Data mining in geosciences

Conference Paper · November 2003

DOI: 10.1109/TELSKS.2003.1246283 · Source: IEEE Xplore

7 authors, including:

David Pokrajac
Delaware State University
117 PUBLICATIONS **648** CITATIONS

SEE PROFILE

Kagya A Amoako
University of Michigan
11 PUBLICATIONS **84** CITATIONS

SEE PROFILE

Data Mining in Geosciences

D. Pokrajac[1], K. Amoako[1], H. Patel[1], J. Brooks[1], N. Cenat[1], K. Marcus[1], S. Darden[1]

Abstract—**In this paper we discuss various aspects of data mining in geosciences. We provide overview of data formats, public datasets and data mining algorithms that can be used for knowledge discovery in geoscience data.**

Keywords—**Data Mining, Geosciences, Satellite Data, Metereological Data, Distribution Learning.**

I. Introduction

Technology has greatly increased our ability to assemble spatial-temporal databases. Many commercial and government remote sensing satellites imaging the Earth and global positioning systems have made it easy to georeference readings from a large range of on-the-ground sensors. Unfortunately, this abundance of data collecting techniques has not been accompanied with the availability of methods to analyze large spatial-temporal databases with many attributes, which currently limits the use of these data. Existing tools that could help in analyzing spatial data include geographic information systems and geographic imaging packages. These tools facilitate visualization and preprocessing of spatial data and provide query and feature manipulation functionality. However, they do not include nor do there exist many algorithmic techniques to find patterns and associations within spatial-temporal data. There is a growing research activity in the area of data mining, but little of this addresses the unique characteristics of spatial-temporal data.

Commercial applications of data mining in areas such as e-commerce and fraud detection have taken central focus in research and applications for a long time period. However, there is a need for a significant amount of innovative data mining work to take place in the context of scientific applications [1]. The advent of fast computational facilities makes possible systematic reanalysis and simulation based on non-uniformly collected heterogeneous data. The NCEP/NCAR Reanalysis Project, as a joint project of National Center for Environmental Prediction (NCEP) and the National Center for Atmospheric Research (NCAR), has a goal to produce new atmospheric analyses using historical data as well to produce analyses of the current atmospheric state [2]. This effort results with 55GB/year of processed data, containing several temporal climate and weather attributes at a regular 3D spatial grid for 50+ years of atmospheric fields. The data have been employed in various domains including climatology, forestry and environmental sciences as well to create "snapshot" annual CD-ROMs containing digests of original reanalysis data.

Technological advances, along with the increased impact of non-military applications, has led to rise of satellite remote sensing for civil purposes, including geosciences and environmental sciences [3]. In addition to a network of geostationary (e.g. GOES-I series, METEOSAT) and polar orbiting (e.g NOAA-14–NOAA-16) weather and meteorological satellites, novel series of satellites have been planned that provide steady data streams from multiple sensors to accomplish deeper understanding of climate and environmental changes. NASA Earth Observation System—EOS [4] consists of several low-altitude satellites and is the first system to offer integrated measurements of the Earth's processes. EOS supports a coordinated series of polar-orbiting and low-inclination satellites for long-term global observations of the land surface, biosphere, solid Earth, atmosphere, and oceans. These systems are characterized with a huge amount of daily produced data. Thus, Landsat 7 instruments have a data rate of 150Mbps [5], while Terra produces data in order of 1TB daily [4]. These massive amounts of information are to be handled using the specially developed storage and processing distributed systems such as Earth Observing System Data and Information System (EOSDIS). It is potentially very useful to develop and evaluate techniques for determining relationship among data of different origin and physical nature, in order to improve an understanding of underlying processes and enhance knowledge discovery for various related fields [6].

In spite of significant work in related domains (simulation techniques, remote sensing, databases) and considerable attempts at spatial and spatial-temporal data mining [7], there is an urgent need for additional activities for both theoreticians and practitioners in order to make a genuine application of the emerging technologies a reality. In Section II of this paper, we discuss data formats and publicly available data sets, while Section III presents several potentially promising techniques applicable in data mining of geoscience data. We believe that the proposed techniques could significantly improve our understanding of geoscience processes which on the other hand could lead to further improvements in spatial-temporal data mining methodology.

