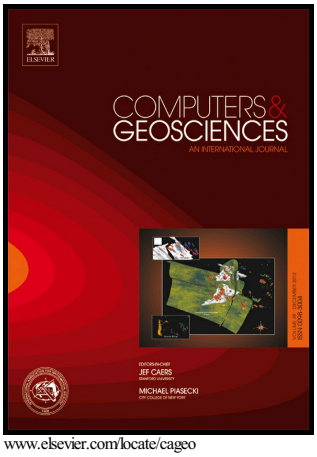
Decision Forests for Machine Learning Classification of Large, Noisy Seafloor Feature Sets

Ed Lawson, Denson Smith, Donald Sofge, Paul Elmore, Frederick Petry



www.elsevier.com/locate/cageo

PII: S0098-3004(16)30584-2
DOI: http://dx.doi.org/10.1016/j.cageo.2016.10.013
Reference: CAGEO3858

To appear in: *Computers and Geosciences*

Received date: 24 April 2015
Revised date: 1 September 2016
Accepted date: 25 October 2016

Cite this article as: Ed Lawson, Denson Smith, Donald Sofge, Paul Elmore and Frederick Petry, Decision Forests for Machine Learning Classification of Large Noisy Seafloor Feature Sets, *Computers and Geosciences* http://dx.doi.org/10.1016/j.cageo.2016.10.013

Decision Forests for Machine Learning Classification of Large, Noisy Seafloor Feature Sets

Ed Lawson[1], Denson Smith[2], Donald Sofge[1], Paul Elmore[2], Frederick Petry[2*]

[1]Naval Research Laboratory, Information Technology Division, Washington, DC

[2]Naval Research Laboratory, Marine Geosciences Division, Stennis Space Center, MS

[*]Corresponding author: 228 688 4948. fpetry@nrlssc.navy.mil

Abstract

Extremely randomized trees (ET) classifiers, an extension of random forests (RF) are applied to classification of features such as seamounts derived from bathymetry data. This data is characterized by sparse training data from by large noisy features sets such as often found in other geospatial data. A variety of feature metrics may be useful for this task and we use a large number of metrics relevant to the task of finding seamounts. The major significant results to be described include: an outstanding seamount classification accuracy of 97%; an automated process to produce the most useful classification features that are relevant to geophysical scientists (as represented by the feature metrics); demonstration that topography provides the most important data representation for classification. As well as achieving good accuracy in classification, the human-understandable set of metrics generated by the classifier that are most relevant for the results are discussed.

Keywords - bathymetry, topography, seamounts, random forests, extremely randomized trees

1. Introduction

The basic approach of classification algorithms is based on feature metrics characterizing classification goals, in our case identification of oceanic seamounts and ridges. Selected data is used in which objects to be identified are labeled and used to train the particular classifier. Then the trained classifier can be used to identify objects in unlabeled data.

It is not unusual for classifications of geospatial features such as found in seafloor bathymetry data to involve sparse training data characterized by large noisy features sets. We show that extremely randomized trees (ET) classifiers can perform well in such an environment.

Classification of undersea geomorphological structures is typically performed by oceanographers producing bathymetric charts using seafloor depths measured from both higher-resolution acoustic readings and low-resolution satellite altimetry. Coverage of scenes are often sparse, due to cost of acquiring data. It's generally believed that less than 10% of the seafloor has been mapped with high-resolution bathymetric charts [Becker, 2009].