

## Algorithm Development and Mining (ADaM) System for Earth Science Applications

Rahul Ramachandran \*, Helen Conover, Sara Graves and Ken Keiser

Information Technology and Systems Center  
University of Alabama in Huntsville  
Huntsville, AL - 35899

### 1. INTRODUCTION

Data Mining has an enormous potential as a processing tool for earth science data. It provides a solution for extracting information from massive amounts of data. However, designing a data mining system for earth science applications is complex and challenging. The two key issues that need to be addressed in the design are (1) variability of data sets and (2) operations for extracting information. Data sets not only come in different formats, types and structures; they are also typically in different states of processing such as raw data, calibrated data, validated data, derived data or interpreted data. The mining system must be flexible to handle these variations in data sets. The operations needed in the mining system vary for different application areas within earth science. Operations can range from general-purpose operations such as image processing techniques or statistical analysis to highly specialized data set-specific science algorithms.

The Algorithm Development and Mining (ADaM) system, developed at the Information Technology and Systems Center at the University of Alabama in Huntsville, is one such mining tool. The system provides knowledge discovery and data mining capabilities for data values, as well as for metadata, and catalogs the information discovered. It contains algorithms for detecting a variety of geophysical phenomena to address the needs of the earth science community. This paper describes this system and its applications to different projects in earth science.

### 2. ADaM FEATURES

ADaM was developed in response to the need to mine large scientific data sets for geophysical phenomena detection and feature extraction. It provides a variety of processing tools, which allow easy integration of spatial and temporal variables of earth science data sets.

Algorithms that detect a variety of geophysical phenomena were added to the system to address specific needs of the earth science community

#### 2.1 Design

ITSC developers applied the latest object oriented software design concepts while developing the ADaM system. The key features of the design of the system are:

- **Portability** - In order to realize a high degree of platform independence, the system was written using widely available, standard tools such as the C++ programming language and ANSI C libraries. This has allowed the Center to develop the same system for Windows and UNIX operating systems.
- **Network Accessibility** - The client-server architecture of the system allows ease in network accessibility. This feature of the system allows it to be used as an application at a data archiving center or on a user's desktop workstation.
- **Extensibility** - The ADaM system consists of three basic types of modules: input filters (readers for different data formats), processing modules (general-purpose algorithms and user-defined algorithms) and output filters (writers for different data formats). Since the number of data sets and algorithms for analysis continues to increase and change, the system was designed to be extensible by the use of plug in modules. New modules can be added to the system easily. Thus, the system is designed to evolve along with the demand.

#### 2.2 Processing Architecture

The ADaM system architecture utilizes a data pipeline approach. Mining is broken down into a series of steps with results from each step passed to the next one in line. Figure 1 illustrates both ADaM's data processing stream, as well as the three basic types of modules: input, processing, and output. The use of data input filters, specialized for a variety of data types, has been instrumental in simplifying the development of the processing and output operations. The selected input filter translates the

---

\*Corresponding author Address:  
Rahul Ramachandran  
Information Technology and Systems Center  
University of Alabama in Huntsville,  
Huntsville, AL 35899  
email: ramachan@itsc.uah.edu

data into a common internal structure so that the processing operations can all be written for a single data representation. This allows the addition of new operations to the system without having to address input data format problems. Similarly, the addition of a new input filter provides access to the entire suite of processing operations for the data type in question. The mining system currently allows over 100 different operations to be performed on the input data stream. These operations vary from specialized atmospheric science data set specific algorithms to generalized image processing techniques. The last step in the mining process is the selection of the output format. Since the input data has been converted to ADaM's internal format, the output modules allow the user the option to select either the input format or a different format for the final data product. In the same manner as the input modules, the output filters effectively insulate the processing operations from having to support all the possible output formats. Details about the ADaM System can be found in Keiser et al (1999).

