LLNL-TR-416780

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Machine Learning and Data Mining for Comprehensive Test Ban Treaty Monitoring

S. Russell, S. Vaidya

September 14, 2009

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Machine Learning and Data Mining for Comprehensive Test Ban Treaty Monitoring

Stuart Russell, Computer Science Division, University of California, Berkeley, CA
Sheila Vaidya, Global Security Directorate, Lawrence Livermore National Labs, CA

Abstract

The Comprehensive Test Ban Treaty (CTBT) is gaining renewed attention in light of growing worldwide interest in mitigating risks of nuclear weapons proliferation and testing. Since the International Monitoring System (IMS) installed the first suite of sensors in the late 1990's, the IMS network has steadily progressed, providing valuable support for event diagnostics. This progress was highlighted at the recent International Scientific Studies (ISS) Conference in Vienna in June 2009, where scientists and domain experts met with policy makers to assess the current status of the CTBT Verification System.

A strategic theme within the ISS Conference centered on exploring opportunities for further enhancing the detection and localization accuracy of low magnitude events by drawing upon modern tools and techniques for machine learning and large-scale data analysis. Several promising approaches for data exploitation were presented at the Conference. These are summarized in a companion report. In this paper, we introduce essential concepts in machine learning and assess techniques which could provide both incremental and comprehensive value for event discrimination by increasing the accuracy of the final data product, refining On-Site-Inspection (OSI) conclusions, and potentially reducing the cost of future network operations.

Introduction

The International Monitoring System (IMS) of the CTBTO is comprised of physical sensor stations (seismic, hydroacoustic, and infrasound) connected by a worldwide communications network to a centralized processing system in Vienna, the International Data Center. The IDC operates continuously and in real time, performing *station processing* (analysis and reduction of raw seismic sensor data to detect and classify signal arrivals) and *network processing* (association of phase arrivals with hypothesized events). Fully automated processing of the signals to produce a reliable catalogue of event reports is currently beyond the state of the art, so the IDC analysts must post-process the output from the automated system to generate higher quality event bulletins for further distribution. Errors in automated processing include false detections and missed detections caused by station noise; incorrect classification of arrivals; and incorrect associations. Thus, opportunities exist at all levels of the IDC pipeline to apply techniques from *machine learning* to improve the accuracy of the final output.[1]

We begin by explaining the basic ideas of machine learning, with special emphasis on *data-driven* and *model-driven* methods. We clarify how these methods may be applied to improve the performance of various parts of the IDC processing pipeline. Multiple teams at the ISS Conference presented preliminary results that demonstrated improvements in phase classification as well as rejection of spurious associations via some of these methods (Kuzma *et al*., 2009).

The second section of the paper proposes a more radical revision of the IDC data processing approach using a model-driven Bayesian methodology to perform probabilistic inferencing from the signal evidence with a vertically integrated probabilistic model of the entire signal generation process (from event to waveform). This approach has several potential advantages, including globally optimal association; proper handling of non-detections as evidence; improved low-amplitude signal detection and noise rejection; continually self-calibrating sensor model; and optimal fusion of multiple sensor modalities.

We conclude that incorporating machine learning methods into the IDC framework could indeed improve the detection and localization of low-magnitude events, provide more confidence in the final output, and reduce the load on human analysts. The principal obstacles to rapid instantiation of machine learning methods within an operational context, however, are the availability of raw data for testing during algorithm development and the difficulty of evaluating and benchmarking the impact of local improvements on the overall system. We outline a programmatic construct for overcoming these hurdles by proposing to coordinate and drive data-related R&D initiatives through a virtual Data Exploitation Center (vDEC), under the auspices of the CTBTO, for the evolution and prove-in of next generation data processing methods for CTBT verification.

