International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979, Impact Factor: RJIF 5.72
www.allsubjectjournal.com
Volume 3; Issue 6; June 2016; Page No. 192-195

Knowledge discovery and weather prediction of India using machine learning

Sujata Ashok Jadhav, Potdar SP

Department of Information technology Sinhgad college of Engineering Pune, India

Abstract
Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [17]. NetCDF is network common data format which is self-describing and converting various heterogeneous data into some common format or portable [4]. The proposed system mainly focuses on extraction of knowledge using NetCDF datasets with the help of well suited Machine Learning techniques. Machine learning techniques are better computing technologies continue to revolutionize the capability to solve larger and more complex problems in science and engineering. Machine learning optimizes a performance criterion using example data or past experiences or training datasets [7]. Moreover Machine learning for NetCDF datasets yields better statistical data prediction using R programming language.

Keywords: CDL, prediction, Scientific data, training datasets, knowledge extraction, NetCDF, R programming

1. Introduction

Knowledge discovery in databases is the process of discovering useful knowledge from a collection of data. It is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. Knowledge discovery is the process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed result. There are different approaches to discovery, which includes inductive learning [17].

Machine learning is a scientific discipline that deals with the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions [7]. Designing a Learning System includes Problem Description, Choosing the Training Experience, Choosing the Target Function, Choosing a Representation for the NetCDF is network common data form which is developed by Unidata program center normally used in atmospheric research. NetCDF is a platform independent format for representing multi-dimensional array-orientated scientific data. NetCDF is new to the GIS community but widely used by scientific communities for around many years [12]. The purpose of the Network Common Data Form (NetCDF) interface is to support the creation, efficient access, and sharing of data in a form that is self-describing, portable, compact, extendible, and archivable [4, 7].

This paper is organized as follows: section II describes literature survey about NetCDF data model characteristics and components. It also focuses on some machine learning techniques. Section III presents overall idea about proposed system. Conclusion is presented in section IV.

2. Literature Survey

A) Netcdf Data Model

NetCDF dataset contains scientific data. In NetCDF data model a scientific data is conceptually modeled with a set of objects, operations, and rules that determine how the data is represented and accessed [3].

i) Characteristics

a) Self-Describing - A NetCDF file includes information about the data it contains. The header (metadata) describes the body (data)
b) Portable - A NetCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
c) Scalable - A small subset of a large dataset may be accessed efficiently. Both command-line programs (CDO) and data servers (Open DAP) allow quick and easy sub setting (and super setting).
d) Appendable - Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
e) Archivable - Access to all earlier forms of netCDF data will be supported by current and future versions of the software.
f) Initially Annoying - The spin-up is not trivial, but the payoffs are sweet.
g) Direct Access - A small subset of a large dataset may be accessed efficiently, without first reading through all the preceding data.
h) Sharable - One writer and multiple readers may simultaneously access the same NetCDF file [5].

ii) Components

A typical NetCDF file has three sections
a) Variables: variables are the basic unit of data in a NetCDF dataset. These are array of data. There can be multiple variable with different data types.
b) Dimensions: When a variable is defined, its shape is specified as a list of dimensions Dimension is integer parameter which defines the structure or shape of the data array stored in NetCDF files (e.g. Time, Depth, Latitude, and Longitude). Dimension may have attached attributes. Multidimensional data is represented in fig 1.
c) Attributes: Attributes are 1-dimensional array of value.

Users are responsible to decide what attributes should include in netCDF files. There are two types of attributes:

- Global Attributes: Describe the contents of the file
- Variable Attributes: Attributes defines descriptive data associated with variable [1, 11]

iii) Common Data Language

CDL (Common Data Language) is text notation for NetCDF objects and data. CDL is network Common Data form Language described as follow:

NetCDF name {Dimensions: ...//Contains metadata
Variables: Global attributes
Attributes: Local attributes}

CDL ensures interoperability with the help of utilities like ncdump, ncgen –b, ncgen –c and so on as shown in fig.

