactions on Geoscience and Remote Sensing* 49.2, pp. 717–725.

Li, X. and B. Liu (2003). "Learning to classify texts using positive and unlabeled data". In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. IJCAI'03. Acapulco, Mexico: Morgan Kaufmann Publishers Inc., pp. 587–592.

Lin, G. C. and T. C. Minter (1976). "Bayes estimation on parameters of the single-class classifier". In: *Proceedings of Symposium on Machine Processing of Remotely Sensed Data*. 3A–22–3A–27. West Lafayette, IN, USA.

Liu, B., Y. Dai, X. Li, W. L. Lee, and P. S. Yu (2003). "Building text classifiers using positive and unlabeled examples". In: *Third IEEE International Conference on Data Mining*. IEEE Computer Society, pp. 179–186.

Liu, C., M. White, and G. Newell (2013). "Selecting thresholds for the prediction of species occurrence with presence-only data". In: *Journal of Biogeography* 40.4. Ed. by R. Pearson, pp. 778–789.

Mack, B. (2014). *oneClass: One-class classification in the absence of test data*. R package version 0.1-1.

Malenovský, Z., H. Rott, J. Cihlar, M. E. Schaepman, G. García-Santos, R. Fernandes, and M. Berger (2012). "Sentinels for science: Potential of Sentinel-1, -2, and -3 missions for scientific observations of ocean, cryosphere, and land". In: *Remote Sensing of Environment* 120, pp. 91–101.

Mantero, P., G. Moser, and S. Serpico (2005). "Partially Supervised classification of remote sensing images through SVM-based probability density estimation". In: *IEEE Trans. Geosci. Remote Sensing* 43.3, pp. 559–570.

Marconcini, M., D. Fernandez-Prieto, and T. Buchholz (2014). "Targeted Land-Cover Classification". In: *IEEE Trans. Geosci. Remote Sensing* 52.7, pp. 4173–4193.

Minter, T. C. (1975). "Single-Class Classification". In: *Proceedings of Symposium on Machine Processing of Remotely Sensed Data*. 2A-12–2A-15. West Lafayette, IN.

Morán-Ordóñez, A., S. Suárez-Seoane, J. Elith, L. Calvo, and E. de Luis (2012). "Satellite surface reflectance improves habitat distribution mapping: a case study on heath and shrub formations in the Cantabrian Mountains (NW Spain)". In: *Diversity and Distributions* 18.6, pp. 588–602.

Moser, G. and S. B. Serpico (2013). "Combining Support Vector Machines and Markov Random Fields in an Integrated Framework for Contextual Image Classification". In: *IEEE Trans. Geosci. Remote Sensing* 51.5, pp. 2734–2752.

Muñoz-Marí, J., F. Bovolo, L. Gómez-Chova, L. Bruzzone, and G. Camp-Valls (2010). "Semisupervised One-Class Support Vector Machines for Classification of Remote Sensing Data". In: *IEEE Transactions on Geoscience and Remote Sensing* 48.8, pp. 3188–3197.

Munoz-Mari, J., G. Camps-Valls, L. Gomez-Chova, and J. Calpe-Maravilla (2007). "Combination of one-class remote sensing image classifiers". In: *2007 IEEE International Geoscience and Remote Sensing Symposium*. Institute of Electrical & Electronics Engineers (IEEE).

Muzzolini, R., Y.-H. Yang, and R. Pierson (1998). "Classifier design with incomplete knowledge". In: *Pattern Recognition* 31.4, pp. 345–369.

Ortiz, S., J. Breidenbach, and G. Kändler (2013). "Early Detection of Bark Beetle Green Attack Using TerraSAR-X and RapidEye Data". In: *Remote Sensing* 5.4, pp. 1912–1931.

Phillips, S. J., R. P. Anderson, and R. E. Schapire (2006). "Maximum entropy modeling of species geographic distributions". In: *Ecological Modelling* 190.3-4, pp. 231–259.