Machine learning

The field of machine learning covers all computational methods for improving performance based on experience. The range of methods and settings is too vast to be sketched here in completeness, but there is a small set of key questions that must be answered to constrain the possibilities for choosing a learning method:

- Which component of the overall system must be improved?
- How is that component represented – *e.g.*, a weighted linear function, a complicated decision tree, or an impenetrable chunk of machine code?
- What existing data are relevant to that component?
- Do the data include the “right answers” – *i.e.*, correct outputs for the component given the inputs?
- What knowledge is already available to constrain and inform the design of the component?

This paper examines just two families of methods. The first, *supervised model-free learning* is appropriate for cases where data are plentiful and correct outputs are available, but little is known about the correct design of the component(s). The second, *Bayesian model-based learning* is effective in situations when significant prior knowledge is available but does not require advance knowledge of the correct outputs for each component.

Supervised model-free learning

The key idea here is many hundreds of years old: find a hypothesis that maximizes some combination of simplicity and degree of fit to the data. For example, suppose the component to be learned is responsible for classifying detected seismic signals as P waves or S waves. An unknown function f determines the true classification given the signal. In the supervised setting, we assume that the learning algorithm is provided example signals x_i along with the correct label $f(x_i)$ for each – perhaps obtained from the final Reviewed Event Bulletin (REB) or other authoritative source. The goal of learning is then to find a hypothesis h that is "close" to f in a precise sense: given a sufficient *training set* of examples, h should agree with f on the classification of almost all members of a previously unseen *test set* of examples (that are supplied without labels). The framework of machine learning provides theoretical guarantees on the ability of learning algorithms to meet this criterion and predicts the amount of data required to be effective.

This seemingly simple task encompasses a large range of activities, roughly characterized by the nature of the inputs, outputs, and the family of hypotheses considered. For example, x_i might be a purported sentence of English, $f(x_i)$ the label "ungrammatical," and h a grammar; or x_i might be an image, $f(x_i)$ the label "giraffe," and h a kernelized linear separator applied to the outputs of a fixed battery of feature extractors on the image. Other popular hypothesis classes include decision trees, neural networks, logistic regression functions, nearest-neighbor classifiers, and various forms of ensemble classifiers that generate and combine multiple hypotheses.

Supervised machine learning methods are readily applicable to IDC data sets for assisting the final diagnosis. Such methods were illustrated at the ISS Conference last June. Several posters showed the value of incorporating off-the-shelf learning and classification methods to improve the accuracy of phase discrimination in station processing and to detect spurious events proposed during network processing. Examples of the benefits from data fusion were plentiful, and design concepts were presented for improving seismic database query processing, borrowing ideas from the Web-search environment. The Best Poster award at the Conference went to a team that trained neural networks to detect false events in the SEL1 bulletin. These approaches, many of which could significantly enhance the current IDC pipeline, are elaborated upon in a separate report (Kuzma *et al*., 2009).

However, none of these supervised learning methods, as currently conceived, are likely to overcome the fundamental limitations of bottom-up, localized processing of signals and detections. Seismic data analysis, on a global scale, cannot decompose into *independent* local decisions about detections and associations; the ambiguities inherent in the data are best resolved by a comprehensive analysis of the kind offered by integrated probabilistic inference methods. Moreover, such methods can easily integrate the best earth models as well as detailed models of sensor artifacts and failures, and missing data. Such an approach is discussed in the following sections.

Bayesian model-based learning

When there is substantial prior knowledge available – for example, that of seismic phases and signal propagation – this knowledge can often constrain the space of hypotheses considered and thereby improve prediction accuracy and reduce data requirements.
An approach that achieves these goals is Bayesian model-based learning. When applied to problems involving sensor data, two generative models are developed:

- $P_\theta(world)$ describes a prior distribution over events of interest in the world; it may include a prior over the model parameters and structure, allowing these to be updated in the light of additional data.
- $P_\phi(signal \mid world)$ describes the sensor model, *i.e.*, the process by which events in the world generate sensor measurements.

