Int J Appl Earth Obs Geoinformation 60 (2017) 83–98

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

journal homepage: www.elsevier.com/locate/jag

Class imbalance in unsupervised change detection – A diagnostic analysis from urban remote sensing

Tobias Leichtle[a,b,c,*], Christian Geiß[b], Tobia Lakes[c], Hannes Taubenböck[b]

[a] Company for Remote Sensing and Environmental Research (SLU), Kohlsteiner Straße 5, 81243 München, Germany
[b] German Aerospace Center(DLR), German Remote Sensing Data Center (DFD), Münchner Straße 20, 82234 Weßling, Germany
[c] Humboldt-Universität zu Berlin, Geography Department, Rudower Chaussee 16, 12489 Berlin, Germany

ARTICLE INFO

Keywords:
Change detection
Object-based image analysis (OBIA)
Very-high resolution (VHR) remote sensing
Class imbalance
Clustering
Urban environment

ABSTRACT

Automatic monitoring of changes on the Earth's surface is an intrinsic capability and simultaneously a persistent methodological challenge in remote sensing, especially regarding imagery with very-high spatial resolution (VHR) and complex urban environments. In order to enable a high level of automatization, the change detection problem is solved in an unsupervised way to alleviate efforts associated with collection of properly encoded prior knowledge. In this context, this paper systematically investigates the nature and effects of class distribution and class imbalance in an unsupervised binary change detection application based on VHR imagery over urban areas. For this purpose, a diagnostic framework for sensitivity analysis of a large range of possible degrees of class imbalance is presented, which is of particular importance with respect to unsupervised approaches where the content of images and thus the occurrence and the distribution of classes are generally unknown a priori. Furthermore, this framework can serve as a general technique to evaluate model transferability in any two-class classification problem. The applied change detection approach is based on object-based difference features calculated from VHR imagery and subsequent unsupervised two-class clustering using k-means, genetic k-means and self-organizing map (SOM) clustering. The results from two test sites with different structural characteristics of the built environment demonstrated that classification performance is generally worse in imbalanced class distribution settings while best results were reached in balanced or close to balanced situations. Regarding suitable accuracy measures for evaluating model performance in imbalanced settings, this study revealed that the Kappa statistics show significant response to class distribution while the true skill statistic was widely insensitive to imbalanced classes. In general, the genetic k-means clustering algorithm achieved the most robust results with respect to class imbalance while the SOM clustering exhibited a distinct optimization towards a balanced distribution of classes.

1. Introduction

Change detection is one of the most intrinsic capabilities of remote sensing due to its system-inherent repetitive character of image acquisition (Singh, 1989). Especially with respect to recently available satellite-based multispectral images with very-high spatial resolution (VHR), change detection remains a challenging task and an active field of research (Bruzzone and Bovolo, 2013). With increasing level of detail in VHR images, traditional pixel-based methods for change detection become less effective and techniques from object-based image analysis (OBIA) are utilized more frequently (Hussain et al., 2013). These methods allow for characterization and associated detection of changes of entire image objects which are directly related to meaningful real-world objects. OBIA techniques are beneficial in particular over urban environments, where VHR images comprise a wealth of detail due to the large spatial heterogeneity of (mostly man-made) objects. Changes within the complex urban environment can be of various types (e.g., construction of buildings, setup of infrastructure, reconstruction of buildings, etc.), whereas the most basic use case of change detection is the dissociation of changed and unchanged areas (Ridd and Liu, 1998). This two-class discrimination (i.e., binary classification of changes) is the focus in most applications of unsupervised change detection in remote sensing (Bruzzone and Bovolo, 2013). Dependent on the spatial resolution of the data, object-based binary change detection applications range from classification of urban and non-urban land cover based on medium and high resolution remote sensing images (e.g., Taubenb & ck et al. (2012)) to discrimination of changed and unchanged buildings using VHR imagery (e.g., Tian et al. (2014)). In

* Corresponding author at: Company for Remote Sensing and Environmental Research (SLU), Kohlsteiner Straße 5, 81243 München, Germany.
E-mail address: tobias.leichtle@slu-web.de (T. Leichtle).

http://dx.doi.org/10.1016/j.jag.2017.04.002
Received 23 December 2016; Received in revised form 31 March 2017; Accepted 11 April 2017
Available online 30 April 2017
0303-2434/

particular, unsupervised two-class change detection approaches focusing on buildings in VHR remote sensing images have become a vital research field (e.g., Huang et al. (2014), Wang et al. (2015)). With persistently increasing availability of VHR remote sensing imagery, automated concepts for change detection are highly required (Bruzzone and Bovolo, 2013). In order to achieve high automation, unsupervised methods for change detection are preferred over supervised methods since they do not require any prior knowledge on changes between the images (Hussain et al., 2013). However, for unsupervised techniques the distribution of classes is unknown a priori and can be heavily imbalanced (e.g., the unchanged class exhibits dominantly more samples compared to the changed class in numerous applications).