Phillips, S. J. and M. Dudík (2008). "Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation". In: *Ecography* 31.2, pp. 161–175.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Rabe, A., B. Jakimow, M. Held, van der Linden, S., and P. Hostert (2014). *EnMAP-Box, Version 2.0: software available at www.enmap.org.*

Richards, J. (2005). "Analysis of remotely sensed data: the formative decades and the future". In: *IEEE Trans. Geosci. Remote Sensing* 43.3, pp. 422–432.

Roscher, R., B. Waske, and W. Forstner (2012). "Incremental Import Vector Machines for Classifying Hyperspectral Data". In: *IEEE Trans. Geosci. Remote Sensing* 50.9, pp. 3463–3473.

Roy, D., M. Wulder, T. Loveland, W. C.E., R. Allen, M. Anderson, D. Helder, J. Irons, D. Johnson, R. Kennedy, T. Scambos, C. Schaaf, J. Schott, Y. Sheng, E. Vermote, A. Belward, R. Bindschadler, W. Cohen, F. Gao, J. Hipple, P. Hostert, J. Huntington, C. Justice, A. Kilic, V. Kovalskyy, Z. Lee, L. Lymburner, J. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R. Wynne, and Z. Zhu (2014). "Landsat-8: Science and product vision for terrestrial global change research". In: *Remote Sensing of Environment* 145, pp. 154–172.

Russell G. Congalton, K. G. (2008). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC PR INC. 200 pp.

Sánchez-Azofeifa, A., B. Rivard, J. Wright, J.-L. Feng, P. Li, M. M. Chong, and S. A. Bohlman (2011). "Estimation of the Distribution of Tabebuia guayacan (Bignoniaceae) Using High-Resolution Remote Sensing Imagery". In: *Sensors* 11.12, pp. 3831–3851.

Sanchez-Hernandez, C., D. S. Boyd, and G. M. Foody (2007). "One-Class Classification for Mapping a Specific Land-Cover Class: SVDD Classification of Fenland". In: *IEEE Trans. Geosci. Remote Sensing* 45.4, pp. 1061–1073.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson (2001). "Estimating the Support of a High-Dimensional Distribution". In: *Neural Computation* 13.7, pp. 1443–1471.

Shahshahani, B. and D. Landgrebe (1994). "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon". In: *IEEE Trans. Geosci. Remote Sensing* 32.5, pp. 1087–1095.

Stohlgren, T. J., P. Ma, S. Kumar, M. Rocca, J. T. Morisette, C. S. Jarnevich, and N. Benson (2010). "Ensemble Habitat Mapping of Invasive Plant Species". In: *Risk Analysis* 30.2, pp. 224–235.

Stuffler, T., K. Förster, S. Hofer, M. Leipold, B. Sang, H. Kaufmann, B. Penné, A. Mueller, and C. Chlebek (2009). "Hyperspectral imaging—An advanced instrument concept for the EnMAP mission (Environmental Mapping and Analysis Programme)". In: *Acta Astronautica* 65.7-8, pp. 1107–1112.

Tax, D. M. J. (2001). "One-class classification - Concept-learning in the absence of counter-examples". PhD thesis. Delft University of Technology.

Trevor Hastie Robert Tibshirani, J. F. (2009). *The Elements of Statistical Learning*. Springer-Verlag New York Inc. 767 pp.

Waske, B. and M. Braun (2009). "Classifier ensembles for land cover mapping using multitemporal SAR imagery". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 64.5, pp. 450–457.

