ELSEVIER

Computers & Geosciences 33 (2007) 875–897

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Hybrid image classification and parameter selection using a shared memory parallel algorithm

Rhonda D. Phillips*, Layne T. Watson, Randolph H. Wynne

Departments of Computer Science, Mathematics, and Forestry, Virginia Polytechnic Institute and State University, USA

Received 27 February 2006; received in revised form 25 September 2006; accepted 31 October 2006

Abstract

This work presents a shared memory parallel version of the hybrid classification algorithm IGSCR (iterative guided spectral class rejection) to facilitate the transition from serial to parallel processing. This transition is motivated by a demonstrated need for more computing power driven by the increasing size of remote sensing data sets due to higher resolution sensors, larger study regions, and the like. Parallel IGSCR was developed to produce fast and portable code using Fortran 95, OpenMP, and the Hierarchical Data Format version 5 (HDF5) and accompanying data access library. The intention of this work is to provide an efficient implementation of the established IGSCR classification algorithm. The applicability of the faster parallel IGSCR algorithm is demonstrated by classifying Landsat data covering most of Virginia, USA into forest and non-forest classes with approximately 90% accuracy. Parallel results are given using the SGI Altix 3300 shared memory computer and the SGI Altix 3700 with as many as 64 processors reaching speedups of almost 77. Parallel IGSCR allows an analyst to perform and assess multiple classifications to refine parameters. As an example, parallel IGSCR was used for a factorial analysis consisting of 42 classifications of a 1.2 GB image to select the number of initial classes (70) and class purity (70%) used for the remaining two images.

Keywords: Parallel processing; Remote sensing; Landsat; Forest area; High performance computing; Data parallel computing

1. Introduction

As remote sensing data sets continue to increase in size, there is a demonstrated need for faster computing resources to decrease processing time. Furthermore, when dealing with certain classification algorithms, more accurate results may be obtained by using slightly different input parameters. A significantly faster (parallel) implementation of these classification algorithms would allow an analyst to make several runs using different parameters in the equivalent time required to make one serial run, potentially producing more accurate classification results. Although there are increasingly more parallel computers available to the research community, porting existing serial applications to a parallel environment is usually non-trivial. This paper discusses specific changes that are made to the IGSCR (iterative guided spectral class rejection) classification algorithm to produce a shared memory parallel algorithm with accompanying

*Corresponding author. Tel.: +1 540 231 6931; fax: +1 540 231 6075.

E-mail addresses: rdphllps@vt.edu (R.D. Phillips), ltw@cs.vt.edu (L.T. Watson), wynne@vt.edu (R.H. Wynne).

0098-3004/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2006.10.014

pseudocode for the classification algorithms (Wayman et al., 2001; Musy et al., 2006). A further goal of this implementation is to create source code that is both portable and open source. The final parallel IGSCR code runs on multiple hardware platforms and operating systems, and it does not have the same "black box" that is associated with commercial software libraries. A final goal of this work is to demonstrate the utility of the parallel IGSCR implementation by accurately and efficiently classifying Landsat data covering the state of Virginia into forest and non-forest land use informational classes.

The following is a description of the outline of the rest of the paper. The second section contains a comprehensive review of the work that has lead up to this work as background information. The third section includes detailed descriptions of the serial IGSCR algorithm and the serial *K*-means and maximum likelihood algorithms that comprise the IGSCR algorithm. The fourth section includes a description of the Hierarchical Data Format version 5 (HDF5) and Application Programming Interface (API) and how these are used in this implementation. The fifth section describes the modifications that are made to serial IGSCR to produce a parallel IGSCR algorithm, and the sixth section demonstrates the parallel results and accompanying analysis. The seventh section wraps up the paper with a discussion of conclusions reached.

2. Background

2.1. IGSCR

Unsupervised classification is a process by which all pixels or objects with similar spectral values (spectral classes) are identified (clustering) and then subsequently labeled with respect to informational classes (labeling). Supervised classification, in contrast, requires analyst identification of the spectral classes within each informational class beforehand (training). Remaining pixels or objects are then assigned to a spectral class using a decision rule (classification). As with unsupervised classification, the resulting map must be labeled with respect to informational classes, but for supervised classifications this is trivial since the informational class to which each spectral class belongs was identified in the training stage.

IGSCR is an example of a hybrid classification method, a classification method that exhibits characteristics of both unsupervised and supervised classification (Richards and Jia, 1999). Hybrid classification methods combine multiple classifiers to reduce the workload of the human analyst, most often in the training phase. Bruzzone and Prieto (2001) use unsupervised classification (clustering) to modify the spectral signatures generated from a supervised classification so the same training data can be used on images of the same landscape acquired on different dates. Byeungwoo and Landgrebe (1999) use a hybrid approach to create a one class classification where the analyst need only train for the class of interest and then unsupervised classification is used to generate signatures for a supervised classification of the original image. Guided clustering requires a user to select training data to represent predefined informational classes, automatically identifies spectral classes (clusters of pixels with similar brightness value vectors) within each informational class (category, such as deciduous forest or row crops) using a clustering algorithm, and then uses the resulting spectral class signatures to perform a supervised classification (Bauer et al., 1994). This method is advantageous because accurate results are produced while allowing for a greater amount of automation (Bauer et al., 1994). Guided clustering cannot be entirely automated, however, as user interaction is required after the training process to oversee spectral class creation and refine parameters. IGSCR is more disposed to automation as no user interaction is required after the training phase (Wayman et al., 2001). IGSCR uses a process called "cluster busting" first introduced by Jensen et al. (1987) to refine spectral classes iteratively prior to application of a decision rule. Each spectral class produced by clustering is assigned the value of the majority informational class if that spectral class is statistically determined to be sufficiently pure. In practice, IGSCR-derived area estimates were shown to exceed established precision standards in the USDA Forest Service Forest Inventory and Analysis (FIA) program (Musy et al., 2006). IGSCR was also used recently for a study on Sudden Oak Death where Kelly et al. (2004) demonstrated that the IGSCR hybrid classification method outperformed both supervised and unsupervised methods alone.

2.1.1. Automation

The IGSCR algorithm requires a clustering algorithm and a means by which brightness value vectors are assigned to clusters (the decision rule),