AGU December 2015 H43E-1542

Application of Data Cubes for Improving Detection of Water Cycle Extreme Events

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

Arif Albayrak and William Teng

NASA Goddard Space Flight Center (ADNET Systems, Inc.)

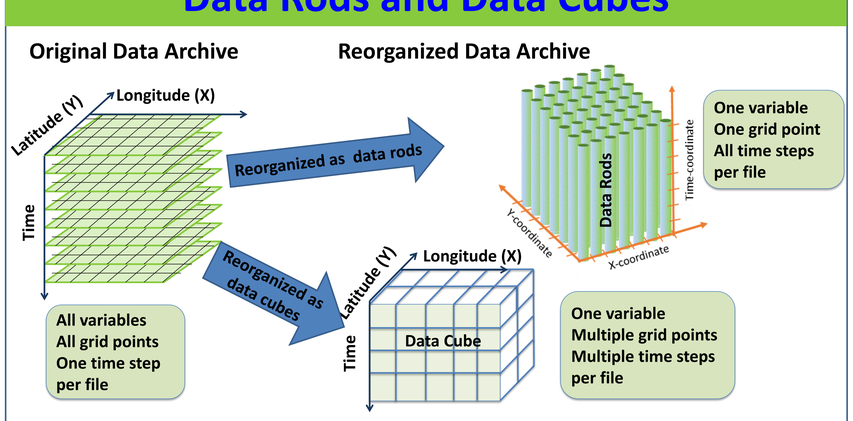
Emails: Arif.Albayrak@nasa.gov, William.L.Teng@nasa.gov

Abstract

As part of an ongoing NASA-funded project to remove a longstanding barrier to accessing NASA data (i.e., accessing archived time-step array data as point-time series), for the hydrology and other point-time series-oriented communities, "data cubes" are created from which time series files (aka "data rods") are generated on-the-fly and made available as Web services from the Goddard Earth Sciences Data and Information Services Center (GES DISC).

Data cubes are data as archived rearranged into spatio-temporal matrices, which allow for easy access to the data, both spatially and temporally. A data cube is a specific case of the general optimal strategy of reorganizing data to match the desired means of access. The gain from such reorganization is greater the larger the data set. As a use case of our project, we are leveraging existing software to explore the application of the data cubes concept to machine learning, for the purpose of detecting water cycle extreme events, a specific case of anomaly detection, requiring time series data. We investigate the use of support vector machines (SVM) for anomaly classification. We show an example of detection of water cycle extreme events, using data from the Tropical Rainfall Measuring Mission (TRMM).

Data Rods and Data Cubes



Data needs for machine learning

In order to apply Machine Learning algorithms for detecting extreme events, data have to be brought to a specific format, as follows:

$$\begin{pmatrix} y_1 & x_{11} & x_{12} & \cdots & x_{1n} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_k & x_{k1} & x_{k2} & \cdots & x_{kn} \end{pmatrix},$$ Where y_k are labels and x_{kn}, e.g., $(x_{11}, x_{12}, x_{13}, \ldots, x_{1n})$, are the feature vectors.

Input data need to be in spatio-temporal space, which is the case for most data in NASA data archives. However, most data, as archived, are in the form of time-step arrays. Thus, for one year of daily data, the search algorithm needs to be applied to 365 files of MxN matrices, for co-location purposes. Such a search is inefficient, and, thus, a more efficient format is needed.

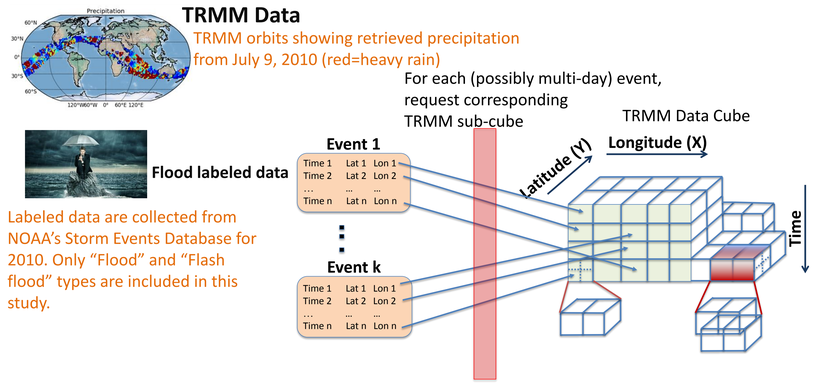
Comparison of different methods of data organization (for one year's data)

	Cons	Pros	Notes
As-archived files	Time series requires opening up to 365 files.	Fast search in spatial space	T number of time steps implies T number of files.
Data rods	Spatial search requires opening up to MxN files.	Fast search in temporal space	Each grid "point" corresponds to one file.
Data cubes	Larger file sizes	Fast search in spatial and temporal spaces. If memory allows, fast calculations for specific applications	Each file has the TxMxN size.

Leveraging Existing TRMM Data Cubes

Data collection and formatting

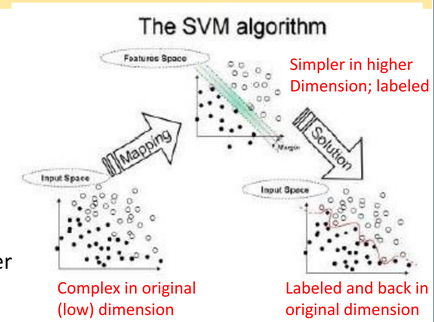
TRMM precipitation data are used to retrieve anomalies in the time series. For each recorded flood event, a request is made to the TRMM data cube. The retrieved sub-cube is then labeled as flood/no flood, based on NOAA's Storm Events Database.



