Analysis of Common data form using Machine Learning Mechanism

Priyanka Bhurkunde
Sinhgad College Of Engineering, Pune
Pune, India
psbhurkunde1917@gmail.com

K.S. Korabu
Sinhgad College Of Engineering, Pune
Pune, India
kskorabu.scoe@sinhgad.edu

***Abstract*— Scientific organizations are creating a vast data store which is very useful for future predictions. With the increasing importance on analysis of large scale scientific data, and with growing dataset sizes, number of new challenges increases. Scientific communities' data are commonly stored in multidimensional arrays, but owing to the evolution of scientific instruments and simulations, scientists are facing the problem of data flood. Particularly, different data accessing solutions are needed. Such array-based scientific data is accessed by using a machine independent and portable interface known as Network Common Data Form (NetCDF). Visualization is an increasingly important activity for understanding complex geosciences data, and for communicating results to a variety of audiences. Machine Learning provides an environment for analysis and visualization of such multidimensional data. The aim of proposed system is to develop an effective system for analysis and visualization of scientific data with the use of machine learning technique.**

Keywords- multidimensional data, NetCDF, Machine Learning, visualization.

I. INTRODUCTION

Analysis of data is a process of inspecting, cleaning, transforming data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data is collected and analyzed to answer questions, test hypotheses or disprove theories. Once the data is analyzed, it may be reported in many formats to the users of the analysis to support their requirements. The users may have feedback, which results in additional analysis. When determining how to display the results, the analysts may consider data visualization techniques to help clearly and efficiently communicate the message to the audience in the form of information graphics. The data, in graphical format, helps to obtain additional insight regarding the messages within the data. A wide range of application software has been written which makes use of netCDF files. NetCDF is an Application Programming Interface (API) that provides methods for accessing array-oriented data and a freely-distributed collection of software libraries for C, FORTRAN, C++, Java, and Perl that implement this interface. The obtained portable dataset can be used to train the application by using machine learning techniques [9][12][13]. Machine learning is a scientific discipline that deals with the construction and study of algorithms that can learn from data. In Statistical Data Analysis, the main focus on getting insight into data, understanding complex phenomena through partial observations, and making informed decisions in the presence of uncertainty. In Machine Learning, analysis and processing of data is achieved using statistical methods. However, the goal is not necessarily to understand the data, but to learn from it.

This paper is organized as follows: Section II describes Literature survey in which NetCDF data model and its components and Machine Learning techniques are discussed. Section III consists of Proposed system. The conclusions are presented in section IV.

II. LITERATURE SURVEY

A. NetCDF Data Model

NetCDF (Network Common Data Format) are not Database Management Systems. Relational database system is not suitable for complex large data access. First, existing database systems that support the relational model do not support multidimensional (arrays) or hierarchical objects as a basic unit of data access [1]. A quite different data model is needed for such data to facilitate its retrieval, modification, mathematical manipulation and visualization. Related to this is a second problem with general-purpose database systems: their poor performance on large objects. Collections of satellite images, scientific model outputs and long-term global weather observations are beyond the capabilities of most database systems to organize and index for efficient retrieval. Finally, general-purpose database systems provide, at significant cost in terms of both resources and access performance, many facilities that are not needed in the analysis, management, and splay of array-oriented data [4].

B. COMPONENTS

A netCDF dataset contains dimensions, variables, and attributes, which all have both a name and an ID number by which they are identified. These components can be used together to capture the meaning of data and relations among data fields in an array-oriented dataset.

The netCDF library allows simultaneous access to multiple netCDF datasets which are identified by dataset ID numbers, in addition to ordinary file names. The classic netCDF data model uses dimensions, variables, and attributes, to capture the meaning of array-oriented scientific data. UML diagram below represents data models visually.

Data types: The NetCDF interface defines data types – char, byte, short, integer, float, and double. These types were chosen to provide a reasonably wide range of trade-offs between data precision and number of bits required for each value [10]. These external data types are independent of whatever internal data types are supported by a particular machine and language combination [11]. Existing methodologies are HDF (Hierarchical Data

Format) and HDF5. This is a physical file format for storing scientific data.