Given a signal, we can compute a posterior distribution over the events of interest given that signal, according to Bayes' Rule:

$$P(world \mid signal) = \alpha P_\phi(signal \mid world)\ P_\theta(world)$$

where α is a normalization constant. Of course, the complexities of actual models mean that this computation is often far from trivial.

Typically, Bayesian learning methods can continuously adapt the model parameters to improve the degree of fit to the data, as a side effect of performing the inferences required to interpret the data according to the equation above. This adaptation requires no "ground truth" (unlike supervised learning methods) and hence provides a technical foundation for continuous self-calibration and sensor diagnostics.

Speech recognition is perhaps the best-known example of Bayesian model-based learning and inference. In speech recognition, $P_\theta(world)$ is a generative model of word sequences and $P_\phi(signal \mid world)$ is a generative model of acoustic features given words, mediated by complex pronunciation models for words in terms of their constituent sounds. The parameters of such models are estimated from thousands of hours of speech data, leading to very high performance in many commercial applications. Interestingly, training on isolated words does not work, because in real speech, the pronunciation of a word depends strongly on the words preceding and following it due to physical constraints on the motion of the lips, tongue, jaw, etc. Because of these low-level inter-word dependencies, as well as high-level constraints on plausible word sequences, each word helps disambiguate other words. Thus, a bottom-up approach does not work for speech signals; and as we will see, the same lesson applies for seismic analysis.

Vertically integrated seismic analysis

While the current IDC data analysis pipeline is functioning effectively, we believe that its overall serial nature imposes unnecessary limitations on system performance that can be largely overcome by a *vertically integrated* probabilistic approach. Recent advances in modeling capabilities and in general-purpose inference algorithms such as Markov Chain Monte Carlo (MCMC) suggest that it is in fact possible to address problems as complex as nuclear detonation detection via a completely integrated, model-based probabilistic system derived from first principles. A research prototype system (VISA-CV, for Vertically Integrated Seismic Analysis for CTBT Verification) is currently under development with the goals of testing it within the IDC domain (Arora *et. al*., 2009a, 2009b). The VISA-CV generative model begins with a two-component generative process for events: natural events occur according to a Poisson process with spatially varying intensity (augmented with secondary processes for aftershocks) and a Gutenberg-Richter magnitude distribution, while man-made events are assumed to occur with a uniform spatial distribution. Signals from the events propagate according to a travel-time model (initially IASPEI91 with added Gaussian uncertainty) and are selectively attenuated in different frequency ranges and phases. The detected signal consists of arriving waveforms and local noise, modeled by station-specific noise models and response functions. Sources of uncertainty include local station noise, drifting station response function, miscalibration and sensor malfunction, uncertainty in the waveform travel time model, and frequency and phase absorption, all of which are spatially varying; as well as uncertainty in the model of event and aftershock locations, times, magnitudes, and waveform generation. All these uncertainties can be represented explicitly and estimated over time.

Once data samples – currently, just the IDC arrival detections but eventually the full waveforms – are supplied to the system, MCMC probabilistically infers a posterior distribution over possible event locations, times, and magnitudes. In essence, MCMC efficiently samples over hypothetical worlds to obtain estimates that converge to the true posterior given the evidence (see Figure 1). The fact that MCMC computes posterior probabilities – the best possible answers given the data – takes the algorithm itself off the table; to get better answers, one must either improve the model or add more sensors.

