Machine Learning in Benthic Habitat Mapping

Justin Ting — Honours Student

Simon O'Callaghan — NICTA Researcher

NICTA

SIT

2016-05-06

└Machine Learning in Benthic Habitat Mapping

Machine Learning in Benthic Habitat Mapping

Justin Ting — Honours Student

Simon O'Callaghan — NICTA Researcher

NICTA

SIT

Introduction

- Less than 10% of the world's oceans are mapped compared to 99% of Earth's topology mapped (low resolution)[1]
- To map more of the world's oceans at a high resolution, we employ benthic habitat mapping techniques
- Marine habitat mapping cuts across marine biology, geology, hydrography, oceanography, geophysics (Brown et al., 2011), along with habitat mapping

[1] http://www.wired.com/2009/06/nasa-satellite-maps-99-of-earths-topography/

Introduction
- Less than 10% of the world's oceans are mapped compared to 99% of Earth's topology mapped (low resolution)[1]
- To map more of the world's oceans at a high resolution, we employ benthic habitat mapping techniques
- Marine habitat mapping cuts across marine biology, geology, hydrography, oceanography, geophysics (Brown et al., 2011), along with habitat mapping

[1] http://www.wired.com/2009/06/nasa-satellite-maps-99-of-earths-topography/

- 0.5 minutes
- What is benthic habitat mapping?
- Benthic - ecological region at the lowest level of a body of water
- benthos - things related to the benthic layer
- Habitat mapping - based on a small amount of high resolution data, and a considerably larger amount of lower resolution data, a relationship is created to correlate the data in overlapping regions, which is then projected to the regions without the high resolution data to create a 'habitat map'
- The high resolution data is generally actual sediment samples or organism samples at the benthos
- Low resolution data is generally some sort of acoustic data representing basic properties of the seafloor

Problem Statement

- Much research in benthic habitat mapping generates deterministic maps using as-is machine learning techniques/implementations
- We need to be able to monitor marine habitats on a large scale to assess human impact over time to be able to make informed management decisions

Problem Statement

- Much research in benthic habitat mapping generates deterministic maps using as-is machine learning techniques/implementations
- We need to be able to monitor marine habitats on a large scale to assess human impact over time to be able to make informed management decisions

- 1 minute
- Use of 'vanilla' algorithms such as random forests in Lucieera et al. (2013), Seiler et al. (2012), Hasan et al. (2014)

Solution - Overview

- Probabilistic approach allows us to state certainty about a particular mapped area
- Use of Gaussian Processes requires a matrix inversion step with $O(n^3)$ complexity - attempt to overcome this by making our covariance matrix sparse and hence the inversion step computationally feasible for large datasets

└─Solution - Overview

Solution - Overview

- Probabilistic approach allows us to state certainty about a particular mapped area
- Use of Gaussian Processes requires a matrix inversion step with $O(n^3)$ complexity - attempt to overcome this by making our covariance matrix sparse and hence the inversion step computationally feasible for large datasets

- 1 minute

Solution - Gaussian Processes

TODO - (very) brief overview of what GPs are here, without going into the maths

2016-05-06

Solution - Gaussian Processes

Solution - Gaussian Processes

TODO - (very) brief overview of what GPs are here, without going into the maths

- 1.5 minutes

Results

We would like to compare the results obtainable from use of Gaussian Processes in generating habitat maps to those of more naive methods. Note that measurements are taken by averaging over 10-fold cross validations.

This first set of results is using 24 separate habitat classes as the labels.

Algorithm	F1 score	Accuracy
KNN (5)	0.04823	0.14192
Logistic Regression	0.00900	0.17119
Random Forest	0.04960	0.15240
SVM	0.00890	0.17119

2016-05-06

└Results

Results

We would like to compare the results obtainable from use of Gaussian Processes in generating habitat maps to those of more naive methods. Note that measurements are taken by averaging over 10-fold cross validations.

This first set of results is using 24 separate habitat classes as the labels.

Algorithm	F1 score	Accuracy
KNN (5)	0.04823	0.14192
Logistic Regression	0.00900	0.17119
Random Forest	0.04960	0.15240
SVM	0.00890	0.17119

- 15 seconds just highlighting results here

Results

This second set of results is using a higher level mapping of the habitats created in collaboration with an expert - bringing the number of classes down to 5. The accuracy increases notably when the aggregation of visually matching habitat classes is performed.