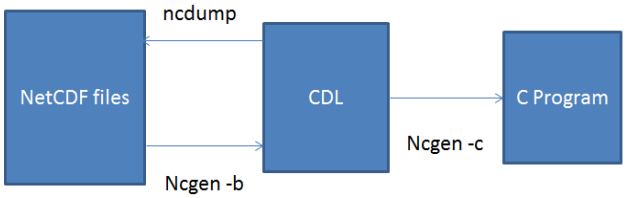


Fig 1: CDL notation and NetCDF utilities

iv) 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 [2].

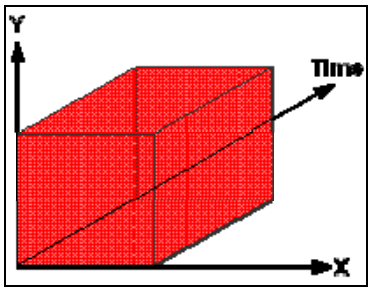


Fig 2: Multidimensional data representation

B) Advantage over Existing System

1. One traverse
2. Fast data access
3. Easy to apply logic
4. Cost : freely available
5. Hardware requirement is very low
6. Easy to take a backup of file

C) Machine Learning Techniques

Learning is used when Solution needs to be adapted to particular cases or Solution changes in time.

Machine learning generally falls into categories

i) Supervised Machine Learning Technique

SL is a machine learning mechanism which is more supervised learning technique that first finds a mapping between inputs and outputs based on a training dataset, and then makes predictions to the inputs that it has never seen in training [4]. Supervised Learning is based on statistics like Classification and Regression [18, 7, 12].

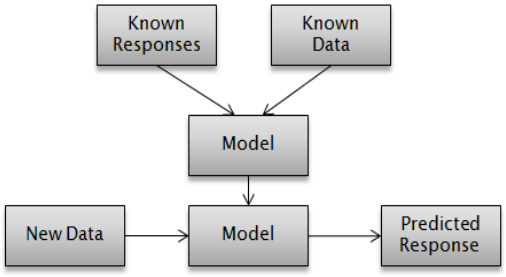


Fig 3: Supervised Machine learning

ii) Unsupervised Machine Learning Techniques

UL is based on what normally happens. UL inspired by the brain's ability to extract patterns. Clustering is the most important form of UL. It deals with data that have not been pre classified in any way, and does not need any type of supervision during its learning process. The most well-known example of clustering algorithm is k-means clustering [4]. In UL model not provided with correct results during the training. UL uses statistical properties to cluster the data. Normally UL is used when datasets involved hundreds of thousands of variables. This is the new technique of machine learning because most big datasets do not come with labels [7].

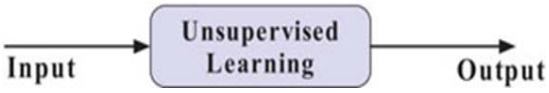


Fig 4: Unsupervised machine learning

iii) Reinforcement Machine Learning Technique

The reinforcement learning (RL) approach enables an agent to learn a mapping from states to actions by trial and error so that the expected cumulative reward in the future is maximized. RL is powerful since a learning agent is not told which action it should take; instead it has to discover through interactions with the system and its environment which action yields the highest reward [4, 7].

Fig 5: Reinforcement machine learning

3. Proposed System

Propose system is going to performed knowledge discovery for NetCDF datasets using machine learning techniques. Hence an effective data extraction is performed from these NetCDF datasets. System architecture shown in fig: 7. First scientific data collected from various sources, and then this data is provided to NetCDF environment where it gets converted into NetCDF format [12]. Dimension and variable components of NetCDF files stored metadata while Attributes stored actual data. These NetCDF data is further provided to machine learning technique. There are different machine learning techniques available but propose system select best solution technique by comparing all and apply on NetCDF data.