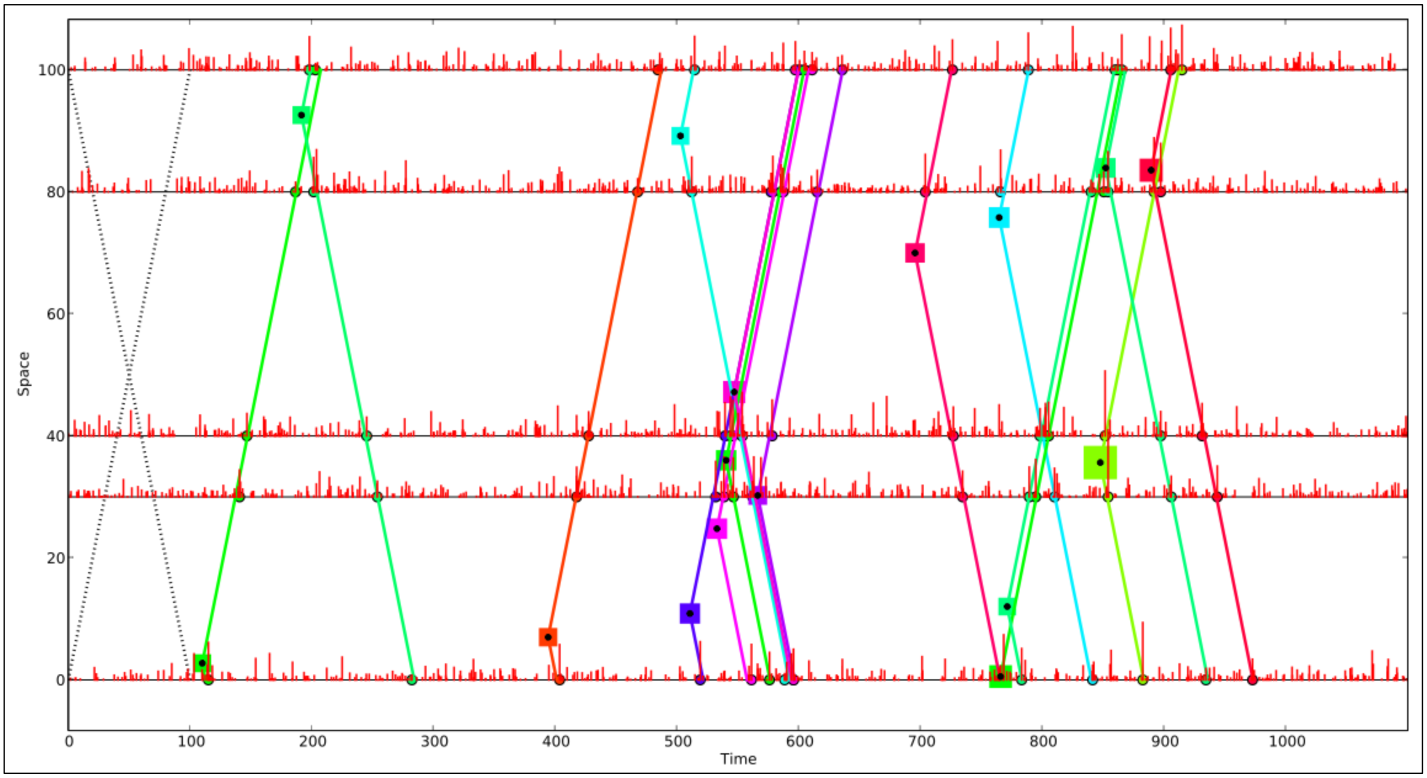

Figure 1: A sample from the VISA-CV MCMC inference process applied to seismic traces from a simulated one-dimensional world. The x-axis shows time and the y-axis shows position in the 1-D space. For the five stations considered here, each signal trace is shown as a series of impulses of different magnitude. Hypothesized events are shown as squares with area proportional to event magnitude; waves propagated along the color-coded rays, generating impulses when they intersect the stations. MCMC generates samples by adding, deleting, and moving events, adjusting propagation times and arrival magnitudes, and changing associations among events and detections. Although the vast majority of the detected impulses are noise-induced – some much larger than the real detections – the system is able to recover the true events correctly.