II. Geoscience Data

Data preprocessing is important but frequently neglected step of data mining. Knowledge of formats data are stored is crucial for proper choice of data mining tools. Here, we discuss the most common geoscience data formats—HDF, GRIB and GeoTIFF.

[1] All authors are with Delaware State University, Computer and Information Science Department, 1200 N Dupont Hwy, Dover, DE 19901, USA, E-mails: dragoljub.pokrajac@verizon.net, kagya_a@hotmail.com, josettebrooks@hotmail.com, n_cenat@hotmail.com, sedrdn00@hotmail.com, kelvin_marcus@hotmail.com, patelhe@dsc.edu

0-7803-7963-2/03/$17.00 ©2003 IEEE

HDF. Hierarchical data format (HDF) is a "library and multi-object file format for the transfer of graphical and numerical data between machines" [8,9]. It was created by the National Center For Supercomputing Applications (NCSA) and extensively used by NASA-EOS. The reasons for HDF popularity include system independence and portability. Due to its hierarchical structure, HDF is particularly suitable for on-line analytical processing and data mining of scientific data. The HDF format is supported on various platforms (Linux, Windows, Mac,etc) and through different vendors of public and commercial software [9] which support data analysis, visualization and manipulation. The HDF format puts main emphasize on storage and I/O efficiency. HDF files contain the collections of scientific data objects organized into datasets and groups. The datasets are here defined as multidimensional arrays and support the metadata while the groups are directory-like structures containing datasets or other groups. The components of datasets include *arrays* (identically typed data specified by indices), dataspaces (containing information about size, shape and selected regions of array), users-defined attribute list and storage options.

GRIB. GRIB (GRidded In Binary) format is a World Meteorological Organization (WMO) standard for the efficient exchange of gridded binary data [10]. GRIB was initially designed to accomplish fast and efficient broadcast and dissemination of weather data generated by numerical forecast models. However, the format is also frequently used for meteorological satellite images. In addition to WMO, GRIB is also used by National Weather Service, Air Force Global Weather Central and Fleet Numerical Meteorology and Oceanography Center for the exchange of gridded data files. A GRIB file typically consists of a short Indicator section serving as a file header, followed by Product Definition Section that identifies the file content (e.g. time, source, units), and a Grid Description Section identifying data geometry (e.g. projection type, grid size). Since GRIB files may contain missing or corrupted values, the GRIB format may include an optional bitmap section that serves to mask missing/non-relevant data. Finally, a GRIB file contains the data in packed format followed by a key sequence used to denote the end of a GRIB product. Software to access, visualize and convert GRIB files is publicly available though the National Center for Atmospheric Research (NCAR), European Centre for Medium-range Weather Forecasts, and UCAR Unidata Program Center.

GeoTIFF. GeoTIFF [11] is built on Tag Image File Format (TIFF), a widely used image data format. GeoTIFF fully complies with the TIFF 6.0 specifications, and uses a small set of TIFF tags to store georeferencing information, serving geographic as well as projected coordinate systems needs (UTM, US State Plane and National Grids) on top of raster images. It supports geographic projection types (e.g. Transverse Mercator, Lambert Conformal Conic, etc). GeoTIFF takes advantage of TIFF platform-independent data format to enhance data interchange, since GeoTIFF keys are designed analog to standard TIFF tags. The format is flexible such that new keys may be defined as necessary. GeoTIFF uses pre-defined codes to specify metadata (projection types, coordinate systems, etc). The GeoTIFF information content is designed to be compatible with the data decomposition approach employed by the National Spatial Data Infrastructure (NSDI) of the U.S. Federal Geographic Data Committee. GeoTIFF format is supported by leading GIS, mapping and digital photometry software vendors and packages (e.g., ESRI ArcInfo and ArcView, MapInfo, SocetSet, etc). However, since GeoTIFF is actually a superset of TIFF, ordinary image processing software will simply ignore the GeoTIFF tags and display the images as regular TIFF. Therefore, when geographic referencing is not a priority, GeoTIFF data can be analyzed through general-purpose data analysis environments (e.g. Matlab).

Numerous institutions collect spatial-temporal geoscience data. In the following paragraphs we discuss two publicly available data sets.

Datasets from EOS satellites are collected by NASA/JPL through Enhanced Thematic Mapper Plus instrument [4,5]. Landsat7 remote sensing data can be obtained through NASA Distributed Active Archive Center (http://edcdaac.usgs.gov/main.html). Radiometrically and systematically corrected Level1R data products are in HDF and GeoTIFF formats [5]. Typical data size for one $30*30m^2$ resolution scene is 35MB/band [3].

