Available online at www.sciencedirect.com
SCIENCE DIRECT®
Remote Sensing of Environment 98 (2005) 468 – 480
Remote Sensing of Environment
www.elsevier.com/locate/rse

On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification

Thomas G. Van Niel [a,d,*], Tim R. McVicar [b], Bisun Datt [c]

[a] CSIRO Land and Water, Private Bag No 5, Wembley, WA 6913, Australia
[b] CSIRO Land and Water, PO Box 1666, Canberra, ACT 2601, Australia
[c] CSIRO Earth Observation Centre, GPO Box 3023, Canberra, ACT 2601, Australia
[d] Cooperative Research Centre for Sustainable Rice Production, Yanco, NSW 2703, Australia

Received 24 January 2005; received in revised form 23 August 2005; accepted 29 August 2005

Abstract

The number of training samples per class (n) required for accurate Maximum Likelihood (ML) classification is known to be affected by the number of bands (p) in the input image. However, the general rule which defines that n should be $10p$ to $30p$ is often enforced universally in remote sensing without questioning its relevance to the complexity of the specific discrimination problem. Furthermore, identifying this many training samples is often problematic when many classes and/or many bands are used. It is important, then, to test how this generally accepted rule matches common remote sensing discrimination problems because it could be unnecessarily restrictive for many applications. This study was primarily conducted in order to test whether the general rule defining the relationship between n and p was well-suited for ML classification of a relatively simple remote sensing-based discrimination problem. To summarise the mean response of n-to-p for our study site, a Monte Carlo procedure was used to randomly stack various numbers of bands into thousands of separate image combinations that were then classified using an ML algorithm. The bands were randomly selected from a 119-band Enhanced Thematic Mapper-plus (ETM+) dataset comprised of 17 images acquired during the 2001–2002 southern hemisphere summer agricultural growing season over an irrigation area in south-eastern Australia. Results showed that the number of training samples needed for accurate ML classification was much lower than the current widely accepted rule. Due to the asymptotic nature of the relationship, we found that 95% of the accuracy attained using $n = 30p$ samples could be achieved by using approximately $2p$ to $4p$ samples, or $\leq 1/7$th the currently recommended value of n. Our findings show that the number of training samples needed for a simple discrimination problem is much less than that defined by the general rule and therefore the rule should not be universally enforced; the number of training samples needed should also be determined by considering the complexity of the discrimination problem.

Keywords: Crop classification; Dimensionality; Training sample; Time-series; Multi-temporal; Maximum likelihood

1. Introduction

The 'curse of dimensionality' is the tendency for model accuracy to initially increase as the number of variables (e.g., bands, p) used increases, but then reach a limit where accuracy decreases—the point where the model is overfit (Hand, 1981; Hughes, 1968; Pal & Mather, 2003). This phenomenon is called 'peaking' in the pattern recognition literature (Jain & Waller, 1978) and in the remote sensing literature has been referred to as the Hughes (1968) effect (Foody & Arora, 1997; Pal & Mather, 2003). Peaking is caused by a poor estimation of the class probability density function (pdf) by the training data (Hand, 1981). This poor estimation of the class pdf will commonly occur in remote sensing classification when too many bands are used with respect to training sample size (n). If n is too small, the class pdf will not have enough precision to accurately estimate the too complex feature space. This phenomenon has been shown to affect the maximum likelihood (ML) classifier (Pal & Mather, 2003) where the common practice of modelling non-Gaussian remote sensing class pdf's with a single Gaussian distribution would exacerbate the issue.

In order to avoid peaking, it is common practice in remote sensing to ensure that n be comprised of at least 10 to 30 times

* Corresponding author. CSIRO Land and Water, Private Bag No 5, Wembley, WA 6913, Australia. Tel.: +61 8 9333 6705.
E-mail address: Tom.VanNiel@csiro.au (T.G. Van Niel).

0034-4257/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.rse.2005.08.011

the number of discriminating wavebands (i.e., $n=10p$ to $30p$ Mather, 1999; Piper, 1987). Most likely this n-to-p relationship was originally intended to be a 'rule of thumb', but for most, has turned into a generic rule to be applied universally. This is partly because this relationship is very often presented as a definitive statement or rule without qualification (James, 1985; Jensen, 1986; Mather, 1999; Piper, 1992; Pal & Mather, 2003). As such, this heuristic rule is often enforced without question, even though common sense dictates that the number of training samples required to achieve optimal accuracy will ultimately depend upon the discrimination problem, which in turn depends upon the characteristics of the data, the site, and the resultant classification level desired.