One important benefit of the vertically integrated approach is that signals need not be analyzed at each station in isolation. Suppose that a hypothetical event has been formed from detections at three other stations, such that the event's location, time, and magnitude imply an arrival at a fourth station in the time interval [t - δt, t +δt]. If a signal is present – even well below the usual SNR threshold – it can be picked and associated with the event. On the other hand, if no signal is present, the event is disconfirmed by the (absence of) evidence. The smaller the value of δt, the more pronounced this effect will be. Thus, a strong, and thus far unexploited, interaction exists between the accuracy of the travel time model and the ability to pick signals from noise at a particular station. This interaction is demonstrated empirically in the simulated model used in Figure 1.

The VISA-CV research prototype has been tested only on a small 2-hour segment of *parametric* data from the IDC (*i.e.*, above-threshold P-wave detections, rather than raw waveforms). The segment includes three events that generated 3 or more arrivals, and the prototype recovers all three perfectly. In comparison, the IDC SEL3 bulletin includes three additional events which are not well supported by the evidence (see Figure 2).

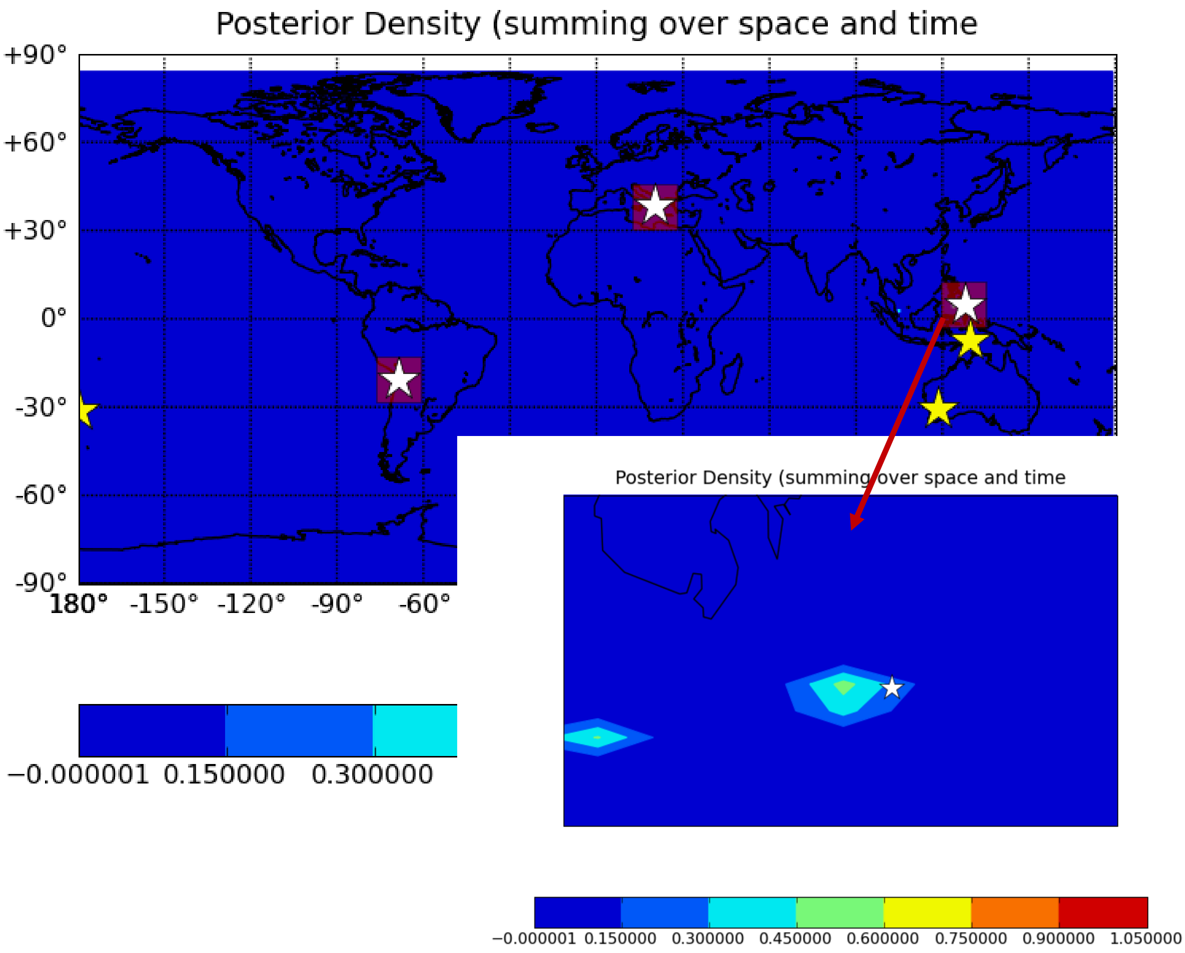

Figure 2: Display showing preliminary results from the prototype VISA-CV system. White stars indicate true events, yellow stars indicate additional spurious events proposed in SEL3, and red squares show events proposed by our research prototype. The inset shows the posterior event distribution near the Sulawesi coast; the posterior is *bimodal* due to uncertainty in the association between events and detections.

vDEC

Based upon the above discussions, we believe that the CTBTO could benefit greatly from a strategic thrust focused on improving techniques for processing IMS and OSI data sets, taking into consideration the state of the art in machine learning, the advances in data structures and query techniques, and the shaping of sensor data for more accurate exploitation and inference. The long term goal of such an effort should be to assist the CTBTO analyst in making more robust and expedient decisions, aided by a historical perspective, in the face of rapidly growing multi-sensory information and the imperative for more accurate and timely event characterization.