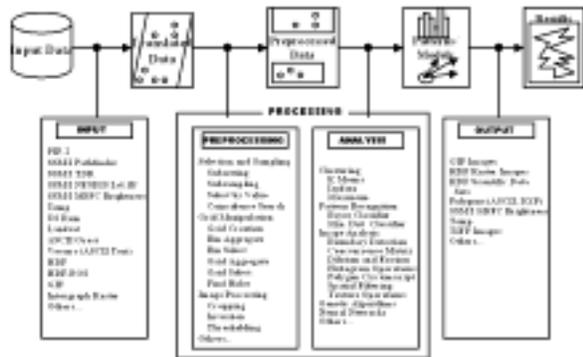


Figure 1: Schematic diagram depicting the processing architecture of the mining system

### 3. TYPES OF MINING PROBLEMS

There are three types of mining problems. They are:

**Type 1: Known Phenomenon, Known Algorithm:** The users know exactly what they are looking for and what sets of algorithms to use. The mining plan is then set up to perform those operations over the data sets of interest. The Mesoscale Convective System (MCS) detection using the SSM/I data set is one such example of known phenomenon/known algorithm mining performed at ITSC (Hinke et al (1997)).

**Type 2: Known Phenomenon, Learned Algorithm:** This is known as an algorithm development approach where the end user knows what phenomenon to target but is unsure of the characteristics of the phenomenon or what sequence of algorithms to apply. This is iterative in nature and allows end users the flexibility of fine tuning their algorithm for the event of interest. Cloud mask detection using GOES-8

data is an example of known event, learned algorithm. Details about this product are described later in the paper.

**Type 3: Search for Unknown Relationships:** This category is called target independent data mining. The user searches the data sets for transient events with thresholds. The main idea is to mine for trends depicting anomalies in the data sets. This approach could be used to check the quality of the data set by searching for spurious values or just mining for rare phenomena which would be hard to detect because of the size of the data set. Hinke et al (1997) describe such an application of target independent data mining and the impact in data reduction while retaining all the transient phenomena information.

Mining can occur at various times in a data production workflow. The time chosen depends upon the application. The possible times include:

**Real Time:** In some cases data mining must occur in real time in order for the data to be of any use in the application. This is a common requirement for applications that predict future events such as severe weather.

**On Ingest:** Certain types of data mining are always performed on data to make it useful. Such mining or processing could include navigation, quality control, etc.

**On Demand:** Mining is done to provide custom order processing for the end user. Different users may want gridding, subsetting, subsampling and other operations to be performed to suit their specific needs.

**Repeatedly:** The mining system can be used for fine tuning algorithms for scientific analysis. Thus, multiple passes would be required for scientific discovery and exploration through the data as the algorithms are developed and refined in an iterative fashion.

### 4. EARTH SCIENCE APPLICATIONS

ADaM has been utilized in a variety of earth science applications. A short summary of some of these different applications is given below:

#### 4.1 Cumulus Cloud Detection

An accurate cloud mask product is required in many different areas of atmospheric science for variety of reasons. The accuracy of the earth radiation budget estimates derived from satellite-based measurements is highly dependent on the variability of cloud cover. Thus, cloud cover is one of the most important variables affecting the radiation

balance and the global climate. For certain retrieval algorithms, detection of cloud pixels is vital. If the scene is contaminated with cloud pixels the algorithms give spurious results. Figure 2 illustrates the mining scheme used in creating the cumulus cloud mask.

