University of Saskatchewan

Visualization, Geometry & Algorithms Lab

Towards a Machine Learning Approach for the Prediction and Compression of WRF Datasets

Jarin Tasnim, Gazi MD Hasnat Zahan, Dr. Carl Gutwin, Dr. Kevin A. Schneider, Dr. Debajyoti Mondal
Department of Computer Science, University of Saskatchewan
jat923@usask.ca, gazi.hasnat@usask.ca, gutwin@cs.usask.ca, kevin.schneider@usask.ca, dmondal@cs.usask.ca

ABSTRACT

The challenges of storing, retrieval and analysis of big weather datasets from cloud storage poses a great barrier to researchers and decision makers. Efficient data compression techniques can help in mitigating storage constraints. One approach is to maintain a precomputed summary in a local machine and generate enough data using prediction models, which allows users to visualize data in a small machine, even when the user is offline. To achieve this goal, we first designed features and corresponding machine learning models, and then analyzed how well the models can approximate the visualization compared to the visualization computed from the real dataset (WRF model output).

We built two models: The first model learns from ALBEDO, EMISS, GRDFLX features over a geographical area and predicts the SOIL MOISTURE. The second model learns from four features, where we assume that some SOIL MOISTURE data is missing. Thus the first (second) model works with all (resp., 50%) SOIL MOISTURE data missing, and reduces the storage space requirement by 25% (resp., 12.5%). The predicated visualization was reasonable approximation to the original.

MOTIVATION

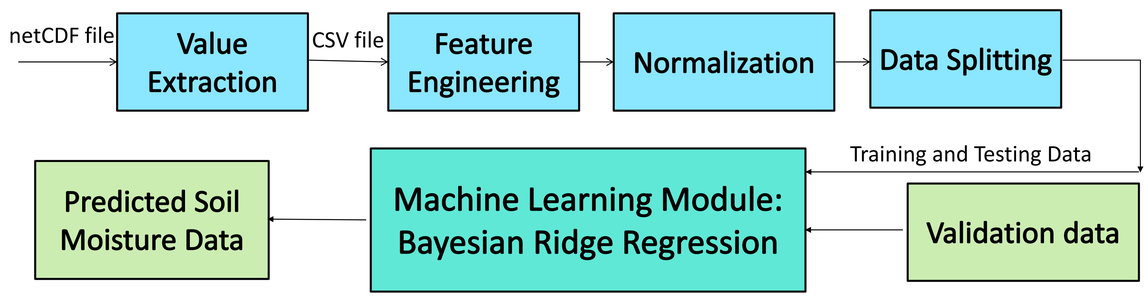
The climate change is a growing concern, which also poses a challenge to manage and preserve the global water resources. Developing robust models that can predict the weather changes is important to cope with such challenges [1].

The technological advancement in data gathering, storing, monitoring, analysis and prediction will be useful towards the management and protection of the natural resources. The concept of flawless disaster warning is one of the ultimate objectives. GWF's plan to create Canada's first national water disaster warning system by breaking knowledge gaps and technology barriers is where we are headed.

What happens after the prediction is done and decision has been taken? There is a big question of how smooth the management strategies will run when these situations arise. There is a need of adaptive scenario based risk-management tools to tackle such events.

Each of the datasets are large (gigabytes) and kept in the cloud storage. They are troublesome to download and often to work with in a resource driven environment. This motivated us to create a model that can learn enough to produce accurate synthetic data that can simulate the overall geographic process to be used ubiquitously and on the fly.

MODEL ARCHITECTURE



MODEL 1: DATA PREDICTION

Each feature vector at a latitude and longitude pair containing 3 variables: ALBEDO, EMISS and GRDFLX along with 8 of their neighbors: total 9 values each. No SOIL MOISTURE value was given while training the model. The training and testing datasets had data of 15 and 5 days of the year 2014 respectively. The objective was to predict SOIL MOISTURE of all the points of the dataset.

