An Interactive Data Mining Framework for EarthCube

Joseph B.H. Baker and J. Michael Ruohoniemi
Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg Virginia

Naren Ramakrishnan
Department of Computer Science, Virginia Tech, Blacksburg, Virginia

Vision: An EarthCube Data Mining Framework for Geosciences Research and Life-Long Learning

The goal of EarthCube is to transform the conduct of geosciences research by supporting the development of community-guided cyberinfrastructure to integrate data and information for knowledge management. Compiling the multitude of geosciences datasets into one single “EarthCube” will undoubtedly increase scientific productivity by making it easier for researchers to access multiple datasets quickly. However, a major challenge in geosciences is the reliable extraction of recurrent features from the massive archive of multi-dimensional datasets, many of which exhibit a large degree of spatiotemporal sparseness. This challenge exists today, and will only become much more pronounced once EarthCube is fully operational. It is our thesis that the majority of geosciences datasets are currently under-utilized - not simply because of issues with data access and distribution - but rather because the geosciences community does not yet have sufficient computational capabilities in automated event detection and feature classification to extract the fullest quantitative information from the datasets. What is really required to allow EarthCube to become more than just a comprehensive data clearing house is the development of sophisticated new tools for interactive visualization and mining of multiple datasets for physical content. Such an effort will require an intellectual partnership between the geosciences and computer science research communities. Several existing databases can function as test-beds for development of new EarthCube data mining algorithms, allowing the effort to proceed in parallel with the roll-out of EarthCube cyberinfrastructure. Once fully incorporated into EarthCube the new algorithms will automate extraction of important recurrent features across multiple datasets and thereby improve scientific productivity. Furthermore, the new data mining framework will also function as a flexible machine learning environment for students of all ages, allowing them to become “citizen scientists” in much the same way that the internet has enabled the rise of a generation of “citizen journalists” (i.e. bloggers). In summary, if EarthCube is to be fully successful in producing transformative change in the conduct of geosciences research and education, then the design needs to explicitly include plans to develop a dedicated interactive framework for cutting-edge data mining.

Geosciences: An Emerging Front in Data Analytics

Over the years, the NSF Geosciences Directorate has made substantial investments in research infrastructure to collect a wide variety of geosciences sensor data, as well as funding for the research community to analyze the datasets and advance understanding. However, the sheer number and volume of datasets has quickly grown to become beyond the reach of traditional data analysis tools. In short, the field is now awash in data, and an injection of new ideas from the computational science and data analytics communities is needed to make further scientific progress. A related problem is data distribution and access. The traditional approach for providing access to geosciences datasets has been to post preview products of sensor data to a website where it can be browsed by interested users and perhaps downloaded for later analysis. However, the actual interest of the scientific user usually does not lie in the measurements themselves, but rather in their deeper physical content, and how that content may relate to physical content residing in other datasets. Gaining access to that deeper content is often obstructed by the sheer complexity and volume of data presented to the user. A skilled user, who has gained experience working with a particular dataset, develops the necessary skills required to identify features of interest and the tools to extract the physical content for detailed study. However, the novice user can only gain access to that expertise through time-consuming personal collaboration with a skilled user. Even the skilled user often lacks the required knowledge in modern computer-aided data analysis techniques to extract the fullest information from the dataset. Further progress requires a mechanism for lowering the threshold for gaining access to the deepest physical content residing in all geosciences datasets. The EarthCube initiative represents an opportunity to make bold progress in this critical area of data analytics by weaving a robust data mining framework into the supporting EarthCube cyberinfrastructure to increase research productivity.

Definition of User Requirements: Extracting Spatiotemporal Features from Geosciences Datasets

The best way to describe the user requirements for an EarthCube data mining framework is to provide a simple representative example of the challenges associated with extracting complex features from geosciences datasets. In this section we describe these issues in the context of the Super Dual Auroral Radar Network (SuperDARN).