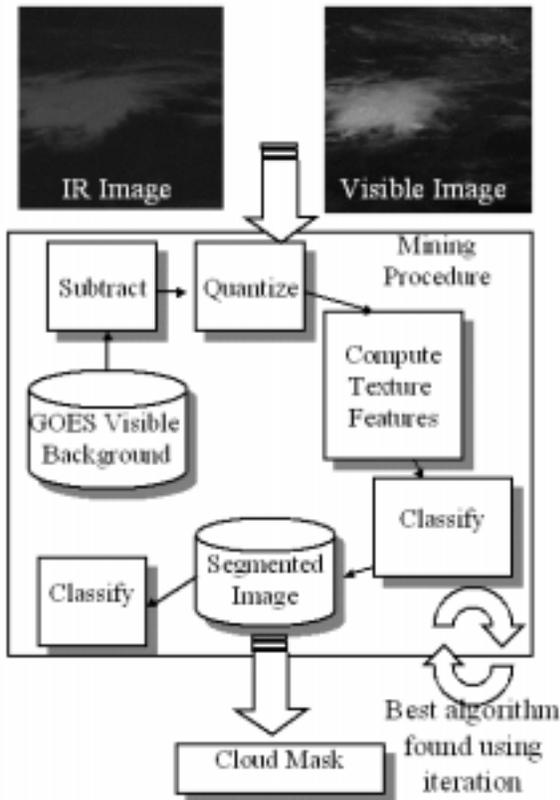


Figure 2: Mining scheme utilized to create cumulus cloud mask

Nair et al (1999) describes the importance of detecting cumulus cloud fields in satellite imagery. Since land use change has significant impact on the formation of cumulus convection, the variation due to land use change becomes an important aspect to monitor on temporal scales to assess human impact. With these factors as a driving force, a general-purpose cloud mask was created using ADaM. ADaM's different image processing techniques were applied to Geostationary Operational Environmental Satellite (GOES) data iteratively until the best set of algorithms were reached that gave optimum results and performance times. Boundary layer cumulus clouds over land are difficult to detect in satellite data, due to low contrast in both visible and infrared channels. For GOES satellite data the problem becomes severe, as the infrared channel resolution is 4km, compared to 1km in the visible channel. A study was conducted analyzing three of the different image processing and pattern recognition techniques available in ADaM for cumulus cloud detection. These

were classifiers based on 1) texture and spectral features, 2) edge detection and spectral features, and 3) purely spectral features.

### 3.2 Phenomena Detection

The ability of ADaM to search or mine for particular data values or geophysical phenomena within a specified data product has been actively utilized on several projects. The Data Mining Center at the Information Technology and Systems Center daily mines EOS Special Sensor Microwave/Imager (SSM/I) Brightness Temperature Swaths from DMSP F13 and F14 satellites for the purpose of detecting MCS events globally. The Global Hydrology Resource Center (GHRC) gets the SSM/I data from the Fleet Numerical Meteorology and Oceanography Center (FNMOC) through the National Environmental Satellite and Data Information Service (NESDIS). Figure 3 illustrates the processing scheme for extracting MCS from SSM/I data sets.

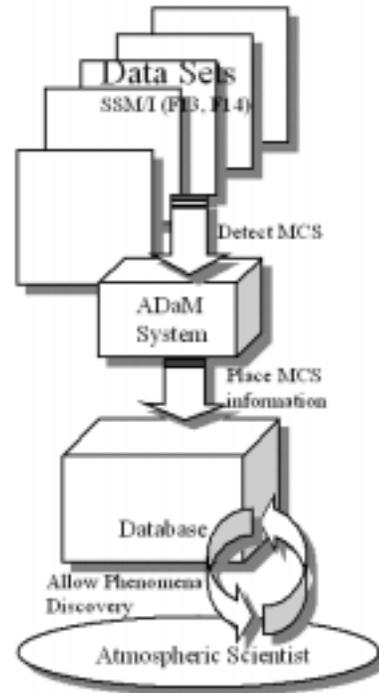


Figure 3: Process of Mining and storing MCS from SSM/I data-set

The data are processed at the GHRC within hours of its reception to produce full resolution "swath" brightness temperatures (Tb). A compressed copy of the SSM/I swath data is electronically transferred to the ITSC's Data Mining Center where the data is mined for MCS events. Special Sensor Microwave/Imager (SSM/I) data is being mined during ingest to detect and locate Mesoscale Convective Systems. An algorithm developed by Devlin (1995) is used for detecting MCS events from

SSM/I data. The mined MCS's are placed in a database to provide a powerful discovery system which combines the results of the ADaM system with a data management engine to provide scientist with a tool to extract phenomena of interest from large data archives.