Chapter IV

Mapping raised bogs with an iterative one-class classification approach

ISPRS Journal of Photogrammetry and Remote Sensing, vol. 120, pp. 53-64

Benjamin Mack, Ribana Roscher, Stefanie Stenzel, Hannes Feilhauer, Sebastian Schmidtlein and Björn Waske

https://doi.org/10.1016/j.isprsjprs.2016.07.008

Chapter V

Synthesis

1 Findings

In this thesis one-class classification, an advanced learning paradigm for classifying remote sensing data, has been investigated from the practitioner's perspective. The main objective was to enhance the usability of one-class classification methods for users with limited background in fields such as computing methodologies, machine learning, pattern recognition and statistical learning. With one-class classification methods it is possible to map a specific class of interest (or positive class) by training a model with labeled samples of this class but without the need of a representative set of counter-class (or negative class) labeled sample. Obviously, this is particularly interesting when reference data collection is expensive which is often the case.

In Chapter I it has been shown theoretically that OCC models based on positive labeled and unlabeled data (PU-learner) can be as good as fully supervised binary classifiers. Instead, the potential performance of models based on positive data only (P-learner) is lower for typical classification problems. This has been confirmed empirically in Chapter II based on a variety of classification problems. On average over all these experiments the potential performance of the biased SVM (a PU-learner) has been shown to be similar to a fully supervised binary SVM while OCSVM (a P-learner) performed significantly poorer. The potential performance of MaxEnt (a PU-learner) was of special interest since it is one of the most frequently used approaches in applied OCC studies. It performed better than OCSVM but not as good as the biased SVM. However, the potential performance of the base algorithms could not be achieved by any of the various parameter and threshold selection combinations that have been investigated for each of the three base classification algorithms. This shows how critical model (parameter and threshold) selection is in OCC applications and how unreliable fully-automatic approaches are.

In Chapter III a user-oriented strategy for OCC has been presented which is based on the visualization and interpretation of the outcome of any one-class classifier. A diagnostic plot has been introduced which fosters the understanding of the result in terms of class separability and the suitability of a particular threshold. Comparing such diagnostic plots allows to identify poor models and improve the results by manually selecting more suitable parameter and/or threshold settings. The informativeness of the diagnostic plot and the suitability of the strategy for improving poor automatic selections has been demonstrated based on exemplary one-class classification problems. The tools for the presented analysis strategy together with an interface to the one-class classifiers OCSVM, biased SVM and MaxEnt has been implemented in the open-source R package *oneClass* (Mack, 2017).

Finally, in Chapter IV the focus was on a particularly challenging application of one-class classification. First, the number of labeled training samples for the class of interest (raised bogs) was very small (31). Second, the class of interest was rare in the study site ($< 1\,\%$ of the mapped area) leading to the complicating situation of imbalanced classes and eventually small negative disjuncts overlapping with the class of interest (see Chapter I Section 2.4). In the study an algorithm has been proposed which is particularly designed for such situations. In an iterative pre-classification step, easy to classify negatives are classified and removed from subsequent analysis. As a consequence, the dataset remaining in the final classification stage is much more balanced which allows for the application of parameter and threshold selection approaches that would not be possible otherwise. In the presented study a joint parameter and threshold selection approach based on a normal mixture model has been developed which automatically selected an accurate and suitable model. As a useful by-product it provides an extension of the diagnostic plots developed in Chapter III which helps the user to better rate the outcome of the automatic algorithm in the absence of complete and representative test data.

2 Conclusions and Recommendations

OCC algorithms potentially perform as good as fully supervised binary classifiers. However, in order to realize the full potential of OCC, it is in many cases necessary to use OCC algorithms that are considered more difficult to handle from the user's perspective[1]. Thus, it is not a good idea to blindly rely on fully-automatic model selection algorithms but instead critically analyze the classification outcomes. The strategy and tools developed in this thesis are useful in this respect.

