Machine Learning Engineer Nanodegree

Capstone Proposal

YuanFu Yang

March 04, 2018

Icerberg Classification (from Kaggle Competition)

Domain Backgroud

During the winter and spring, drifting icebergs in the North Atlantic ocean present threats to navigation and activities in areas such as offshore of the east coast of Canada. It pose a risk to people's lives and oil and gas equipment, as well as present challenges to shipping and other activities. Current methods for monitoring iceberg conditions is aerial reconnaissance, supplemented by platform and vessel-based monitoring. However, in remote areas with particularly harsh weather, these methods are not feasible, and the only viable monitoring option is via satellite. The data in this paper was detected by synthetic aperture radar (SAR). After detection, additional processing is needed to distinguish between ships and icebergs. The discrimination between the two classes is carried out through feature extraction and target classification steps. In previous studies [1], intensity and polarimetric parameters are used as features to a Support Vector Machine (SVM) classifier, in order to discriminate ships from icebergs in simulated, dual polarized, medium resolution SAR data. Recently, Convolutional Neural Networks (CNN) have been successfully adopted for demanding SAR classification tasks, as in [2][3]. Both SVM and CNN will be used in this study for iceberg classification.

Problem Statement

The goal of this project is to train a model that it able to recognize image as an iceberg or not, and it is a binary classification in machine learning terms. The model should be able to recognize arbitrary iceberg with a high degree of accuracy. There are four inputs in this dataset: HH, HV(Image data), angle, and label (is iceberg or not). We need to train a model that would be able to classify image into these two classes.

Datasets and Inputs

The data had collected by Sentinel-1, a space mission funded by the European Union and carried out by the European Space Agency (ESA) within the Copernicus program, consisting of a constellation of two satellites. The payload of Sentinel-1 is a synthetic aperture radar in C band that provides continuous imagery (day, night and all weather). They are not light-dependent and can see through the darkness, clouds, fog and most weather systems, taking snapshots over very large areas (up to 250 X 250 kilometers). It handles multi-inputs which are one meta data input and one image input for image data with two channels: **HH** (transmit horizontally and receive horizontally) and **HV** (transmit horizontally and receive vertically). In addition, the incidence **angle** of which the image was taken is another inputs in this datasets. The datasets has been provided by the Centre for Cold Ocean Resource Engineering (C-CORE) on Kaggle (https://www.kaggle.com/c/statoil-iceberg-classifier-challenge/data) and presented in json format. It consists of a list of images and other information. In the training file, there are 1604 data with 1 label (is iceberg or not), and 4 inputs (HH/HV/Angle/image_id). The labels are provided by human experts and geographic knowledge on the target. The distribution of labels seems to balance (Is_iceberg: 753, not_iceberg: 851). The images are 75x75 images with two bands (HH/HV). In the input of angle, there are 133 missing data need to handle before training.

Solution Statement

In this datasets, there are 2 type of predictive factors: imagical (HH/HV) and numerical (Angle) data. I will use 4 supervised learning methodology, Support Vector Machine (SVM), Random Forests (RF), Adaptive Boosting(AdaBoost), and eXtreme Gradient Boosting (XGboost) with all factors to predict test data. Before doing this, some feature engineering should be done. Then I will use deep learning - convolutional neural networks (CNN) to predict test data and compared with previous 4 supervised learning. In CNN, it only can be use imagical factors, so the "angle" will be dropped before model training. Furthermore, CNN model needs three input channels but we only have two (band_1 and band_2) in the datasets. I will take the mean of band_1 and band_2 to be used as the third channel.

Benchmark Model

For compare with final result, I choose "Random Forest" as my benchmark model. The score (precision/recall/F-score) of first run is 0.7465/0.6839/0.7331. It seems like solvable because the performance is acceptable in the absence of data extraction.

Evaluation Metrics

I use three index to calculate performance of my model: precision, recall, and F-score. The final result will be decided by F-score with beta = 0.5.

Project Design

- **Programming language** : Python 3.6
- **Libraries :** Keras, Tensorflow, Scikit-learn, XGboost (for model set up), Pandas (for data review and pre-process), Matplotlib, Seaborn, Plotly (for data visualization)
- **Workflow :**

1. Data Exploration: To overall inspection the datasets: missing data review (by Pandas), label balance check (using bar chart by Seabon), and data attribute check (by Pandas).

2. Data Pre-process: Dropping the missing data, normalizing numerical features, and transforming skewed continuous features.
3. Initial Model Evaluation: Before data engineering, I will properly evaluate the performance of each model (SVM, Random forests, Adaboost, XGBoost, CNN).
4. Feature Extraction: In CNN, the feature can be extracted in convolution process. In other supervised learning (SVM, Random forests, Adaboost, XGBoost), the feature will be extracted by statistics.
5. Data Augmentation: In order to bring some data augmentation to my model I will use keras's ImageDataGenerator functions and see if there is any improvement in performance (accuracy, precision, recall, F-score).
6. Model Tuning: To improve performance, I will use GridSearchCV by Scikit-learn to find the optimization parameters in each supervised learning model.

Finally, based on the above process, I will evaluate the performance of each model and choose the best as the result of this study.

References:

[1] Michael Denbina, Michael J Collins, and Ghada Atteia: On the detection and discrimination of ships and icebergs using simulated dual-polarized radarsat constellation data. Canadian Journal of Remote Sensing, (just-accepted):00¡V00, 2015.

[2] Carlos Bentes ; Anja Frost ; Domenico Velotto ; Bjoern Tings: Ship-iceberg discrimination with convolutional neural networks in high resolution SAR images. IEEE 05 Sep 2016, ISBN: 978-3-8007-4228-8

[3] Carlos Bentes, Domenico Velotto, and Susanne Lehner. Target classification in oceanographic SAR images with deep neural networks: Architecture and initial results. In IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2015, page 4, 2015.