Machine Learning in Benthic Habitat Mapping - Outline of Research Approach
Research Methods - INFO5993 Assignment 2

Justin Ting, 430203826

April 2016

1 Introduction

This study will be about benthic habitat mapping, which is the ”spatial representation of physically distinct areas of seafloor that are associated with particular groups of plants and animals” (Harris and Baker, 2012). Its significance lies in the fact that there is an increasing need by regulatory bodies such as governments, etc. to better manage bodies of water which are being directly affected by human activity to preserve their state and prevent further damage, etc. Many studies have already collected a range of different data source and tried varying techniques at creating habitat maps using such data. To date, there have not yet been attempts to combine the predictive power of Gaussian proceses (Bender et al., 2012) fully with the increasing amount of bathymetric data that modern technology can collect. The purpose of this study will hence be to capitalise all the data we have access to to create better benthic maps to allow decision making bodies to make more informed decisions.

2 Aims and Research Questions

Our aim is to create a habitat mapping scheme that can provide more accurate habitat maps than current state of the art methods by attempting to solve some of the obstacles that are present in them. For our proposed research, we are specifically looking at exploring the following questions:

- How can we improve the quality of benthic habitat mapping?
- How can we put the existing data we have to better use (i.e. not needing to go on expeditions to obtain new data)?
- How can we use Gaussian Processes (henceforth GP) and the available existing data to improve accuracy of the mapping process?

The former two questions essentially cumulate to form the third, which will be the focus question for our research.

3 Proposed Methodology

3.1 Data Collection

Bathymetric and truthing data from at least Scott Reef and O'Hara Bluffs will be used to verify the performance of the algorithms being tested. Data for both these locations have been used in previous studies(Bender et al. (2012), Ahsan et al. (2011)).

Datasets from different marine contexts are needed as the distribution and variance of properties in different areas can vary considerably, causing some algorithms to work better in certain (physical) conditions. This can be seen in (Ahsan et al., 2011), where there were differences in the accuracy metric depending on the location of the benthic habitat in question - O'Hara, Chevron, and Scott Reef.

3.2 Pre-processing

To address the question, we will begin by exploring possible methods to apply GPs to large datasets. The obstacle faced with the use of GPs is in applying the matrix inversion to the raw covariance matrix which has a worst case complexity of $O(n^3)$ where n is the input size - this is a major bottleneck as it does not scale, preventing use with larger datasets.

One method of overcoming this that has been explored in literature is to 'transform' the full matrix into a sparse one at the inversion step - (Melkumyan and Ramos, 2009) and (Furrer et al., 2006) detail a method that involves a 'cut-off' point within the covariance matrix such that rather than the covariance values tapering off and approaching (but never reaching) zero, after a certain point it is instead actually set to zero. Once this step is done, there are known ways to invert such matrices in much less than $O(n^3)$ time, with a range of libraries in differnet languages implementing such methods. Other ways of doing this will be explored as well as necessary. To supercede the current state of the art in benthic habitat mapping however, it would be prudent to implement the future work suggested (Quinonero-Candela and Rasmussen, 2005) to obtain the best approximation of the large GPs, which involves combining the Partially Independent Training Conditional approximation with "the most powerful selection method for the inducing inputs."

3.3 Measuring Performance of GPs on large dataset

Given that (Bender et al., 2012) is one of the more recent studies employing the use of GPs in benthic habitat mapping, the first metric to compare would be whether including a much larger subset of the original data, if not all of it, improves upon the performance of the work which we are aiming to build on.

Intrinsic accuracy of the method with the stated pre-processing steps will be tested via cross-validation and checked that it exceeds (or not) the probabilistic target least squares classifier (PTLSC) in accuracy.

Performance will also be measured relative to a number of other methods that were found to be highly performant in the *literature review*. These include, in particular, **random forests**, and **boosted decision trees** as well (though the latter method has yet to see much adoption when creating benthic habitat maps).

3.4 Limitations

The two distinct marine locations for which real data is obtainable is likely not representative enough to get a full picture of how well the method proposed will perform in different contexts. During the duration of the study, some effort will be put into obtaining more datasets from other authors of similar studies. Failing that, as a backup/last resort, we will consider generating synthetic data based on information available in past research papers that detail the properties to diversify the data which we use to assess our method.

References

Nasir Ahsan, Stefan B. Williams, and Oscar Pizarro. Robust broad-scale benthic habitat mapping when training data is scarce. 2011.

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

Reinhard Furrer, Marc G. Genton, and Douglas Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15, 2006.

Peter T. Harris and Elaine K. Baker. *Seafloor Geomorphology as Benthic Habitat*. Elsevier Inc., 2012.

Arman Melkumyan and Fabio Ramos. A sparse covariance function for exact gaussian process inference in large datasets. *International Joint Conference on Artificial Intelligence*, 9, 2009.

Joaquin Quinonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959, 2005.