It is also worth stressing the informativeness of reporting the potential performance of a base classifier together with the performance realized by a model selection approach. This information is important in order to better understand the different components of a one-class classifier. In contrast to most, if not all, comparative studies, this information has been included in the in-depth comparison presented in Chapter II. The crucial information derived is the high performance loss due to any of the parameter and threshold selection approaches. Without this information a user might conclude: "Biased SVM with the best parameter and threshold selection approach performs only slightly better than MaxEnt with default parameters and a suitable threshold selection approach. I accept this and spare myself the time to tackle the biased SVM since it

[1] "While for Maxent and BRT, the technique is straightforward, the biased SVM approach required more effort in model selection and parameter tuning, as compared to the other classifiers" (Skowronek, Asner, and Feilhauer, 2017).

is new to me." With this information and being confident in manually analyzing and eventually correcting poor automatic parameters and thresholds selections the conclusion might be: "On average the potential accuracy of biased SVM is substantially higher than the accuracy of MaxEnt. I will take my time to handle this method even though it is new to me."

Sometimes the conclusions of a study might also be misleading without this information. For example, (Chen et al., 2016) investigate the performance of the PUL-algorithm (Elkan and Noto, 2008) dependent on the number of unlabeled training samples. They observe a performance drop when the number of unlabeled training samples increases with respect to the positive labeled training samples and conclude that "a balanced positive and unlabelled sample size is recommended when PUL-SVM is used" (Chen et al., 2016 p.1070). Note that the threshold selection implemented in the PUL-algorithm is a crucial part of the implementation. A natural question is if the performance drop is due to a poor threshold selection or if any potential threshold perform poor. This question can easily be answered by reporting the potential performance. But why is this important to know in order to increase the informativeness and relevance of the study? First, recall the problem of imbalanced data sets and potential overlap with small negative disjuncts (see Section 2.4). It has been argued that in such a case it can be important to sample a large amount of unlabeled training samples for building an accurate model. In such a case, it is possible that the potential accuracy increases with an increasing number of unlabeled training samples even though the accuracy of the classification approach based on a specific threshold selection drops. As a consequence, the conclusion should be that – in case the class of interest is rare in the study area, it might require a large number of unlabeled training samples and an adjustment of the default threshold. Instead, if also the potential accuracy drops with a large number of unlabeled samples the conclusion would be that the PUL-algorithm – as it is implemented in this study[2] – is eventually not suitable for an application where the classes are imbalanced. In the comparative analysis presented by (Chen et al., 2016) the image size is very small (400×400 pixels) and the class distribution is fairly balanced. In many real-world applications this is not the case and it is important to discuss the problem of imbalanced data when making recommendations about the usage of unlabeled samples based on such ideal datasets. The potential performance facilitates such a critical discussion.

Note that in the field of machine learning from imbalanced data (Provost, 2000) stresses

[2] In this study the implementation is based on a SVM using only the default parameter settings of the SVM in ENVI 4.8 which is also questionable since it is known that the accuracy of the SVM is sensitive to the parameter setting.

the large number of studies where up- or down-sampling does or does not solve the imbalance problem. However, "in most of these studies, it never even was asked whether simply setting the output threshold correctly would be sufficient; without doing so the research may be misleading" (Provost, 2000 p. 2). As in the case of OCC studies, reporting the best potential accuracy is not only relevant for users to make more informed decisions when selecting candidate classifiers for an application. It also helps developers to prioritize research that best advances the field of OCC in remote sensing. It is therefore strongly recommended that studies in the field of OCC, particularly when introducing new methods, include the potential performance of the investigated base algorithm(s).

3 Prospect

The usefulness of the user-guided model selection based on the developed diagnostic plot has been shown. Still, there is a high potential to build better analytic tools for solving real-world OCC problems. Certainly, OCC methods are more likely to be used in real-world applications when a powerful and user-friendly system for human-guided model selection is available. Its design should be a research focus and priority of the developer and user community and – from the user's perspective – is more relevant than yet another specific OCC algorithm. For example, (Provost and Fawcett, 2001) proposed a robust classification strategy for imprecise environments. According to the authors, the strategy is "efficient and incremental, minimizes the management of classifier performance data, and allows for clear visual comparisons and sensitivity analyses." The system is based on the receiver operating characteristic curve (ROC) and since it is valid to build the ROC with PU-data and rank the predictive power of different classifiers for a given false negative rate (Phillips, Anderson, and Schapire, 2006) it should be possible to adapt this strategy for OCC. The diagnostic plot introduced in Chapter III is complementary to this strategy and together the two approaches can lead to an improved user oriented system for an efficient user-guided selection of a suitable solution from a large amount of potential one-class classification models.