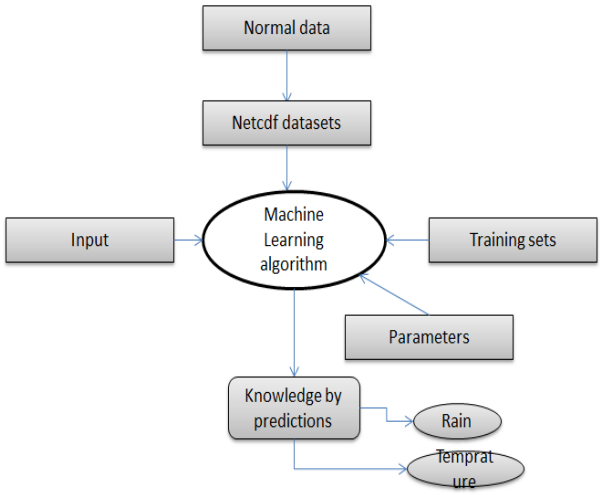
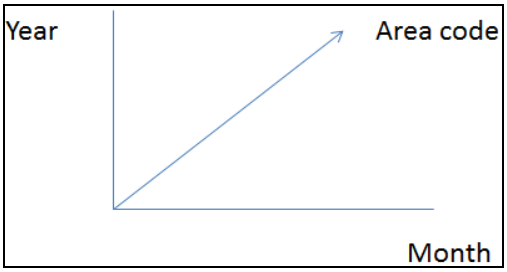


Fig 6: System Architecture

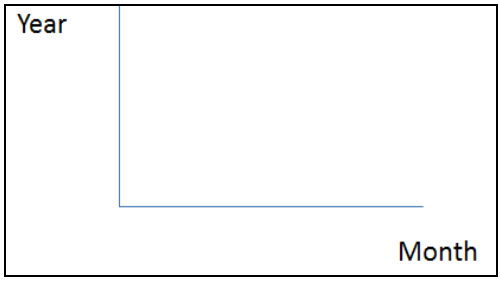
Main function of machine learning technique is to predict output of certain input data by considering past result [7]. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. Machine learning technique create training sets using NetCDF data, and then by extracting feature and characteristics of data statistical model is created. This statistical model generate hypothesis for input data. This hypothesis helps to discover knowledge. At the same time remaining data is considered as testing data, Machine learning techniques then compute model by extracting feature from these testing datasets. Both models get validated that helps for accurate prediction.
Finally we can discover knowledge by prediction with the help of past results and statistic graph generated by system with the help of R language.

Multidimensional parameters for the system to predict rainfall in India

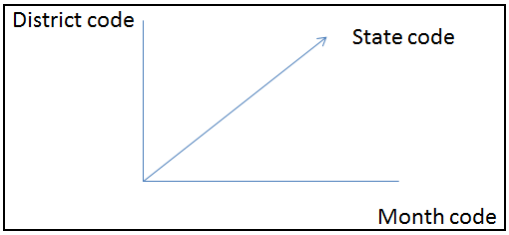
1. Average minimum temperature



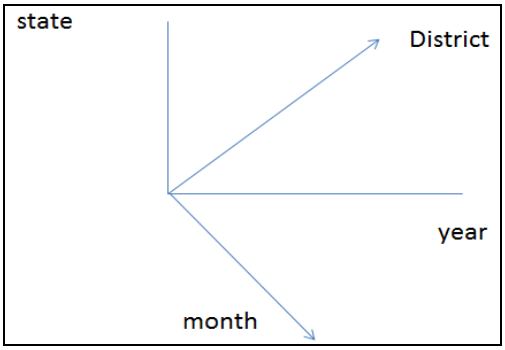
2. India's maximum temperature



3. District rainfall



4. India monthly rainfall data



4. Conclusion

NetCDF is very popular and now globally accepted data representation format. NetCDF is a widely used file format in atmospheric and oceanic research [15]. Machine Learning technique is very well suited and effective for extracting knowledge in any application. Proposed system considered NetCDF and adds all the benefits of common data form in the system instead of other normal data forms. NetCDF programs are written with the help of CDL which ensure platform independency. Hence complex data such as heterogeneous data can be access with very high efficient way. More accurate prediction is done with the help of machine learning technique. The propose system ensures very effective knowledge discovery as it uses machine learning mechanism which includes statistical learning and prediction ability. System access large historical data i.e. over 100 years.