Climate reanalysis dataset [2] is available from NCEP/NCAR. Subsets of data contain attributes at regular spatial grid of $2.5^0 \times 2.5^0$ at 17 altitude levels (each matching a constant atmospheric pressure) and corresponding to a regular 6-hour observational interval. The considered dataset includes u- and v- components of wind and temperature and corresponds to each year of atmospheric observations. The climate reanalysis dataset is available in GRIB format. Initial data mining experiments can be performed on data from Annual Reanalysis CDROMs with further migration to *ds090.0* annual dataset with the larger set of variables corresponding to 17 pressure levels and at regular 6-hour observational interval.

III. ALGORITHMS

To facilitate efficient scientific discoveries in Earth observation databases we propose building and maintaining a library of spatial and temporal patterns identified on historic data (by domain expects or automatically) each represented as a distribution model. For a region of interest from subsequently collected data, the objective is to determine whether it is similar to one of the library patterns in a distributional sense. We propose learning the distribution model of the identified region followed by distributions comparison to the library models, where based on the outcome we can: (i) label the region of interest as corresponding to one of the well characterized library patterns; (ii) use data from the region of interest to better learn the distribution of one of the library patterns; and (iii)

augment the library by introducing a new pattern corresponding to the identified region of interest.

Main issues that should be considered within this approach include (1) Tailoring existing spatial-temporal algorithms to better serve specific needs of spatial-temporal data mining at Earth observation databases; (2) Understanding influence of a data type (binary or continuous values) and underlying data distribution characteristics on performance of specific learning algorithms; and (3) Addressing performance degradations due to temporal and spatial non-stationarity of data.

We propose a "greedy" generic pattern matching procedure consisting of the following steps. For each new temporal layer of spatial data, estimate underlying distribution based on data from a new temporal layer. Compute distributional distance between a new distribution and the distributions corresponding to memorized library patterns and identify the distribution for which the distance is minimal. If the distance is sufficiently small, assign new data to the closest library pattern. Otherwise, include the new distribution to the library as a new pattern. Periodically update estimates of the library distributions by emphasizing recent temporal layers assigned to the corresponding pattern.

Depending on prior knowledge, size of the identified regions of interest and allowed computational time, for learning distributions we assess comparative efficiency of parametric, semi-parametric and non-parametric methods. These will include non-parametric approaches of histogram and kernel–based algorithms [12], parametric approaches of k-means [13] and semi-parametric neural-network based techniques, as explained in the rest of this Section.

Non-Parametric Methods. Non-parametric methods for probability distribution modeling do not make any prior assumption about the form of a distribution function. We consider classic histogram estimation obtained by approximating density as a fraction of data points that fall into each of a pre-specified number of bins positioned equidistantly along each of dimensions of the dataset [12].

k-Means algorithm. Although intuitively appealing, non-parametric techniques suffer from poor scaling with data dimensionality. In addition, due to a large degree of freedom, the non-parametric model estimation is less robust and less accurate as compared to the parametric techniques. Parametric techniques can provide better generalization with the overhead of introducing additional prior information about distribution form. Therefore, we consider learning distributions by the k-means algorithm [13] which is a variant of the expectation-maximization (EM) method [14] aimed to determine the means of Gaussian mixture components [15].

Self-Organized Maps. To avoid strict assumptions of Gaussian data properties, self-organized maps (SOM) can be applied for semi-parametric estimation of distributions corresponding to library patterns and processed temporal data layers. SOM transform a pattern of an arbitrary dimension into a lower dimension discrete map through unsupervised neural-network learning [16]. As in the k-means algorithm, SOM uses Euclidean distance and winner-take-all approach to activate the neuron "closest" to the pattern. However, SOM updates not only the weights of the "winner" but also its neighboring neurons with intensity of the movement inversely proportional to the distance from the activated neuron. This enables construction of a topologically ordered mapping from the input space to the neuron lattice with the neighborhood structure defined in the neurons space. In practice, the learning rate and neighborhood function may vary with the number of examples (thus emphasizing more recent examples), whereas typically the neighborhood function exponentially decreases with the distance between two neurons. In addition to rectangular lattice, additional topologies including hexagonal are proposed.