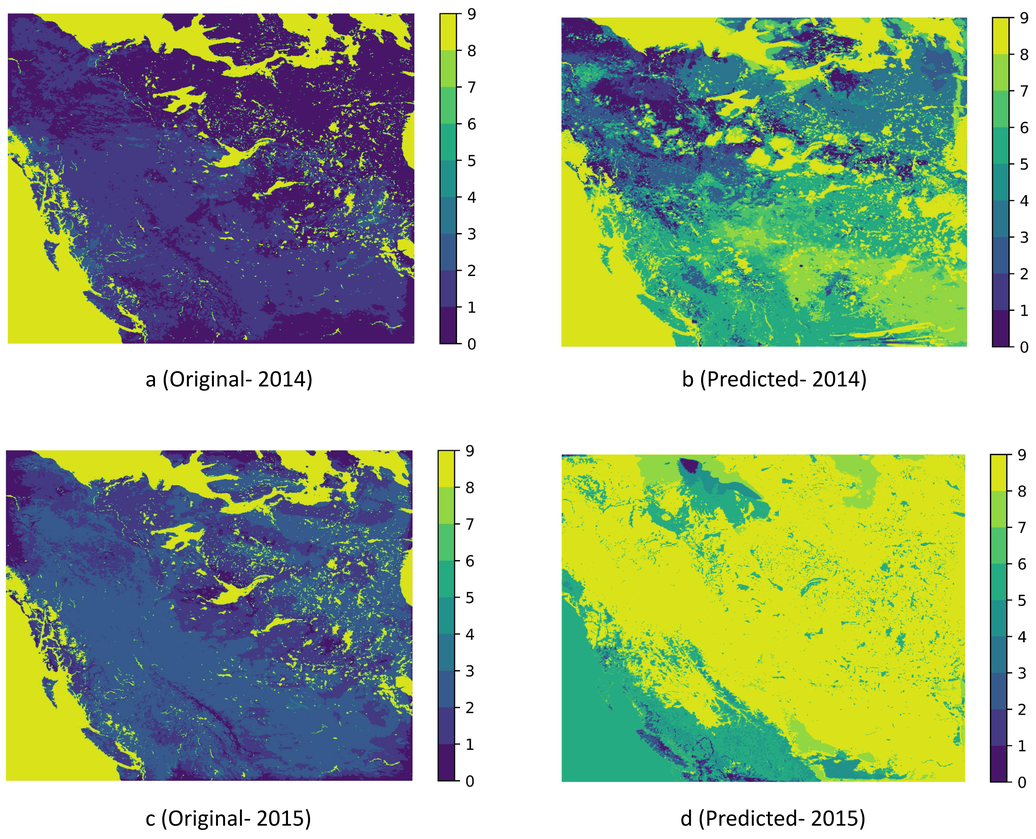


Figure 1: Original Data in the left column (a, c) with the predicted ones in the right (b, d)

MODEL 2: DATA COMPRESSION

Only odd rows and columns were taken as feature vectors, even ones were discarded as the step of data compression.