Acknowledgment

It is great pleasure for me to acknowledge the assistance and contribution of number of individuals who helped me in presenting "Knowledge Discovery and weather prediction of India using Machine Learning" paper.
First and foremost I wish to record my gratitude and thanks to Ms. S. P. Potdar for her enthusiastic guidance and help in successful completion of the paper. I would also like to thank Mr. R. P. Modi for his continuous support. I express my thanks to Mr. V. V. Puri, for his valuable guidance.
I would also extend my gratitude to Mr. P.R. Sonawane, Sonix Nano system, Pune, for being a constant source of inspiration.

5. References

1. Harry L. Jenter and Richard P. singnell: NetCDF- A public domain software solution to access a data problem for numerical modeler.
2. Rew R, Davis G, NetCDF: an interface for scientific data access, IEEE Computer Graphics and Applications. 1990; 10(4):76-82

3. Edward Hartnett*, and R. K. Rew UCAR, Boulder, CO: Experience with an enhanced NetCDF data model and interface for scientific data access.
4. Shouyi Wang, Student Member, IEEE, Wanpracha Chaovalitwongse, Member, IEEE, and Robert Babuˇska: Machine Learning Algorithms in Bipedal Robot Control.
5. Rew G. Davis S. Emmerson, H. Davies, and E. Hartnett. The NetCDF Users Guide, Version 3.6.1. Unidata Program Center. 2006.
6. http://www.unidata.ucar.edu/software/netcdf/
7. Tom M. Mitchell, Machine Learning, Published October 1st by McGraw-Hill. 1997.
8. Peter Barnum and Vinithra Varadharajan: When to Picnic? - The Robotics Institute, Carnegie Mellon University Pittsburgh, PA 15213
9. Yu Su Computer Science and Engineering, The Ohio State University, Gagan Agrawal, Jonathan Woodring CCS-7.: Indexing and Parallel Query Processing Support for Visualizing Climate Datasets - Applied Computer Science Group-Los Alamos National Laboratory.
10. Castroa R, Vegaa J, Ruizb M, De Arcasb G, Barrerab E, Lópezb JM *et al.* NetCDF based data archiving system applied to ITER Fast Plant System Control Prototype http://www.hdfgroup.org/HDF5/.
11. 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.
12. Pavel Michna and Milton Woods: RNetCDF – A Package for Reading and Writing NetCDF Datasets.
13. Rew, G. Davis, S. Emmerson, H. Davies, E. Hartnett, and D. Heimbigner. The NetCDF Users Guide, Version 4.1.3. Unidata Program Center, 2011.
14. Eaton, J. Gregory, B. Drach, K. Taylor, and S. Hankin. NetCDF Climate and Forecast (CF) Metadata Conventions, Version 1.6, 2011.
15. Mark Hall Pentaho Corporation Suite 340, 5950 Hazeltine National Dr. Orlando, FL 32822, USA, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. Witten Department of Computer Science Hamilton, University of Waikato New Zealand- The WEKA Data Mining Software: An Update.
16. William J Frawley, Gregory Piatetsky-Shapiro, Christopher J. Matheus-Knowledge Discovery in Databases: An Overview
17. Nikhil Sethi, Dr. Kanwal Garg. Exploiting Data Mining Technique for Rainfall Prediction, (IJCSIT) International Journal of Computer Science and Information Technologies. 2014; 5(3):3982-3984.