Case Study: An Under-Utilized Resource for Geophysical Waves (SuperDARN)

SuperDARN is an international radar network for studying the Earth's upper atmosphere, ionosphere, and connection into space [e.g. *Greenwald et al.,* 1985*; Ruohoniemi et al*, 1987; *Chisham et al.*, 2007; *Baker et al.*, 2007]. SuperDARN radars operate at HF frequencies (9-18 MHz) to make use of refraction in the ionosphere to "bend" the radar signals and extend the range "over-the-horizon" to several thousand kilometers. All radars operate continuously and in common mode complete azimuth scans every 1-2 minutes. Some radars have been operating more or less continuously for more than 25 years. Backscatter is obtained from a variety of targets including plasma irregularities in the ionospheric *E* and *F* regions, meteor trails at mesospheric heights (80-90 km), and the Earth's surface after reflection from the ionosphere. At the present time there are 18 SuperDARN radars operational in the northern hemisphere and 9 radars in the southern hemisphere. The NSF currently provides funding for the US component of SuperDARN radar operations through the Geospace Facilities (GF), Office of Polar Programs (OPP), and Mid-Size Infrastructure (MSI) programs. Within the geosciences research community SuperDARN is best known for producing widespread measurements of ionospheric plasma drift velocities which are routinely combined to produce patterns of ionospheric plasma convection or "space weather" maps. However, there is a richness of other more exotic geophysical phenomena manifested in the SuperDARN dataset which are currently under-utilized because they are much more difficult to extract from the data stream. For example, SuperDARN radars can monitor the dynamics of atmospheric gravity waves which appear as quasi-periodic enhancements of the ground-backscattered power with periods of 20 to 50 minutes [*Bristow et al.*, 1996]. Gravity waves are generally assumed to be produced by processes occurring in the lower atmosphere and propagate upwards to the thermosphere-ionosphere altitudes. Another under-utilized SuperDARN data resource is the near-range meteor trail measurements obtained at altitudes of 80-90 km which can be used to infer the magnitude and direction of neutral winds in the mesosphere [*Hall et al*., 1997]. By combining measurements from multiple radars across an extended longitude sector it is also possible to examine large-scale processes occurring in the upper atmosphere, such as planetary waves and tides [*Malinga and Ruohoniemi*, 2008]. Although it is recognized that these wave features routinely occur in SuperDARN data, there are several reasons why they haven't been fully utilized thus far. First, the features are manifestations of atmospheric phenomena, rather than ionospheric phenomena, and so they fall outside the immediate scientific domain of most SuperDARN scientists. Second, the atmospheric features themselves are of very large spatial scale and so the individual radar signatures represent only a small spatiotemporal glimpse of the larger phenomenon; stitching together a coherent picture of the large-scale dynamics using the multi-radar dataset tends to be a real challenge. Third, the features tend to be "second order" in amplitude and are overlapped with other more dominant "first order" ionospheric features, which tend to obscure them. The features therefore do not easily lend themselves to (elementary) computer algorithms for event detection (e.g. thresholding) or feature extraction. As a result, there is a tendency for primary SuperDARN research activities to be focused on examination of the larger "events", such as geomagnetic storms and substorms. For these reasons there is a whole zoo of recurrent features in SuperDARN data that is not being fully analyzed, simply because it is too time-consuming to try and extract quantitative information about them from the data stream. Undoubtedly, the same is true in other geosciences datasets too. If we are ever going to unravel the true nature of interconnected complexity within the entire geosciences system, then we need to develop sophisticated new tools for data integration, visualization and data mining that can automatically identify and extract complex features from multiple multi-dimensional datasets that are irregularly gridded in both space and time.

EarthCube Data Mining: An Intellectual Partnership between Computer Scientists and Geoscientists

In the previous section we identified SuperDARN as one example of an under-utilized geosciences dataset because of the difficulties in extracting meaningful quantitative information about complex features in the data stream. The difficulty is not so much that the features cannot be extracted, but rather the expertise in advanced computer-aided data analysis techniques does not reside with the domain geoscientists who are custodians of the dataset. Again, this is likely to be a common problem across the geosciences. If EarthCube is to become more than just a comprehensive geosciences data clearing house it is imperative that the cyberinfrastructure includes sophisticated new tools for interactive visualization and mining of multiple datasets. Such an effort will require an intellectual partnership between the geosciences and computer science research communities. At Virginia Tech such a collaborative partnership already exists between the SuperDARN HF radar group in the Department of Electrical and Computer Engineering and the Data Mining Group in the Computer Science Department. Over the past 12 months, the two groups have forged an active collaboration with the specific objective to develop new algorithms to routinely identify and extract quantitative information about gravity waves and other complex spatiotemporal features in the SuperDARN dataset. In this section we describe some of the data mining tools that are currently being applied to SuperDARN data and which could quickly be adopted by EarthCube. We also provide some suggestions for how development of new data mining algorithms might proceed in the future.