In a another project, Advanced Microwave Sounding Unit (AMSU) data is mined in real time to locate tropical storms and estimate their maximum wind speeds. The Advanced Microwave Sounding Unit is a microwave radiometer that can be used to detect temperature at different levels of the atmosphere. Based on gradients in temperature measurements in a given area, it is possible to estimate maximal sustained radial wind speed. This wind speed estimate can be combined with other factors such as ice scattering and moisture in order to detect tropical cyclones.

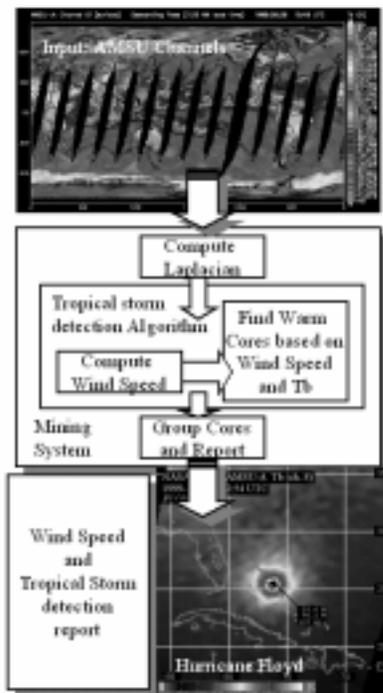


Figure 4: Mining procedure for detection tropical storm and estimating wind speeds

Algorithm described by Spencer et al (1999) is utilized for detecting tropical cyclones and estimating their maximum sustained wind speed. The ADaM data mining system is used to implement this algorithm in near real time production mode. In addition, an image is produced for the AMSU channel 8 brightness temperatures by the same mining application. This image is combined with the text file containing the events to produce an integrated visual result. The processing has been automated and is currently being done as soon as the AMSU data becomes available. This process is illustrated in Figure 4. Results can be viewed on the web at <http://pm-esip.msfc.nasa.gov/cyclone>.

## 5. SUMMARY

The volume of raw data being stored by different earth science archives today defies even the partial manual examination by scientists. Projections have been made that EOS (Earth Observing System) data volumes will reach a terabyte/day by the time all the planned satellites are flown. ADaM and other similar tools will play an important part in extracting valuable information from these large data stores. Continued and strengthened interaction between earth scientists and data mining experts is vital for the successful use of these tools and techniques.

## 6. REFERENCES

- Devlin, K, 1995: Application of the 85 GHz ice scattering signature to a global study of mesoscale convective systems. Master's thesis, Texas A&M University, August 1995.
- Hinke, T., J. Rushing, S. Kansal, S. Graves, H. Ranganath and E. Criswell, 1997. Eureka Phenomena Discovery and Phenomena Mining System. *13<sup>th</sup> International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology*, February 1997.
- Hinke, T., J. Rushing, H. Ranganath and S. Graves, 1997: Target-Independent Mining for Scientific Data: Capturing Transients and Trends for Phenomena Mining, *Proc. Third Int. Conf. On Data Mining (KDD-97)*, Newport Beach, CA, Aug. 14-17, 1997.
- Keiser, K., J. Rushing, H. Conover and S. Graves, 1999. Data Mining System Toolkit for Earth Science Data. *Earth Observation and Geo-Spatial Web and Internet Workshop (EOGEO)-1999*, Washington, Feb 9-11.
- Nair, U.S., J. Rushing, R. Ramachandran, K. S. Kuo, S. Graves and R. M. Welch, 1999: Detection of Cumulus Cloud Fields in Satellite Imagery. Submitted to *The International Symposium on Optical Science, Engineering, and Instrumentation*, 18-23 July 1999, Denver, Colorado.
- Spencer, R.W., and W.D. Braswell, 1999: Tropical Cyclone Monitoring with AMSU-A: Estimation of Maximum Sustained Wind Speeds. *Mon. Wea. Rev.*, submitted.