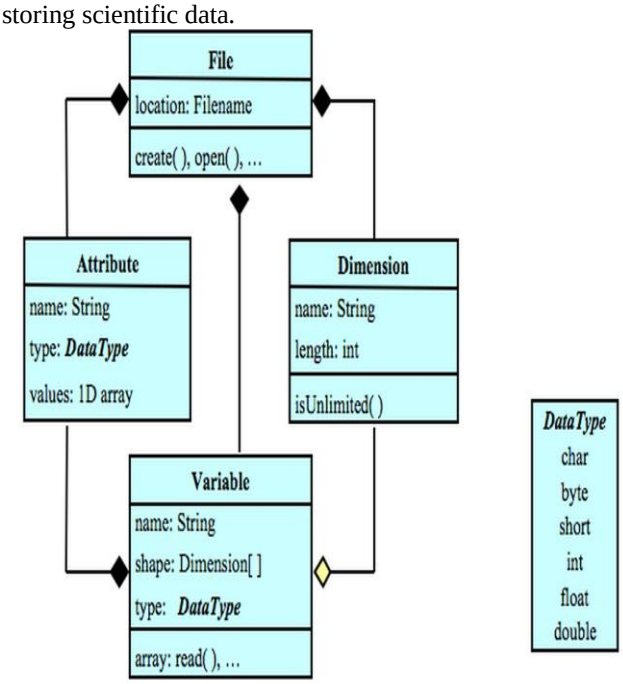

Figure 1: NetCDF data model

. At its highest level, HDF is a collection of utilities and applications for manipulating, viewing, and analyzing data in HDF files. Between these levels, HDF is a software library that provides high-level APIs and a low-level data interface. HDF5 is a new data format, designed to better meet the ever-increasing demands of scientific computing and to take better advantage of the ever increasing capabilities of computing systems [4]. HDF5 is a completely new Hierarchical Data Format product consisting of a data format specification and a supporting library implementation. HDF5 is designed to address some of the limitations of the older HDF product and to address current and anticipated requirements of modern systems and applications [17]. HDF5 is a unique technology suite that makes possible the management of extremely large and complex data collections [16]. One of the goals of netCDF is to support efficient access to small subsets of large datasets. To support this goal, netCDF [3] uses direct access rather than sequential access. This can be much more efficient when the order in which data is read is different from the order in which it was written, or when it must be read in different orders for different applications.

The HDF5 technology suite includes:

- A versatile data model that can represent very complex data objects and a wide variety of metadata.
- A completely portable file format with no limit on the number or size of data objects in the collection.
- A software library that runs on a range of computational platforms, from laptops to massively parallel systems, and implements a high-level API with C, C++, Fortran 90, and Java interfaces
- A rich set of integrated performance features that allow for access time and storage space optimizations.
- HDFView: A visual tool for browsing and editing HDF4 and HDF5 files
- HDF Java Products: All of the HDF Java Products, including HDFView and HDF Java wrappers.

C. Machine Learning Technique

Machine Learning (ML) Technique as a research discipline has roots in Artificial Intelligence and Statistics, and the ML techniques focus on extracting knowledge from datasets. This knowledge is represented in the form of a model which provides description of the given data and allows predictions for new data [2]. This predictive ability makes ML a worthy candidate for bioclimatic modeling [5]. Many ML algorithms are showing promising results in bioclimatic modeling including modeling and prediction of species distribution.

It may be noted that there is no universally best ML method; choice of a particular method or a combination of such methods is largely dependent on the particular application and requires human intervention to decide about the suitability of a method [2]. However, concrete understanding of their behavior while applied to bioclimatic modeling can assist selection of appropriate ML technique for specific bioclimatic modeling applications [6][7]. The inference mechanisms employed by Machine Learning (ML) techniques involve drawing conclusions from a set of examples and extracts knowledge representation from these examples to predict outputs for new inputs. The ML inference mechanism is depicted in Fig. 2