SOM and k-means algorithm have been extensively compared in literature. Thus, Balakrishnan et al, [17] pointed out that k-means algorithm can outperform SOM when the underlying distributions were Gaussian mixtures. Schreer et al. [18] shown that k-means and SOM have similar performance on real-world data. On the other hand, as demonstrated by Yin and Alinson [19], SOM is less sensitive to falling into deep local minima, since the updating in the SOM is a stochastic gradient method instead of a strict gradient like in the k-means. However, it is en empirical question which of the two methods will perform better on Earth observation data since the community is still lacking SOM characterizations on standardized benchmark problems.

Cellular Neural Networks. Cellular networks (CNN) [20] have originally been proposed as highly parallel, non-linear multi-output signal processing systems. Similar as in back propagation networks, an output of a network node is a non-linear function (soft limiter) of the network node state. However, unlike in back propagation networks, here all the nodes are laterally connected and their states are controlled by feedback mechanism. Also, in contrast to backpropagation networks, here each node has memory so the node current state depends on the node states in past. Similar as STUG models [21], CNN exhibit high symmetry in node structure and interconnections, but unlike the former model, non-linearity and feedback give CNN potentially higher processing powers. Finally, the lateral connections of CNN neurons are similar to the connections in SOM.

CNN have been employed for various image processing tasks, including noise removal, non-linear filtering and feature enhancing and extraction. In geoscience data mining, CNN may be applied to learn characteristic properties of a particular scene and then to detect the changes in scene through identifying significant transients. The inherent parallel structure of CNN could result in fast real-time execution of algorithms on parallel computers or specialized hardware. However, to truly apply this concept in data mining, numerous problems have to be solved. One of most important unresolved issue is how to incorporate semantic-preserving non-linear time-space transformations into the CNN paradigm.

ART Networks. Adaptive Resonance Theory (ART), is introduced and developed by Carpenter and Grossberg [22] as

a learning technique based on the winner-take-all concept. The approach also solves the "stability-plasticity" dilemma, which addresses the problem of a learning system preserving its previously learned knowledge (plasticity) while at the same time being able to learn and acquire new knowledge (stability). A useful property of ART networks is the fast and stable learning. These networks come in both supervised and unsupervised modes. The two most common unsupervised models of the ART networks are the ART1 and ART2 models [23]. The supervised models include ARTMAP and Fuzzy ARTMAP [24]. The ART1 model is capable of learning binary input patterns (discrete patterns) while the ART2 model can also learn analog patterns (continuous-valued). The basic architecture for ART networks consists of three groups of neurons—an input processing layer, a cluster layer and the reset and control mechanism. The feedback mechanism that exists between the input and the cluster layers is responsible for the solving the "stability-plasticity" dilemma. Both the ART1 and ART2 models consist of two subsystems, the Attentional and the Orienting subsystem. The Attentional subsystem consists of the comparison and the recognition layers while the Orienting subsystem consists of a reset layer.

ART networks have been applied in pattern classification and pattern recognition processes. They have also been used in clustering remote sensing and Meteorological data [25].

IV. Conclusion

In this paper we discuss various aspects of data mining in geosciences. After providing brief overview of data formats and available satellite remote sensing and metereological datasets, we discuss several machine learning and statistical algorithms that can be used for knowledge discovery of scientific and geoscience data. Work in progress includes extensive experimental evaluation of discussed and novel data mining techniques on remote sensing satellite data.

Acknowledgements

This study was supported in part by DSU PDF award, NIH-funded Delaware BRIN program, and by DoD HBCU/MI Infrastructure Support Program to D. Pokrajac. K. Amoako, N. Cenat, J. Brooks, H. Patel and K. Marcus were partially supported through the Greater Philadelphia AMP program.

References

[1] Kåmath, C., "Introduction to scientific data mining," presented at *Mathematical Challenges in Scientific Data Mining*, 2002, available at http://www.ipam.ucla.edu/publications/sdm2002/sdm2002_ckamath.pdf.

[2] Kalnay, E., et al., "The NCEP/NCAR 40-Year reanalysis project," *Bull. Amer. Meteor. Soc.*, Vol. 76, 437-471, 1996.

[3] Barrett, E.C.,Curtis, F.L., *Introduction to Environmental Remote Sensing*, 4th edn. Stanley Thornes Pub. Ltd, Cheltenham,UK, 1999.

