Statoil/C-CORE Iceberg Classifier Challenge

Tee Kah Hui, Kalkidan Fekadu, Samuel Pegg
Department of Computer Science and Technology
Tsinghua University
30 Shuangqing Rd, Haidian District, Beijing
jh-zheng20@mails.tsinghua.edu.cn, chefirakf10@mails.tsinghua.edu.cn,
peggsr10@mails.tsinghua.edu.cn

Abstract

In this proposal we discuss the Kaggle Competition we wish to undertake for the group assignment. In this competition we are required to identify if a remotely sensed target is a ship or iceberg using machine learning techniques.

1 Introduction

Drifting icebergs present threats to navigation and activities in areas such as offshore of the East Coast of Canada.

Currently, many institutions and companies use aerial reconnaissance and shore-based support to monitor environmental conditions and assess risks from icebergs. However, in remote areas with particularly harsh weather, these methods are not feasible, and the only viable monitoring option is via satellite.

The data set of for competition is provided by Statoil, an international energy company operating worldwide, has worked closely with companies like C-CORE. C-CORE have been using satellite data for over 30 years and have built a computer vision based surveillance system. To keep operations safe and efficient, Statoil is interested in getting a fresh new perspective on how to use machine learning to more accurately detect and discriminate against threatening icebergs as early as possible.

In this competition, the challenge is to build a machine learning algorithm that automatically identifies if a remotely sensed target is a ship or iceberg. Improvements made will help drive the costs down for maintaining safe working conditions.

2 Related Work

Elisa et al. used machine-learning regression methods to estimate the wheat yield across Australia with climate records and satellite images time series. Out of nine base learners and two ensembles,

support vector regression with radial basis function emerged as the single best learner (root mean square error of 0.55 and R^2 of 0.77 at the pixel level).[1]

Swathy et al. estimates high-resolution sea surface temperature (SST) by using three distinct machine learning techniques such as Artificial Neural Networks (ANN), Support Vector Regression (SVR) and Random Forest (RF)-based algorithms. They proposed that the proposed SVR-based algorithm has huge potential to produce operational high-resolution cloud-free SST estimates, even if there is cloud cover in the image.[2] Ashikur et al. studied the performance of different machine algorithm on three different spatial and multi-spectral satellite image classification in rural and urban extents. Random forest, Support Vector Machine (SVM), and their combined strength (stacked algorithms) were applied on Landsat-8, Sentinel-2, and Planet images separately to assess individual and overall class accuracy of the images. Among the three different algorithms, SVM showed comparatively better results with an overall accuracy of 0.969, 0.983, and overall kappa of 0.948, and 0.968, respectively.[3]

Mart et al. used convolutional neural networks method to perform image recognition in high-resolution, multi-spectral satellite imagery. The system consists of an ensemble of convolutional neural networks and additional neural networks that integrate satellite metadata with image features.[4]

3 Methodology

In this project we plan to explore various machine learning methodologies to get the optimal result.

3.1 Data

The data (train.json, test.json) is presented in json format. The files consist of a list of images, and for each image, you can find the following fields:

- id - the id of the image
- band_1, band_2 - the flattened image data. Each band has 75x75 pixel values in the list, so the list has 5625 elements. Note that these values are not the normal non-negative integers in image files since they have physical meanings - these are float numbers with unit being dB. Band 1 and Band 2 are signals characterized by radar back scatter produced from different polarization at a particular incidence angle. The polarization correspond to HH (transmit/receive horizontally) and HV (transmit horizontally and receive vertically). More background on the satellite imagery can be found here.
- inc_angle - the incidence angle of which the image was taken. Note that this field has missing data marked as "na", and those images with "na" incidence angles are all in the training data to prevent leakage.
- is_iceberg - the target variable, set to 1 if it is an iceberg, and 0 if it is a ship. This field only exists in train.json.

3.2 Structure of Image classification task

3.2.1 Image Processing

The aim of this process is to improve the image data(features) by suppressing unwanted distortions and enhancement of some important image features so that our Computer Vision models can benefit from this improved data to work on.[5] For Iceberg classifier contest the images have two channels: HH (transmit/receive horizontally) and HV (transmit horizontally and receive vertically). The background information on the competition website mentions that this can play an important role in the object characteristics, since objects tend to reflect energy differently. Hence, in this stage we will investigate the influence of the two channels and how to pre-process them in away that can improve the performance of the subsequent steps. We intend to use techniques such as histogram equalization, rotation etc.

3.2.2 Object Detection

In this step we will use image segmentation to identify the location of the object, in this case the iceberg or ship, in the image.

3.2.3 Classification

This will be a supervised learning model as the data set for training is labelled. Binary classification method will be used as the requirement is to detect the existence of an iceberg in satellite image. Popular algorithms that can be used for binary classification include:

- Logistic Regression
- k-Nearest Neighbors
- Decision Trees
- Support Vector Machine
- Naive Bayes

We will use image classification techniques such as SVM, ANN, Decision Trees, and CNN. This competition uses the log loss classification metric. The target variable will be set to 1 if an iceberg is detected.

References

[1] Elisa Kamir, François Waldner, Zvi Hochma, Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 160, 2020, Pages 124-135, ISSN 0924-2716.

[2] Swathy Sunder, RAAJ Ramsankaran, Balaji Ramakrishnan, Machine learning techniques for regional scale estimation of high-resolution cloud-free daily sea surface temperatures from MODIS data, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 166, 2020, Pages 228-240, ISSN 0924-2716

[3] Ashikur Rahman, Hasan Muhammad Abdullah, Md Tousif Tanzir, Md Jakir Hossain, Bhoktear M. Khan, Md Giashuddin Miah, Imranul Islam, Performance of different machine learning algorithms on satellite image classification in rural and urban setup, Remote Sensing Applications: Society and Environment, Volume 20, 2020, 100410, ISSN 2352-9385

[4] M. Pritt and G. Chern, "Satellite Image Classification with Deep Learning," 2017 IEEE Applied Imagery P

[5] https://iq.opengenus.org/basics-of-machine-learning-image-classification-techniques/