From a generalized perspective, this phenomenon is referred to as the class imbalance problem (Japkowicz and Stephen, 2002). The imbalanced distribution of classes deteriorates the accuracy of most standard learning and classification methods, which assume a balanced distribution of classes (Japkowicz and Stephen, 2002). Learning from imbalanced data is a recent and highly active field of research in machine learning (Chawla et al., 2004; López et al., 2013). Strictly speaking, any data set that exhibits an unequal distribution of classes (i.e., in case of two classes any deviation from 50:50) may be considered as imbalanced, whereas the common definition of between-class imbalances (minority class against majority class) is in the order of 100:1 or beyond (Haibo and Garcia, 2009). Class imbalance is an intrinsic problem in many classification approaches across a large field of applications (Chawla et al., 2004). Nevertheless, from a remote sensing perspective there exist only few studies that address this phenomenon in a supervised way: A supervised neural network learning method is used in Bruzzone and Serpico (1997) for classification of agricultural land with an imbalanced distribution of classes. Kubat et al. (1998) investigate the application of oil spill detection from radar images, which is treated as an imbalanced two-class supervised machine learning task of discriminating oil slicks (minority class) against the sea surface (majority class). The authors of Williams et al. (2009) employ a modified logistic regression approach for supervised mine classification from remotely sensed data, while García et al. (2011) present an imbalanced multi-class setting using supervised classification on hyperspectral remote sensing images. Thus, no studies were carried out based on remote sensing in an unsupervised context concerning class imbalance.

In general, there are several solutions for the class imbalance problem available in literature, which generally rely on a supervised classification strategy (López et al., 2013). According to Haibo and Garcia (2009), a widely deployed group of solutions for imbalanced learning comprise sampling-based methods. These methods aim at balancing class proportions in the training data set prior to classification. Applicable strategies are, among others, over- and undersampling of the minority and majority classes, respectively, e.g., the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) and its successors. Another important group of supervised solutions to the class imbalance problem are cost-sensitive learning methods. The idea behind these techniques is to develop a hypothesis that minimizes the overall costs of misclassifications in the training data set, whereas the costs of misclassifying a minority example labeled as the majority class is higher than the contrary case (Elkan, 2001). Examples for these methods comprise cost-sensitive boosting methods (Sun et al., 2007), cost-sensitive decision trees, or cost-sensitive neural networks (Kukar and Kononenko, 1998). Furthermore, kernel-based methods can be employed as supervised solutions for the imbalanced learning problem, where the transformation to a higher-dimensional feature space enables proper discrimination of the imbalanced data set (Haibo and Garcia, 2009). In addition, only relevant training samples are utilized to define the model, which may thus not be affected by imbalanced data. Examples of this technique comprise over- and undersampled support vector machines (SVMs) (Akbani et al., 2004), kernel modification methods (Wu and Chang, 2005) or SVMs that include active learning strategies for efficient selection of relevant training samples (Ertekin et al., 2007). Finally, one-class learning or novelty detection methods aim at classifying only a single class (i.e., recognition-based approaches) (Pimentel et al., 2014). Examples of this approach are one-class SVMs (Lee and Cho, 2006) or the autoassociator method which employs neural networks for one-class classification (Japkowicz, 2001).

In contrast to these generally supervised solutions to imbalanced learning, this paper presents the class imbalance problem in an unsupervised context based on the example of urban remote sensing and change detection. In detail, the application of binary change detection of buildings using VHR remote sensing imagery is presented as an example since buildings are the most relevant and one of the most dynamic objects in complex urban environments (Huang et al., 2014). The objectives of this paper are i) presentation of the class imbalance problem in an unsupervised change detection analysis of VHR remote sensing imagery and ii) systematic description of the nature and effects of imbalanced classes in this exemplary application setting. Thus, this study offers a sensitivity analysis framework on the nature and the effects of an arbitrary distribution of classes (i.e., class imbalance) in an unsupervised context for any two-class classification problem, which additionally enables evaluation of transferability as well as verification of the validity of any algorithm with respect to the distribution of classes.

This paper is organized as follows. Section 2 presents the two study areas, the city of Dongying in China and the city of Munich in Germany, respectively, and their individual data settings. In Section 3, a brief review of the employed methodology for change detection is given first, while Section 3.2 entails the proposed framework for evaluation of class imbalance. Section 4 presents experimental results, highlighting the nature and the effects of imbalanced classes in Section 4.2. A detailed discussion is provided in Section 5 while concluding remarks are given in Section 6.

2. Study areas and data sets

The proposed experiments on class imbalance are conducted in two disparate areas of interest with similar number of buildings, each with distinct characteristics concerning the structure of the built environment and a specific ratio of class imbalance (i.e., proportion of changed and unchanged buildings). The first experimental site is located in the dynamic Chinese city of Dongying, while the second test site is situated in the less dynamic German city of Munich. In this paper, the test site of Dongying serves as a representative of simple and less complex urban structure with regular geometric arrangement as it is present in many cities that are strongly influenced by planning. Opposed to Dongying, the morphological structure of Munich is very diverse and the buildings are arranged in a complex configuration. Such a setting is characteristic for historically developed cities. A direct comparison of structural statistics and other relevant numbers to the analyses is given in Table 1.

The experimental site in Dongying city comprises 4119 individual buildings covering an area of 38.5 km^2 in 2013 (Table 1, Fig. 1), which

Table 1
Statistics of the two test sites in Dongying and Munich.

	Dongying	Munich
Number of buildings	4119	4272
Mean height of buildings [m]	14.8	14.5
Mean area of buildings [m^2]	1089	497
Mean volume of buildings [m^3]	17048	7247
Average distance to nearest building [m]	12.0	2.1
Main orientation of buildings	E-W (87%)	N-S (35%)
Area of test site [km^2]	38.5	9.7
Built-up density [%]	17.5	29.2
Temporal scale of remote sensing images	2007–2013	2001–2010
Class proportions [unchanged: changed]	54: 46	93: 7