The spatial, spectral, and temporal characteristics of both the phenomena being mapped and the data itself (in conjunction with the quantization level of the data—Roderick et al., 1996) certainly impact upon discrimination of classes (McVicar et al., 2002; Woodcock & Strahler, 1987), and thus on n. Likewise, the level of detail of the resultant classification (e.g., Anderson level), although often related to both the data and site characteristics, could also alter the relationship between n and p. For example, with a given data source and a constant number of bands, to achieve the same classification accuracy, a species-level classification would be expected to require more training samples than a growth form-level classification of the same site. Likewise, a study site containing low within-class variance and high between-class variance, like irrigated crops with significantly different planting dates, might need less training data than at a site where classes have high within-class and low between-class variance, like many *Eucalypt* forests.

These data- and site-specific differences make it difficult to draw generic conclusions suitable for all studies, where different data sources, study sites, and classification levels are employed. Also, very often, published studies combine only a few images prior to classification (for an exception, see Key et al., 2001), so the large number of images required to describe the n-to-p relationship means that it is not usually studied with remotely sensed data. That is why the fundamental studies on the relationship between training sample size and dimensionality for ML classification which are cited in the remote sensing literature are based on chromosomal data (Piper, 1987, 1992) and probability theory (Hughes, 1968) instead of remote sensing data. The remote sensing-based studies have either concentrated on artificial neural network classification accuracy (Hepner, 1990; Foody et al., 1995; Foody & Arora, 1997), or have not tested a large enough range of n to achieve peaking in ML classification (Dobbertin & Biging, 1996; Pal & Mather, 2003). Subsequently, the relationship between n and p cannot be defined from these remote sensing-based studies, but only partially inferred (Dobbertin & Biging, 1996; Pal & Mather, 2003).

Because only part of the n-to-p relationship has been observed in those studies (where accuracy continues to increase with increasing p), it is also easily misinterpreted. Consequently, little is known about how the generally accepted rule defining the ratio of n-to-p matches common remote sensing discrimination problems. Does this general rule define the minimum requirement needed for a relatively simple remote sensing-based discrimination problem or does it define better what is required for a relatively complex case? We suggest that in the wide range of remotely sensed classification applications, this rule is unnecessarily restrictive for many discrimination problems that potentially would not need as many training samples as $n=10p$ to $30p$ to be accurately classified. We test this hypothesis using broadband multi-temporal classification in a Monte Carlo analysis of a dataset with high temporal density (i.e., 17 ETM+ images in a single growing season). We specifically aimed to determine if the general rule defining that $n=10p$ to $30p$ was required for our relatively simple discrimination problem. Due to the range of n and p studied here, the current work provided a characterisation of multi-temporal accuracy trends where the dependence on the timing of image acquisition relative to the crop phenology was minimised.

2. Methods and data

2.1. Study area and imagery

The study site is the 95,000 ha Coleambally Irrigation Area (CIA), New South Wales, Australia (Fig. 1), where the primary summer crop is irrigated rice. Rice uses the vast majority of available water since it is both permanently ponded between October and March and is planted on more area than any other crop. The other major summer crops are maize, sorghum and soybeans, which all use less water as they are both intermittently furrow-irrigated and grown on much less area. The CIA falls completely within the east–west overlap of two ETM+ scenes, allowing for twice as many image acquisitions, nominally, every 8 days. This provided 17 cloud-free images during the southern hemisphere summer growing season between October 2001 and May 2002; see Table 1. The 17 images provided very good coverage of the entire growing season. In every month except December, at least 2 'cloud-free' images were acquired (Table 1). The mean acquisition interval was 13 days (SD=8.24 days) with a maximum interval of 32 days (twice in the growing season). Using a Monte Carlo approach to combine bands from the 17 images meant that the dense temporal sampling reduced the dependence of subsequent results on specific acquisition dates and allowed the general n-to-p relationship to be assessed. For an in-depth review of remote sensing of irrigated rice as well as the impact of the timing of image acquisition at the site, see Van Niel and McVicar (2004a,b), respectively.

2.2. Validation data

The validation data were acquired from 2 sources: (1) digitised field boundaries from 1.5 m resolution aerial photographs acquired in the 2000–2001 and 2001–2002 southern hemisphere summer growing seasons (used for per-field classification); and (2) landholder surveys providing field-level summer crop type data for 283 fields. Of these 283 fields,