Dataset Guide

RMDS Lab

1 DATA DESCRIPTION

Contestants are asked to create a statistical model to predict tropical cyclone (TC) intensity change in 24 hours using advanced machine learning techniques. There are 2 sets of training data from 1998 to 2010; one is the predictors operationally used in a statistical TC intensity forecast model (SHIPS.nc). The other is satellite observations with spatially and temporally varying features of TC structures (PrcpXXXX.nc). Combined use of two datasets are expected to yield better results than using one dataset only.

SHIPS dataset ranges from 1998 to 2010 and it contains multiple 1-Dimensional data. Each 'Date' represents 1 sample. While PrcpXXXX datasets are one year for each file and contain 4-D variable 'prcp' which reflects the Satellite observations with spatially and temporally varying features of TC structures. The "RLAT" and "RLON" coordinates represent the spatially varying features of TC structures. The "RHOUR" coordinate represents the temporally varying features of TC structures. -12 in RHOUR means 12 hours before current time.

2 DOWNLOAD

You can access the dataset directly through **grmds.org/competition-data**. You must have an account on our platform to participate in this competition.

Data Type Raw Data Data File Type Other Data File				
training_dataset.zip	54.92 MB			
Data Description Announcement: a training session to familiarize our contestants with the data will be hosted by our expert panels at NASA JPL, IBM, LMU. Keep posted for more updates about the competition and training!				
contestants are asked to create a statistical model to predict tropical cyclone (TC) intensity change in 24 hours using dvanced machine learning techniques. There are 2 sets of TC data from 1998 to 2010; one is the predictors operationally sed in a statistical TC intensity forecast model (SHIPS1998-2010.nc). The other is satellite observations with spatially and proposed from the detacate of TC structures (Proposed as). Combined use of the detacate are expected to yield batter.				

3 HOW TO OPEN

The files all have the extension .nc which is created in the netCDF format used to store multidimensional data and is often used for GIS and other spatial-temporal datasets. For each language (Python, R, C++) there is corresponding package or library to open it. Here we will provide the example of using Python to explore the data.

```
In [1]: from netCDF4 import Dataset import muny as mp data = butaset("Prp1998.nc', 'r') data = butaset("Prp1998.nc', 'r') data = butaset("Prp1998.nc', 'r') data = class | netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.netCDF4.net
```

There are three features contained in the Dataset class of netCDF which are 'dimensions', 'variables' and 'groups'. 'Dimensions(size)' describes the length of each variable while 'variables' represents the

value and description of variables as well as the coordinates when it has multiple dimensions. For example for the variable 'prcp':

The values of each variable were stored in a masked array. You can easily convert it to a numpy array and fill the missing value using the given number by default:

Note that -999.0 here signifies missing value.

4 EVALUATION

In the competition, we will focus on predicting TC rapid intensification (RI) occurrence in 24 hours. RI here is a categorical variable that will need to be derived from the data dymax. RI: 1 (dymax≥25) or 0 (dymax<25). For dichotomous categorical forecasts of a RI event, there are only two possible outcomes, Yes or No. Four scenarios for predictions against observations are classified: 1. Observed RIs are correctly predicted, termed "hits"; 2. Predicted RIs do not occur in observations, termed "false alarms"; 3. Observed RIs are not predicted, termed "misses", and 4. Both observations and predictions show no RIs, termed "corrected negatives". The number of cases for each scenario is labeled as a, b, c and d, respectively.

2x2 Contingency Table		Event Observed	
		Yes	No
Event	Yes	a (hits)	b (false alarms)
Forecast	No	c (misses)	d (correct negatives)

The probability of detection (POD) is defined as:

$$POD = \frac{a}{a+c}$$

which ranges from 0 to 1 with 0 being no hits and 1 being perfect. And the False Alarm Ratio (FAR) is defined as:

$$FAR = \frac{b}{a+b}$$

which measures reliability of the prediction, with 0 being perfect and 1 being no skills. A good prediction requires high POD and low FAR. We use the Peirce's skill score (PSS) to combine the values of POD and FAR, as commonly used in operational RI forecast (Kaplan et al. 2015). The PSS is computed as:

$$PSS = \frac{ad - bc}{(a+c)(b+d)}$$

which ranges from -1 to 1 with -1 being no skill and 1 being perfect. The dichotomous categorical forecast predicts a RI occurrence if the predicted dymax value meets the threshold (25 knots in this competition). For a binary dichotomous forecast based on a threshold from a continuous predictor, the PSS is found to be a good scalar measure of the overall skill.

5 CONTACT

If you have other questions about the competition or dataset, please reach out to us through ${\bf cici@rmdslab.com}$.