[4] NASA, 1999 *EOS Reference Handbook, A Guide to NASA's Earth Science Enterprise and the Earth Observing System*, 1999, available at http://eospso.gsfc.nasa.gov/ftp_docs/hand.book99.pdf.

[5] NASA—Goddard Space Flight Center, Landsat Project Science Office, *ScienceData Users Handbook*, available at http://ltpwww.gsfc.nasa.gov/IAS/handbook/handbook_menu.html, 2002b.

[6] Loeb, N.G., Vámai, T., and Davies, R., "Effect of cloud inhomogeneities on the Solar zenith angle dependence of nadir reflectance," *J. Geophys. Res.*, Vol. 102, 9387–9395, 1997.

[7] Roddick, J.F., Hornsby, K., and Spiliopoulou, M., "An updated bibliography of temporal, spatial and spatio-temporal data mining research," *Post-Workshop Proc. Int'l Workshop on Temporal, Spatial and Spatio-temporal Data Mining, TSDM2000*, Lecture Notes in Artificial Intelligence 2007, Springer-Verlag, pp. 147-164, 2001.

[8] NCSA, *HDF5 a New Generation of HDF*, accessed 6/27/03, http://hdf.ncsa.uiuc.edu/HDF5/.

[9] NASA, *HDF-EOS Tools and Information Center*, accessed 6/21/2003, http://hdfeos.gsfc.nasa.gov/hdfeos/softwarelist.cfm.

[10] NCEP, *Office Note 388 GRIB*, accessed 6/21/2003, http://www.nco.ncep.noaa.gov/pmb/docs/on388/.

[11] USGS, *GeoLevel 1 Product Output Files Data Format Control Book*, Vol. 5, Book 2, Revision 5, April 2001, available at http://landsat7.usgs.gov/documents/L7-DFCB-04-L1p-FCB.pdf

[12] Bishop, C., *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[13] Lloyd, S., "Least-squares quantization in PCM," *IEEE Trans. Inf. Theory*, Vol. IT-2, 129-137, 1982.

[14] McLachlan, G., and Krishnan, T., *The EM Algorithm and Extensions*, John Wiley & Sons, Inc., 1996.

[15] Duda,R., Hart, P., and Stork, D., *Pattern Classification*. John Wiley & Sons, New York, 2000.

[16] Kohonen, T., *Self-organized Maps*. Springer-Verlag, 2001.

[17] Balakrishnan, P.V., Cooper, M.C., Jacob, V.S., and Lewis, P.A., "A study of the classification capabilities of neural networks using unsupervised learning: a comparison with k-means clustering," *Psychometrika*, Vol. 59, 509-525, 1994.

[18] Schreer, J.F., O'Hara Hines, R. J., and Kovacs, K. M. "Classification of dive profiles: A comparison of statistical clustering techniques and unsupervised artificial neural networks," *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 3, 383-404, 1998.

[19] Yin, H., Allinson, N. M. "Comparison of a Bayesian SOM with the EM algorithm for Gaussian mixtures," *Proc. WSOM'97: Workshop on Self-Organising Maps*, pp. 118-123, 1997.

[20] L. O. Chua, *CNN: A Vision of Complexity*, World Scientific, Singapore: River Edge, NJ, 1998.

[21] Pokrajac, D., Hoskinson, R., and Obradovic, Z., "Modeling spatial-temporal data with a short observation history", *Knowledge and Information Systems*, in press.

[22] Carpenter, G.A., Grossberg, S., "Art 2: selforganisation of stable category recognition codes for analog input patterns", *Applied Optics*, Vol. 26, pp. 4919-4930, 1987.

[23] Aleshunas, J. J., St. Clair, D.C., and Bond, W.E., "Classification Characteristics of SOM and ART2," *Applied Computing 1994: Proc. 1994 ACM Symposium on Applied Computing*, pp. 297-302, 1994.

[24] Bálya, D., Roska, T., "Supervised and unsupervised ART-like classifications of binary vectors on the CNN universal machine, *Proc. CNNA 20002*, available at lab.analogic.sztaki.hu/cnna_papers/ DBalya_ARTCNN.pdf.

[25] Vlajic, N., Cord, H.C. "Vector quantization of images using modified adaptive resonance algorithm for hierarchical clustering," *IEEE-Trans. Neural Networks*, Vol. 12, 1147-1162, 2001.