One of the most interesting recent trends in remote sensing is large area mapping at high resolution (10m–30m) (Ozdogan, 2015; Gómez, White, and Wulder, 2016) which is possible thanks to freely available Landsat, Sentinel-2 and Sentinel-1 data. This opens the door to new remote sensing based mapping applications, many of them focusing on one or a few specific classes. State-of-the-art approaches rely on spatially contiguous best-available-pixel composites (Griffiths et al., 2013; White et al., 2014), spectral-temporal metrics (Potapov et al., 2012) and/or constructed dense time-series (Hermosilla et al.,

2016; Vuolo, Ng, and Atzberger, 2017). The creation of training datasets required for successfully classifying such features over large areas can be very demanding.In general, the class distributions in such feature spaces are much more complex compared to those in datasets of smaller areas and full-scene features. This is a consequence of the higher variability of the manifestation class of interest on the ground. Furthermore, the features are less homogeneous compared to the full-scene features mainly dependent on the number and acquisition time of the valid observations they are derived from. As a consequence, the classification accuracy can decrease critically when the distance between the training data and the classified data increases (Pelletier et al., 2016).

It is worth to distinguish two typical reference data scenarios for large area mapping: On the one hand, existing large area covering in-situ reference databases (Mack et al., 2016) or existing (out-dated and/or coarse resolution) maps (Zhu et al., 2016) exist and can be used to derive the training dataset for the class of interest and the counter-class. In this case, OCC is probably not a good choice but instead the focus should be on selecting a suitable number of samples from the high amount of possible ones (Zhu et al., 2016) and/or on cleaning the dataset from label noise. In another case, no auxiliary reference dataset is available that can be used, e.g. because the class of interest is a new phenomena in the area (e.g. invasive species) or the spatial location of the class is not persistent over time (e.g. annual crops) which makes out-dated products uninformative for the current situation. In such a situation OCC is obviously attractive since it spares the user from generating a representative negative training set. However, due to the above mentioned reason, even the generation of a representative positive class training set is challenging for large areas. Active learning (Tuia, Pasolli, and Emery, 2011) is another advanced machine learning paradigm that is worth implementing in combination with OCC. The idea of active learning is that the learning algorithm starts with a small reference data set to build a model. Then a query algorithm selects the most informative sample(s) to be labeled by an expert. The newly labeled samples can then be used in order to improve the model. These steps are repeated until a suitable stop criteria is reached. Combining active learning and OCC is a particularly interesting research direction which might critically reduce the cost for single-class mapping over large areas. However, substantial adaptation is required to combine the two learning paradigms and make them suitable for the large datasets to be processed in large area applications. One strategy might be to first derive a representative positive training set using active learning and a P-classifier (e.g. similar to Furlani et al., 2012). Next, the derived P-dataset can be used to derive a subset of the whole data such that the true positive rate is high while a considerable false positive rate can be accepted (e.g. similar to the pre-classifier in Chapter IV). Then the final classification can be solved on the subset by any suitable one-class classification approach or by building a binary classifier

based on active learning. It needs to be investigated how the individual core elements as well as the overall processing workflow need to be designed in order to cope with the immense computational cost of large area datasets.

References

Chen, X., D. Yin, J. Chen, and X. Cao (2016). "Effect of training strategy for positive and unlabelled learning classification: test on Landsat imagery". In: *Remote Sensing Letters* 7.11, pp. 1063–1072.

Elkan, C. and K. Noto (2008). "Learning classifiers from only positive and unlabeled data". In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*. Association for Computing Machinery (ACM).