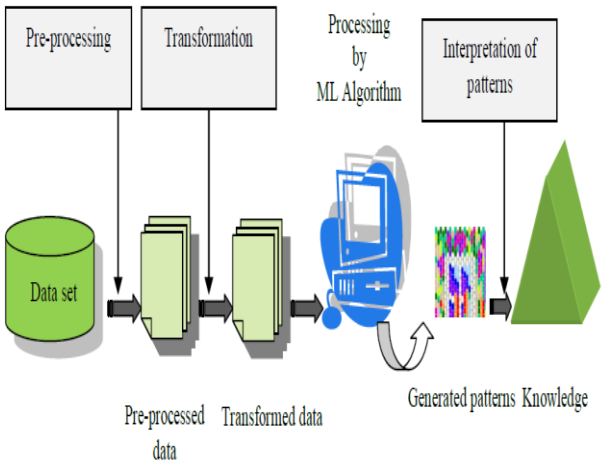

Figure 2: Machine learning inference mechanism

1) *Types of Machine Learning*:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised learning is one of the key ML inference mechanisms and is of particular interest in prediction of geographic ranges. In supervised learning the information about the problem being modeled is presented by datasets

comprising of input and desired output pairs. In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables.

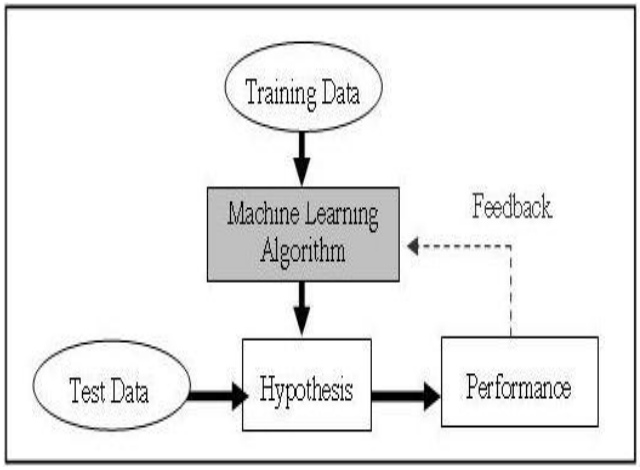

Figure 3: Supervised Machine learning

The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable as it is done in regression analysis. To apply directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set. Unsupervised learning is closer to the exploratory spirit of Data Mining as stressed in the definitions given above. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. Depending upon the NetCDF data extracted the suitable machine learning technique is used [15].

III. PROPOSED SYSTEM

The architecture is highlighted in Figure below. The system contains three modular components that allow users to perform model evaluations using remote sensing data from scientific and other agencies. The extracted information (latitude, longitude, time, value, height) forms a tuple. The data is made available via a spatio-temporal web service to NetCDF module.

The data are then temporally regridded in hourly, daily, or monthly fashion, and then available for metrics computation in Machine Learning environment [8].

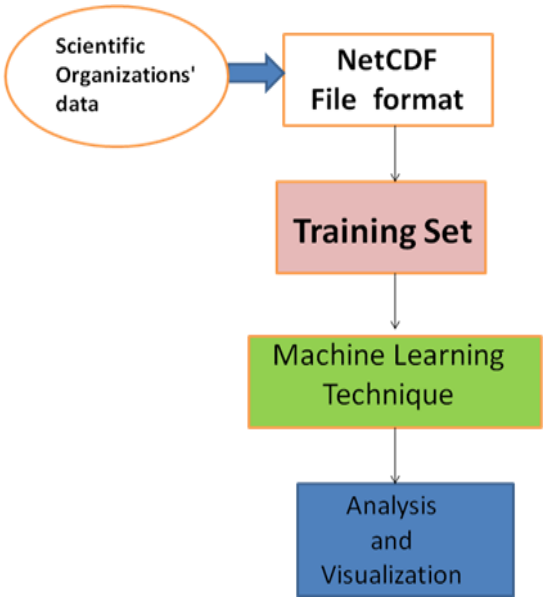

Figure 4: System Architecture

Various statistics including probabilistic distribution functions as well as user-defined metrics are computed. After metrics computation, the metrics can be visualized using Machine Learning methods. Machine Learning Technique is applied for which the training set is obtained from the analysis and computation and an automated system can be obtained.