Background: Knowledge Discovery and Data Mining (KDD)

Knowledge discovery and data mining (KDD) [*Fayyad and Uthurusamy,* 2002] is a thriving sub-discipline of computer science that aims to extract interesting and actionable patterns from multi-dimensional datasets. KDD techniques are typically aimed at "unstructured" discovery tasks, such as finding *all* patterns (or "motifs") of a certain class of phenomenon from a given dataset, rather than an easily recognizable one. Examples include repeatable stock market fluctuations, or frequently purchased items in a supermarket, or recurrent idioms from a text corpus. The key advantages of KDD techniques are (1) scalability and (2) the ability to find patterns without explicit supervisory direction. The past 15 years has witnessed tremendous growth in the adoption of modern KDD techniques in several scientific domains [*Han et al.,* 2002]. Much of this growth has been driven by the availability of massive digital repositories such as the human genome, the Sloan digital sky survey, and the explosion of online media. Some geosciences communities have embraced the data rush and started to incorporate the latest algorithmic methods for data analysis to actively drive their evolution [e.g. *Das and Parthasarathy,* 2009; *Ganguly and Steinhauser,* 2008; *Braverman and Fetzer*, 2005]. These previous studies have acknowledged the necessity of automating large scale data and image understanding, but have also identified significant challenges. Approaches focused on "supervised learning" (e.g., artificial neural networks and logistic regression) require training data which is costly to obtain and might not readily generalize to new classes of phenomena outside the purview of the training dataset. The breadth of patterns manifested in geosciences data is potentially broad, encompassing oscillatory, repeating, patterns of temporal behavior across years of aggregated data recordings; or vortices over particular regions of interest in a narrower time frame; or complex interdependencies between different variables that manifest only at certain times and under certain conditions. Writing specialized computer routines for each pattern of interest is practically unfeasible. What is needed is a "plug-and-play" compositional approach to data mining that enables the geoscientist to interactively define new classes of patterns from simpler building blocks and identify such patterns in the data without explicit supervisory input. This approach will facilitate the application of knowledge discovery techniques to geosciences datasets which by their very nature are impractical to traditional feature extraction, either by visual inspection or by automated search. The Virginia Tech data mining group has the specialized KDD expertise necessary to design such an interactive data mining framework for EarthCube, having been successfully engaged in mining scientific datasets for the past 12 years, in diverse applications such as computational fluid dynamics, aircraft design, data center chillers, land use change analysis, wireless communications, and bioinformatics. Algorithms have been developed for identification of well known pattern classes [*Ramakrishnan et al.* 2009], as well as new undiscovered classes of patterns [*Ramakrishnan et al.* 2004] that are of interest to domain scientists. Key innovative contributions that can be quickly applied to EarthCube include data mining expertise using multi-level spatial aggregates [*Ramakrishnan and Bailey-Kellogg.* 2003], compositional mining of multi-relational datasets [*Jin et al.* 2008], and temporal motif mining [*Patnaik et al.* 2009].

EarthCube Data Mining Design, Development, and Integration

The development of a new interactive framework for mining EarthCube datasets will involve application and extension of several data mining approaches. In this section we introduce but a few broad approaches that we see as particularly relevant. The goal will be to extract the underlying spatiotemporal coherences which are embedded both within a single EarthCube dataset and across multiple datasets. A trained scientist can often discern matches between features across data slices and correlate them, but automating this on a large scale is necessary to extract the fullest information embedded in the massive EarthCube database.

