Digital Typhoon: Large-Scale Heterogeneous Data Archives for Machine Learning Research

Asanobu KITAMOTO
National Institute of Informatics (NII), Japan

1 Introduction

Digital Typhoon is a project to organize the world’s typhoon information and make it universally accessible and useful. We have been collecting and organizing typhoon information for more than a decade, and have built information infrastructure to search and analyze typhoon information. The main data collection, among others, is geostationary satellite (Himawari series) imagery archived since December 1978. From a full disk image taken by satellites, we built a “typhoon image collection,” which consists of about 134,000 images of typhoons since 1978. The uniqueness of this image collection is in fixed framing and map projection, satisfying properties as follows; 1) the center of the image is always aligned with the center of the typhoon based on Japan Meteorological Agency (JMA), 2) the map projection method is chosen so that the size of the image is always constant, namely 2600km on the earth, and 3) the distortion of shape is minimal for a rounded shape like typhoons. This image collection is one of the largest collections in the world, with additional 35,000 images for Australian cyclones.

This image collection has potential to be used for machine learning research as a testbed for various types of algorithms for the following reasons. Firstly, each typhoon image is made by a systematic approach and come with metadata such as location and intensity given by JMA. Secondly, we also developed heterogeneous typhoon data collection such as ground-based observation data, mass media news, social media text (still small), disaster data, and numerical weather prediction data. All data can be linked by various keys such as date and location, allowing machine learning researchers to dig into data collection from many viewpoints. Thirdly, most of the data is accessible on our websites, Digital Typhoon (http://www.digital-typhoon.org/) and Vertical Earth (http://earth.nii.ac.jp/), and users can freely browse the data to search the dataset. Those websites have attracted about 150 million page views since the first opening in 2003, now considered as one of the most useful real-time data sources for typhoons.

2 Approach

We already tried several approaches for the analysis of typhoon image collection. Firstly, we applied content-based image retrieval for searching typhoons based on the similarity of cloud patterns [1, 2]. Image similarity is defined using principal component analysis (in the meteorological terminology, empirical orthogonal function), which is chosen not as the best method, but as a simple method to demonstrate the applicability of content-based image retrieval. We developed an image search engine for hierarchical matching of time-series image sequences, and also implemented a simple query language for searching typhoons by various conditions. The demonstration is available on the website above.

Secondly we focused on detecting rapid intensification of typhoons using support vector machines (SVM) [3]. Early warning of rapid intensification is an important issue for disaster reduction, because this is the most dangerous event for typhoons. We applied SVM for the classification of rapidly-intensifying typhoons against otherwise using cloud patterns of typhoons, and the result showed that the error rate of misclassification was about 25%. Unfortunately, this was not a useful result for an operational use.

Thirdly we tried to understand the evolution of the typhoon using clustering algorithms for time-series [4]. The idea is to characterize the life cycle of typhoons using “prototype cloud patterns” obtained from clustering of typhoon image collection, and state transition between them. We applied k-means clustering, self-organizing maps (SOM), and mixture of Gaussians, and generative topographic mapping...
(GTM) with hidden Markov model (HMM). The result showed that spatio-temporal clustering methods outperformed spatial clustering methods in capturing the temporal structures of the evolution of typhoons.

3 Research Directions

As mentioned above, we studied several image processing and mining algorithms about 10 years ago as a research of what we called “meteoinformatics.” Digital Typhoon project itself is still ongoing, but our focus shifted from machine learning for images to the analysis of other types of data, such as sensor data, textual data, and others. Another topic has been to make information infrastructure for real-time information dissemination and real-time information collection from various sources such as social media. This is why our research on typhoon image collection has not made much progress in these 10 years, but we still believe that this dataset can contribute to typhoon analysis and prediction by machine learning.

Firstly, the dataset can be used for the reanalysis of historical data. This is important to understand how climate change in the past affected the statistics of typhoons, and this is also crucial for studying the change of typhoon statistics under global warming environment. Current statistics of typhoons, however, are obtained from “Dvorak method,” which is based on human judgment on typhoon observation data such as cloud patterns, and it is known to have a problem of reliability from subjective human judgment and historical changes on the method and available data. Although important, reanalysis of past data is a tedious task, so machine learning can be used to help this task.

Secondly, the dataset can be used for improving the forecasting of typhoons. One of the problems in the simulation of typhoons is to set an appropriate initial condition when insufficient observation data is available, especially when a typhoon is in the middle of the ocean. The quality of an initial condition significantly affects the result of forecasting, but the only way to get observation data is to actively fly into the typhoon, especially to a sensitive area for forecasting. Here historical data collection might help to improve the initial condition by understanding the characteristics of the typhoon.

Thirdly, the dataset can be used for effective warning of typhoons. For the general public, absolute values, such as the central pressure of a typhoon in hecto-pascal, or millimeter of rainfall, are difficult to understand. On the other hand, an analogue-based warning, such as “this typhoon is the strongest in history,” or “this typhoon is similar to a famous typhoon in the past,” can facilitate the understanding of the risk. This requires searching similar events in the database using the current situation as a query (search-by-situation) and this is where multimedia search technology can help to provide useful information quickly in emergency situations.

Last but not the least, all ideas discussed in this paper can be applied and extended directly to hurricanes or tropical cyclones in the world. Our datasets are limited to typhoons (plus Australian cyclones) due to the availability of the data. Extension of our approach to the worldwide tropical cyclone database is an interesting future direction of research.

Acknowledgements

This work is partially supported by PRESTO program, Japan Science and Technology Agency (JST).

References