IV. CONCLUSION

The proposed system aims to analyze data in different formats and to perform quality control if the data is not within certain standard ranges. The machine learning technique trains the application to get the future predictions.

Current trends in ML applications for geo- and environmental sciences deal with nonlinear dimensionality reduction and data visualization; analysis and modeling of data in high-dimensional geo-feature spaces; fast modeling of physical and other processes in hybrid models; spatio-temporal patterns/structures extraction, modeling and predictions.

ACKNOWLEDGEMENT

It is great pleasure for me to acknowledge the assistance and contribution of number of individuals who helped me in presenting “Analysis of Common data form using Machine Learning Mechanism” paper.

First and foremost I wish to record my gratitude and thanks to Mrs. K. S. Korabu for her enthusiastic guidance and help in successful completion of the paper. I would also like to thank Ms S. P. Potdar for her continuous support. I express my thanks to Mrs. K. S. Thakre, for her valuable guidance.

I would also extend my gratitude to honorable Mr. P. R. Sonawane, Sonix Nano System, Pune, for being a constant source of inspiration.

REFERENCES

[1] Unidata Program Center, 1991. NetCDF User's Guide: An Interface for Data Access. Unidata Program Center, Boulder, Colorado. 150 pages.

[2] Shouyi Wang, Student Member, IEEE, WanprachaChaovalitwongse, Member, IEEE, and Robert Babu˘ska, "Machine Learning Algorithms in Bipedal Robot Control",2012, vol 42, p 728 – 743.

[3] NetCDF home web page. http://www.unidata.ucar.edu/software/netcdf/

[4] Peter Barnum and VinithraVaradharajan, "When to Picnic?", The Robotics Institute, Carnegie Mellon University Pittsburgh, PA 15213.

[5] Yu Su Computer Science and Engineering, The Ohio State University, GaganAgrawal, Jonathan Woodring CCS-7, "Indexing and Parallel Query Processing Support for Visualizing Climate Datasets", Applied Computer Science Group-Los Alamos National Laboratory.

[6] R. Castroa, J. Vegaa, M. Ruizb, G. De Arcasb, E. Barrerab, J.M. Lópezb, D. Sanzb, B. Gonc¸alvesc, B. Santosc, N. Utzeld, P. Makijarvid.: NetCDF based data archiving system applied to ITER Fast Plant System Control Prototype.

[7] COARDS NetCDF Convention http://ferret.wrc.noaa.gov/noaa_coop/coop_cdf_profile.html.

[8] Yi Wang ; Wei Jiang ; Agrawal, G, "Improving Data Analysis Performance for High-Performance Computing with Integrating Statistical Metadata in Scientific Datasets"- Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium.

[9] R. Rew and G. Davis, "The Unidata netCDF: Software for Scientific Data Access," Sixth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography and Hydrology, Anaheim, CA, February 1990..

[10] PavelMichna and Milton Woods,"RNetCDF – A Package for Reading and Writing NetCDF Datasets",

[11] Rew, G. Davis, S. Emmerson, H. Davies, E. Hartnett, and D. Heimbigner. The NetCDF Users Guide, Version 4.1.3. Unidata Program Center, 2011

[12] Eaton, J. Gregory, B. Drach, K. Taylor, and S. Hankin. NetCDF Climate and Forecast (CF) Metadata Conventions, Version 1.6, 2011.

[13] "An XML-based Language to connect NetCDF and Geographic Communities", S. Nativi, L. Bigagli, B. Domenico, J. Caron, E. Davis, IEEE 2006.

[14] HDF4 Home Page. The National Center for Supercomputing Applications. http:// hdf.ncsa.uiuc.edu/hdf4.html.

[15] HDF5 Home Page. The National Center for Supercomputing Applications. http:// hdf.ncsa.uiuc.edu/HDF5/.

[16] Where is NetCDF Used? Unidata Program Center. http://www.unidata.ucar.edu/packages/netcdf/usage.html.

[17] T.J. Jankun-Kelly and K.-L. Ma, "A Spreadsheet Interface for Visualization Exploration," to appear in Proc. Visualization 2000 Conf., ACM Press, New York, Oct. 2000.