Algorithm	F1 score	Accuracy
KNN (5)	0.33278	0.62459
Logistic Regression	0.20285	0.682705
Random Forest	0.20283	0.68258
SVM	0.20284	0.68270

Results

This second set of results is using a higher level mapping of the habitats created in collaboration with an expert - bringing the number of classes down to 5. The accuracy increases notably when the aggregation of visually matching habitat classes is performed.

Algorithm	F1 score	Accuracy
KNN (5)	0.33278	0.62459
Logistic Regression	0.20285	0.682705
Random Forest	0.20283	0.68258
SVM	0.20284	0.68270

- 15 seconds just highlighting results here

These final two set of results are using Gaussian processes with the granular and aggregated habitat classes, respectively.

Algorithm	F1 score	Accuracy
KNN (5)	0	0
Logistic Regression	0	0
Random Forest	0	0
SVM	0	0

Algorithm	F1 score	Accuracy
KNN (5)	0	0
Logistic Regression	0	0
Random Forest	0	0
SVM	0	0

2016-05-06

These final two set of results are using Gaussian processes with the granular and aggregated habitat classes, respectively.

Algorithm	F1 score	Accuracy
KNN (5)	0	0
Logistic Regression	0	0
Random Forest	0	0
SVM	0	0

Algorithm	F1 score	Accuracy
KNN (5)	0	0
Logistic Regression	0	0
Random Forest	0	0
SVM	0	0

- 1.5 minutes (2 minute for results in total)
- NOTE still all zeroes as I will see if I can put in any *actual* results rather than completely faking it as we were told
- notably higher accuracies than with the previous methods for the granular and aggregated habitat classes respectively
- **TODO** add fake data for *other* datasets?
- **TODO** add times taken to run tests - important in terms of time/memory/etc tradeoffs

Discussion and Analysis

- **TODO**

Discussion and Analysis

- TODO

- 2 minutes
- Compare (fake) results with Bender et al. (2012) and explain the improvement (or otherwise) on particular datasets
- highlight (potentially) better results with one dataset, but not in another (e.g. o'hara bluffs having more variation than scott reef)

Bibiolgraphy

Asher Bender, Stefan B., Williams, and Oscar Pizarro. Classification with probabilistic targets. 2012.

Craig Brown, Stephen J Smith, and Peter Lawton. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuarine, Coastal and Shelf Science*, 92, 2011.

Rozaimi Che Hasan, Daniel Ierodiaconou, Laurie Laurenson, and Alexandre Schimel. Integrating multibeam backscatter angular response, mosaic and bathymetry data for benthic habitat mapping. *PLoS ONE*, 9, 2014.

Vanessa Lucieera, Nicole A. Hilla, Neville S. Barretta, and Scott Nichol. Do marine substrates look and sound the same? supervised classification of multibeam acoustic data using autonomous underwater vehicle images. *Estuarine, Coastal and Shelf Science*, 117: 94–106, 2013.

Jan Seiler, Ariell Friedman, Daniel Steinberg, Neville Barrett, Alan Williams, and Neil J. Holbrook. Image-based continental shelf habitat mapping using novel automated data extraction techniques.

2016-05-06

Bibiolgraphy

Bibiolgraphy

Asher Bender, Stefan B., Williams, and Oscar Pizarro. Classification with probabilistic targets. 2012.

Craig Brown, Stephen J Smith, and Peter Lawton. Benthic habitat mapping: A review of progress towards improved understanding of the spatial ecology of the seafloor using acoustic techniques. *Estuarine, Coastal and Shelf Science*, 92, 2011.

Rozaimi Che Hasan, Daniel Ierodiaconou, Laurie Laurenson, and Alexandre Schimel. Integrating multibeam backscatter angular response, mosaic and bathymetry data for benthic habitat mapping. *PLoS ONE*, 9, 2014.

Vanessa Lucieera, Nicole A. Hilla, Neville S. Barretta, and Scott Nichol. Do marine substrates look and sound the same? supervised classification of multibeam acoustic data using autonomous underwater vehicle images. *Estuarine, Coastal and Shelf Science*, 117: 94–106, 2013.

Jan Seiler, Ariell Friedman, Daniel Steinberg, Neville Barrett, Alan Williams, and Neil J. Holbrook. Image-based continental shelf habitat mapping using novel automated data extraction techniques. *Continental Shelf Research*, 45:87–97, 2012.