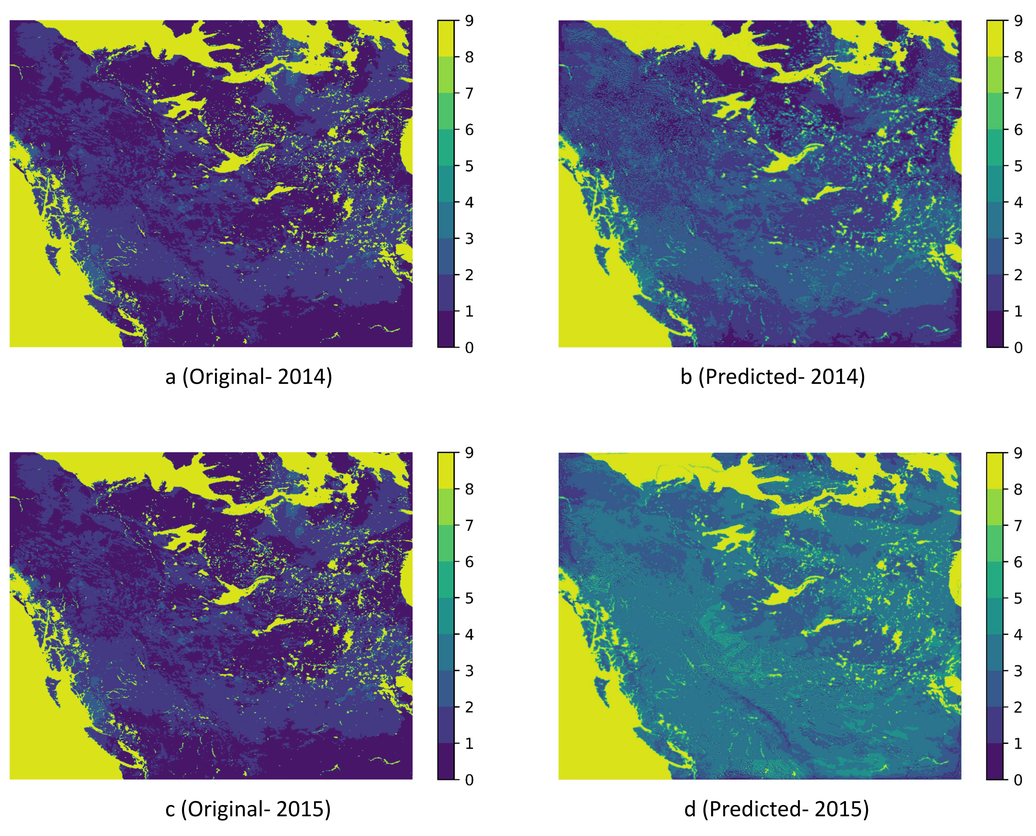


Figure 1: Original Data in the left column (a, c) with the predicted ones in the right (b, d)

OBSERVATION

For evaluating model performance, we considered the coefficient of determination. The best possible score of Coefficient of determination is 1.

The coefficient of determination of Model 1 is 0.68. As shown in figure 1 (b, d) the color mismatch occurs throughout the whole prediction area.

In model 2, 50% of the soil moisture data was used in training. That's why the prediction is expected to be better than model 1. Coefficient of determination of Model 2 is found to be 0.9166 which is reasonably good. The color match throughout the overall area of figure 2 (b, d) shows promising prediction result. The variation of SOIL MOISTURE in different areas in the predicted contour plots in Fig. 2 (b, d) are more similar to the original data than that of the model 1. The model was trained using 2014 data and yet it could predict data from 2015 though the success was not up to expectation because of less number of variables used in learning.

FUTURE WORKS

The data that was used for training the models was from western Canada though the location information have not been fed to the model. Therefore, we will try to predict the weather parameters over the whole Canada regardless of the location information.

We will implement recursive reduction of the data set. Model will be trained to predict missing values some of the data. With these predicted values, the model will be able to predict more internal missing points. This process will repeat and thus the whole dataset can be predicted from a small amount of summary data. The challenge of data prediction will be coupled with the advantage of data compression; the future avenue we want to explore.

We have terabytes of data available for analysis which is tough to analyze. We have already worked with data by significantly reducing the soil moisture variable. In future we will attempt to reduce of all data variables, targeting a reduction from terabytes to megabytes.

The training and the prediction datasets were only using 2014 and 2015 data, whereas we have data from 2008 to 2015. We will test and train our data over the longer period of time to gain better insight.

REFERENCES

[1] Adam, K., Majid, M. A., Fakherldin, M. A. I., & Zain, J. M. (2017). A Big Data Prediction Framework for Weather Forecast Using MapReduce Algorithm. Advanced Science Letters, 23(11), 11138-11143.

[2]MacKay, D. J. (1992). Bayesian interpolation. Neural computation, 4(3), 415-447.