To facilitate such an endeavor, we propose a virtual Data Exploitation Center (vDEC), which will connect international experts (both academic and commercial) in different disciplines with the IDC/OSI framework, to assess, develop and implement upgrades to the current data processing infrastructure for event detection and localization. We envision such a construct will seamlessly tie in to other National Data Centers to access the largest available data sets for training algorithms and cross-checking results, and incorporate the best practices from multiple sources. vDEC's charter will be to advance the state-of-the-art in data processing in coordination with the operational arm of the IDC so as to provide a smooth transition from research into the production environment. A viable business model for vDEC is in discussion.

Conclusions

We have briefly summarized applications of machine learning to CTBT verification, including near-term improvements to components of the current IDC pipeline, as suggested by several posters in the June ISS Conference, as well as a more substantial architectural overhaul based on vertically integrated probabilistic models that connect underlying seismic events to measured signals. Such models could improve seismic phase classification, identify spurious associations through global optimization, characterize station drift/noise, use the absence of detections to disconfirm hypotheses, perform time-localized "sub-threshold" signal detections, combine multiple inputs, and cumulatively, lower the threshold for event detection and localization. Taken a step further, continuous sensor self-calibration could lead to better sensor design and layout and potentially mitigate cost of future network operations.

To coordinate and guide machine learning and data exploitation methods development in support of Treaty verification, we propose a focus center (vDEC) under the CTBTO umbrella, which will leverage multidisciplinary expertise to incubate, test and evolve next generation data solutions for IDC/OSI missions.

References

The popular term *data mining* largely overlaps with data-driven machine learning methods. The fields of *statistics* and *pattern recognition* also cover some of the same territory.

Nimar S. Arora, Michael I. Jordan, Stuart Russell, and Erik B. Sudderth (2009a), "Vertically Integrated Seismological Analysis I: Modeling." Poster abstract, CTBTO International Scientific Studies Conference, Vienna, June 2009.

Nimar S. Arora, Michael I. Jordan, Stuart Russell, and Erik B. Sudderth (2009b), "Vertically Integrated Seismological Analysis II: Inference." Poster abstract, CTBTO International Scientific Studies Conference, Vienna, June 2009.

Heidi Kuzma, Sheila Vaidya, and Ronan Le Bras (2009), "Data Mining for CTBT Verification." In this volume.