Dimensionality Reduction by Feature Extraction

One obvious benefit that will be produced by compiling a large number of geosciences datasets into one monolithic EarthCube will be the new opportunities to do multi-instrument data analysis. However, the sheer volume of data will provide significant computational challenges. One simple data mining approach that can be applied to reduce the dimensionality of the EarthCube will be "feature extraction". This will be particularly valuable for detecting high-level flow features (e.g. vortices and turbulence in vector field measurements). Given a flow field, it is relatively easy for a human eye to visually inspect and aggregate flow directions in order to identify vortices. The challenge will be to mimic such imagistic reasoning in a computational algorithm. EarthCube feature extraction algorithms could build upon experiences with the Spatial Aggregation Language (or SAL) [*Bailey-Kellogg and Zhao,* 2004] to define a small set of generic operators parameterized by domain-specific knowledge. SAL works using an "aggregate-classify-redescribe" methodology and can be applied to both time-domain and frequency-domain features. These operators systematically transduce the spatiotemporal fields into higher-level structures (e.g., velocity vectors into isobar cells into curve segments into troughs into recognizable geophysical features). This approach is more advantageous than defining one monolithic feature detection algorithm because it allows the construction of modular and reusable routines that can use similar processing techniques at different levels of abstraction. Through the systematic development of such building blocks, EarthCube users will be able to evaluate and re-use parts of such aggregators to build their own feature detection routines for identification of particular geophysical phenomena.

Interpolation of Sparse Data by Data Driven Surrogates

Feature extraction is a useful method for reducing the dimensionality of a dataset but it generally works best on data that is regularly gridded. Many geosciences datasets are sparsely populated in space and/or time and hence present significant challenges to the application of traditional feature extraction algorithms. This difficulty can be overcome by first building data-driven surrogates using methods such as Gaussian Processes (or GPs) [*Rasmussen and Williams,* 2005]. Gaussian processes are sophisticated interpolation mechanisms that aim to fit not a desired functional form but a desired covariance function. The ability of GPs to supply suitable dense representations for SAL has already been demonstrated in domains such as aircraft simulation and wireless network characterization [*Bailey-Kellogg and Ramakrishnan,* 2001]. For EarthCube it will be necessary to explore both stationary and non-stationary covariance functions and aim to capture discontinuities. To address these requirements it will likely be necessary to explore mixtures of GPs [*Meeds and Osindero,* 2006]. Accommodating multiple outputs will require the exploration of dependent Gaussian processes [*Boyle and Frean,* 2005] and Bayesian networks factorizing conditional independencies between GPs [*Friedman and Nachman,* 2000]. Each of these formulations has been considered previously in separate contexts but there has been no systematic integration of all of these methodologies in one framework for mining massive spatio-temporal datasets. EarthCube will provide a basis for this sort of integrative data mining.

Identification of Interconnected Geophysical Activity by Subclustering

Subclustering will be a useful KDD tool for unraveling the interconnectivity within and between the various datasets that comprise the EarthCube. Clustering is a well studied technique: it partitions a dataset into different regions of density. Subclustering [*Kriegel et al.* 2009] aims to identify both the right dimensions and clusters along those dimensions. Subclustering has been shown to be very useful in all multidimensional contexts and has high potential to quickly identify "hotspots" of interconnected geophysical activity deeply embedded in the EarthCube. However, again, it is likely that there will be complications associated with the spatiotemporal sparseness and complexity of many geosciences datasets. This is because subspaces are usually defined using simple axis-parallel/isothetic notions or by an arbitrarily oriented subspace (e.g., linear projections or a non-linear embedding). For sparse datasets the space of such subcluster definitions is likely to be as broad as the search for subclusters themselves. Traditional density-based and correlative definitions for subclusters may therefore prove to be of insufficient expressiveness in the geosciences domain. Here again, it will likely be necessary to exploit the Gaussian Process (GP) framework to help organize the EarthCube database and identify regions of promising interest by the use of optimizing functionals. In particular, the GP machinery can be embedded in a trust-region framework to quickly narrow down on important subslices which can then be subjected to feature detection. Once these subslices are identified, a pattern-based approach could be used to generalize across them in order to detect general "transferable" trends across space/time. Together, the use of subclustering and feature detection will lead to an increased supply of higher level feature streams for consumption by the EarthCube user community.