Application to Detection of Extreme Events

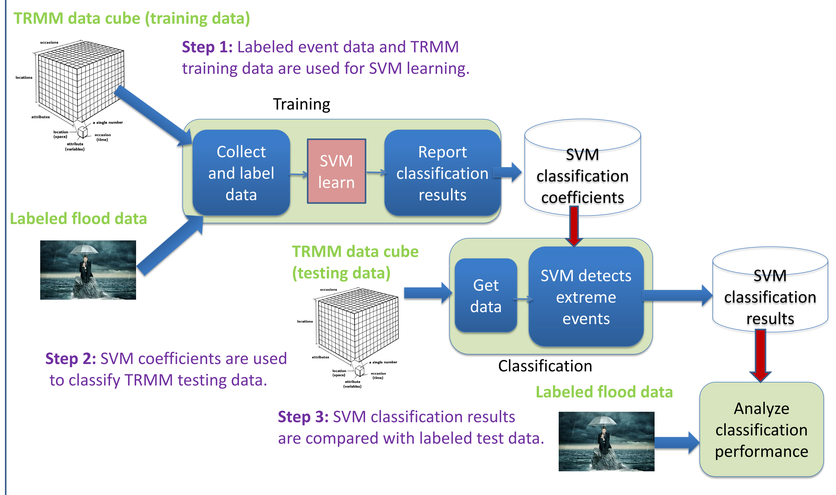
Methods for anomaly detection: SVM

A support vector machine (SVM) is a computer mathematical algorithm that learns by example to assign labels to objects (Noble, 2006). It is a supervised learning method that analyzes data and recognize patterns. An SVM constructs a hyperplane or a set of hyperplanes in a higher-dimensional space, which can be used for classification, regression, or other tasks.



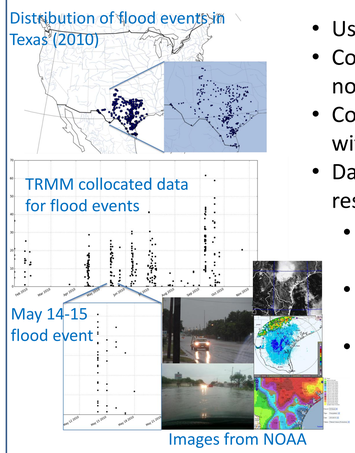
Training and testing process for extreme events (flood detection)

Step 1. SVM training: Use TRMM data collocated with labeled event data.
Step 2. SVM testing: Use coefficients from Step 1 to test data classification.
Step 3. SVM classification: Compare results to original labels (events).



Results

Experiment (SVM) results



- Used Texas state events.
- Collected 401 flood labeled event data and ~ 400 no flood but rain events.
- Collocated corresponding TRMM precipitation data with flood labeled data.
- Data divided 90% and 10%, for training and testing, respectively.
  - ~80 randomly chosen data points used for testing, and 720 points used for training.
  - With training data, correctly classified 699 event/no-event out of 720 data points (~3% loss).
  - With testing data, correctly classified 73 event/no-event (~9% loss).

Data cube performance

Experiment s performed on:

- OS: Linux
- Memory: 12 GB
- Speed: 3.2 GHz
- File type: netCDF4

Data cube for TRMM (2010)

- lat 400x1 (float32)
- lon 1440x1 (float32)
- time 2920x1 (float32)
- precip data 2920x400x1440 (float32)

Action	Variables	Time to load the data	Memory limitations
Load entire data matrix	lat, lon, time, precipitation	real 51.634s user 5.606s sys 12.662s	Requires almost all the memory; cannot perform other jobs
To load subset of precipitation data; indexing is required.	lat, lon, time	real 0.216s user 0.071s sys 0.027s	None
After indexing is completed, opening of 3 examples vectors	1. lon[1000] (lon point) 2. time[:] (time array) 3. precipdata[:,350,1000] (entire column data)	~ 7.59s	None

Summary

- With data cubes, we can create problem-specific smart queries. Indices can easily replace date, lat, and lon fields; which allow fast sub-gridding.
- For more complicated problems such as supervised learning, data have to be in a specific format. In order to complete the task, we can access multiple data cubes (of different variables) at the same time and attach labels to each column of the spatio-temporal type.
- Data cubes in netCDF4 (HDF5) formats provide storing capabilities that minimize memory requirements.
- Because of high intensity nature of precipitation in Texas, our study showed that it is relatively easy to classify an event as flood or no-flood. So, SVM is perhaps an overkill in this specific case. However, for other regions of the U.S., the benefits of this method should be more clear.

References

Noble W.S., 2006. What is support vector machine, Nature BioTechnology, 24, 1565, doi 10.1038/nbt 1206-1565.

Albayrak A. and W. Teng, 2014. Estimating rainfall for index-based agricultural insurance, in Proc. ASPRS 2014 Annual Conference, Louisville, KY.

Blackwell W. and F. Chen, 2009. Neural networks in atmospheric remote sensing, Lincoln Lab. ISBN 978-1-59693-372-9.

Acknowledgment: This work is partly supported by NASA ROSES NNH13ZDA001N-ACCESS.