Furlani, M., D. Tuia, J. Munoz-Mari, F. Bovolo, G. Camps-Valls, and L. Bruzzone (2012). "Discovering single classes in remote sensing images with active learning". In: *2012 IEEE International Geoscience and Remote Sensing Symposium*. Institute of Electrical and Electronics Engineers (IEEE).

Gómez, C., J. C. White, and M. A. Wulder (2016). "Optical remotely sensed time series data for land cover classification: A review". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 116, pp. 55–72.

Griffiths, P., S. van der Linden, T. Kuemmerle, and P. Hostert (2013). "A Pixel-Based Landsat Compositing Algorithm for Large Area Land Cover Mapping". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6.5, pp. 2088–2101.

Hermosilla, T., M. A. Wulder, J. C. White, N. C. Coops, G. W. Hobart, and L. B. Campbell (2016). "Mass data processing of time series Landsat imagery: pixels to data products for forest monitoring". In: *International Journal of Digital Earth* 9.11, pp. 1035–1054.

Mack, B. (2017). *oneClass: one-class classification in the absence of test data*. https://github.com/benmack/oneClass.

Mack, B., P. Leinenkugel, C. Kuenzer, and S. Dech (2016). "A semi-automated approach for the generation of a new land use and land cover product for Germany based on Landsat time-series and Lucas in-situ data". In: *Remote Sensing Letters* 8.3, pp. 244–253.

Ozdogan, M. (2015). "Image Classification Methods in Land Cover and Land Use". In: *Remotely Sensed Data Characterization, Classification, and Accuracies*. Ed. by P. S. Thenkabail. CRC Press. Chap. 11, pp. 231–258.

Pelletier, C., S. Valero, J. Inglada, N. Champion, and G. Dedieu (2016). "Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas". In: *Remote Sensing of Environment* 187, pp. 156–168.

Phillips, S. J., R. P. Anderson, and R. E. Schapire (2006). "Maximum entropy modeling of species geographic distributions". In: *Ecological Modelling* 190.3-4, pp. 231–259.

Potapov, P. V., S. A. Turubanova, M. C. Hansen, B. Adusei, M. Broich, A. Altstatt, L. Mane, and C. O. Justice (2012). "Quantifying forest cover loss in Democratic Republic of the Congo, 2000–2010, with Landsat ETM+ data". In: *Remote Sensing of Environment* 122, pp. 106–116.

Provost, F. (2000). "Machine learning from imbalanced data sets 101". In: *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, pp. 1–3.

Provost, F. and T. Fawcett (2001). "Robust Classification for Imprecise Environments". In: *Machine Learning* 42.3, pp. 203–231.

Skowronek, S., G. P. Asner, and H. Feilhauer (2017). "Performance of one-class classifiers for invasive species mapping using airborne imaging spectroscopy". In: *Ecological Informatics* 37, pp. 66–76.

Tuia, D., E. Pasolli, and W. Emery (2011). "Using active learning to adapt remote sensing image classifiers". In: *Remote Sensing of Environment* 115.9, pp. 2232–2242.

Vuolo, F., W.-T. Ng, and C. Atzberger (2017). "Smoothing and gap-filling of high resolution multi-spectral time series: Example of Landsat data". In: *International Journal of Applied Earth Observation and Geoinformation* 57, pp. 202–213.

White, J. C., M. A. Wulder, G. W. Hobart, J. E. Luther, T. Hermosilla, P. Griffiths, N. C. Coops, R. J. Hall, P. Hostert, A. Dyk, and L. Guindon (2014). "Pixel-Based Image Compositing for Large-Area Dense Time Series Applications and Science". In: *Canadian Journal of Remote Sensing* 40.3, pp. 192–212.

Zhu, Z., A. L. Gallant, C. E. Woodcock, B. Pengra, P. Olofsson, T. R. Loveland, S. Jin, D. Dahal, L. Yang, and R. F. Auch (2016). "Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 122, pp. 206–221.