Recurrent Pattern Identification by Motif Mining

Another KDD technique that will be particularly valuable for identifying recurrent patterns in multidimensional EarthCube datasets is "motif" mining. Motifs can be defined as repetitive patterns of occurrence in higher level tranduced features across time and/or space [*Mueen and Keogh*, 2010]. By contrast, classical frequency domain decomposition techniques are limited in the types of repetitive patterns they can mine. In principle, motif mining should be directly applicable to EarthCube datasets; but again, the massive dimensionality will present unique challenges. However, there are many approaches that can be used to create symbolic representations of the geosciences datasets, and thus reduce their dimensionality. For instance, it should be possible to first perform clustering of the multivariate series and then use the sequence of cluster identifiers as an abstract symbolic representation of the system state. This allows the level of abstraction of the symbol sequence to be raised by encoding the transitions from one symbol to another. Another strategy is to use features extracted from the first stage to provide a greater level of expressiveness for symbol definition. This will provide a sequence of events which can then be mined for repeating patterns using serial episode discovery algorithms. The mining process would then follow the level-wise procedure ala *Apriori*, (i.e., candidate generation followed by counting). The count or frequency measure will be based on non-overlapped occurrences [*Patnaik et al.* 2009]. It should be noted here that the events in the mined patterns correspond to transitions from one symbol to another in the abstract symbolic representation of the time-series data. The motif occurrences represent matching time series subsequences and in addition the episode mining framework allows for robustness to noise and scaling in terms of finding matching subsequences. Motif mining will be valuable for identifying the onset of significant changes in the phase state manifested in EarthCube datasets (e.g. onset of El-Nino or auroral substorm); however, it will be particularly useful in searching for other, as yet undetected, changes in the collective state of the space environment that may not be as noticeable as larger events.

Future Plans: Development of an Interactive Data Mining Framework for Research and Outreach

In the previous section we provided a brief sketch of four broad KDD approaches that could be quickly adopted for efficiently mining EarthCube datasets. However, EarthCube provides new possibilities for transforming not only geosciences - but data mining in general, as well. More specifically, a comprehensive data mining architecture for EarthCube would use all of the above developed taxonomy of pattern classes and mining algorithms (in addition to others!) to define an event model that is sufficiently expressive to allow for true interactive spatiotemporal process discovery in geosciences. A geosciences "event" is a happening of interest and can be stated in terms of either the original data (e.g., a sensor reading exceeding a pre-set threshold) or over a pattern class (e.g., oscillations of frequency greater than a given threshold or an episode that recurs in a symbolic stream). Ideally, it should be possible to leverage existing event description languages from the simulation and data engineering communities to create a representation that will: (1) support the capture of spatial and temporal coherence, (2) model causality information when available, and (3) record large scale associations between events and even parts of events. The goal of this representation would be to support a "process query capability" where the administrator can query streams based on high-level properties derived from the event modeling rather than the low-level sensor data. The event models developed would also directly support simple capabilities like dynamic time warping to compare time series, indexing/retrieval of segments, and piecing together segments to define hypothetical scenarios of usage. The end result would be an entirely novel interactive capability to analyze EarthCube datasets at significantly higher levels of abstraction than the individual data streams. Indeed, one can imagine an EarthCube interface that will allow researchers to interactively mine the datasets by tuning algorithm parameters to pull out features of interest and then retroactively searching the data archive for recurring instances, then downloading an extracted data stream and associated graphical outputs. Such a system would significantly enhance the utility of EarthCube because data users would be able to do true interactive data analysis, on the fly, rather than merely browsing for data. Such a system would also make geosciences data immediately accessible to students of all ages, allowing them to become "citizen scientists".

References

Bailey-Kellogg, C., N. Ramakrishnan: Ambiguity-Directed Sampling for Qualitative Analysis of Sparse Data from Spatially-Distributed Physical Systems, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 43-50, 2001.

Bailey-Kellogg, C., F. Zhao: Qualitative Spatial Reasoning Extracting and Reasoning with Spatial Aggregates, *AI Magazine* 24(4): 47-60, 2004.

Baker, J.B.H., R.A. Greenwald, J.M. Ruohoniemi, K. Oksavik and J.W. Gjerloev, L.J. Paxton, M.R. Hairston, Observations of ionospheric convection from the Wallops SuperDARN radar at middle latitudes, 112, A01303, doi:10.1029/2006JA011982, *J. Geophys. Res.*, 2007.

Boyle, P., and M. Frean, Dependent Gaussian processes. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 217-224. The MIT Press, 2005.

Braverman A. and E. Fetzer, Mining Massive Earth Science Data Sets for Large Scale Structure,Proceedings of the Earth-Sun System Technology Conference, a3p1, 2005.

Bristow WA, Greenwald RA, Villain J-P (1996) On the seasonal dependence of medium-scale atmospheric gravity waves in the upper atmosphere at high latitudes. *J Geophys Res*., 101:15685–15699

Chisham et al, A decade of the Super Dual Auroral Radar Network (SuperDARN): scientific achievements, new techniques and future directions, *Surveys in Geophysics*, 28, 33-109, doi:10.1007/s10712-007-9017-8, 2007.

Das, M., S. Parthasarathy, Anomaly detection and spatio-temporal analysis of global climate system, *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, 2009.

Fayyad, U. M., R. Uthurusamy: Evolving data into mining solutions for insights, *Communications of the ACM*, 45(8): 28-31, 2002.

Friedman, N., and I. Nachman. Gaussian process networks, in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 211-219, Morgan Kaufmann, 2000.

Ganguly, A.R., and K. Steinhaeuser, Data Mining for Climate Change and Impacts, *Proceedings of the IEEE International Conference on Data Mining Workshops*, 2008

Greenwald, R. A., K. B. Baker, R. A. Hutchins, and C. Hanuise, An HF phased-array radar for studying small-scale structure in the high-latitude ionosphere, *Radio Science, 20*, 63-79, 1985.

Hall, G. E., MacDougall, J.W., Moorcroft, D. R., St.-Maurice, J.-P., Manson, A. H., and Meek, C. E.: Super Dual Auroral Radar Network observations of meteor echoes, *J. Geophys. Res.*,102(A7), 14 603–14 614, 1997.

Kriegel, H.-P., P. Kröger, A. Zimek: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, *ACM Transactions on Knowledge*, Vol. 3(1), 2009.

Malinga, S.B., and J.M. Ruohoniemi, "The quasi-two-day wave studied using the Northern Hemisphere SuperDARN HF radars," *Ann. Geophys.*, v.25, p. 1767, 2007.

Meeds, E., and S. Osindero, An alternative infinite mixture of Gaussian process experts. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 883-890, The MIT Press, Cambridge, MA, 2006.

Mueen, A., E. J. Keogh: Online discovery and maintenance of time series motifs, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1089-1098, 2010.

Patnaik, D., M. Marwah, R. K. Sharma, Naren Ramakrishnan: Sustainable operation and management of data center chillers using temporal data mining, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1305-1314, 2009.

Ramakrishnan, N., C. Bailey-Kellogg: Gaussian Process Models of Spatial Aggregation Algorithms, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1045-1051, 2003.

Ramakrishnan, N., D. Kumar, B. Mishra, M. Potts, R. F. Helm: Turning CARTwheels: an alternating algorithm for mining redescriptions, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 266-275, 2004.

Ramakrishnan, N., D. Patnaik, V. Sreedharan: Temporal Process Discovery in Many Guises, *IEEE Computer*, 42(8): 97-101, 2009.

Rasmussen, C.E. and C.K.I. Williams, *Gaussian Processes for Machine Learning*, *MIT Press*, Dec 2005.

Ruohoniemi, J. M., Greenwald, R. A., Baker, K. B., Villain, J. P., and McCready, M. A.: Drift motions of small-scale irregularities in the high-latitude F region - An experimental comparison with plasma drift motions, *J. Geophys. Res*., 92, 4553-4564, doi:10.1029/JA092iA05p04553, 1987.