

High Energy Physics

1. Track Finding, Clustering, Classification, Fitting and Parameter Estimation

Given a set of position and/or momentum measurements provided by a tracking detector, group these measurements together in subsets, each subset containing measurements originating from one charged particle.

For a given set of tracks, track-fitting problem is to find the optimal estimate of a set of parameters uniquely describing the state of the particle somewhere in the detector. The most widely used technique involves some sort of Kalman filtering, a linear recursive method shown to be equivalent to global least square minimization. Provided the track model is truly linear, and measurement errors are Gaussian, the Kalman filter is efficient and optimal. Because it is recursive, it is well-suited for progressive track finding and fitting, and does not involve large matrix computations.

Parallel and combinatorial extensions of the Kalman filter are used to simultaneously build track trajectories, smooth tracks, and estimate parameters; trajectories are typically propagated in parallel. More recently, these deterministic track-propagation algorithms are being compared to adaptive, probabilistic trajectory-building techniques that a better handle dense track environments. A variety of other techniques have been applied to this problem historically, including Hough transforms, decision trees, and neural networks.

The HEP community would like to better understand which techniques should have superior performance, and, ultimately, how to design the track detection sensors in the first place to provide the optimal information to the pattern recognition algorithm(s).

Jim Siegrist, LBNL

F.-P. Schilling, Track Reconstruction and Alignment with the CMS Silicon Tracker, Poster presented at the 33rd Intl. Conference on High Energy Physics, ICHEP 2006 (Moscow, July 2006)

A. Strandlie, Adaptive methods with application to track reconstruction at LHC, CSC 2003.

Climate Modeling

1. Detection of Extreme Events

The quantification of rare, episodic, extreme events or anomalies that are localized in space and time, such as tropical cyclones, volcanic eruptions, and flash floods, are valuable for understanding how global patterns of change may be linked to thresholds triggered on multiple interacting variables. There is a need for efficient, parallel methods for detecting such events in large data sets. Once these events are detected, it would be useful to study the relationships between other global variables and the quantity and intensity of the extreme events, and to draw conclusions about the role played by external and natural forcings.

2. Spatial/Temporal Patterns of Variability

Spatial multivariate analyses such as empirical orthogonal functions (EOF, or PCA) are often performed on simulated data to extract patterns or “fingerprints” of one or more climate variables that may be matched to observed data for model validation. Various extensions of EOF, as well as canonical correlation analysis (CCA), are useful for studying the joint variability of two or more interacting climate data fields.

In the temporal domain, researchers may want to separate periodic oscillations from random fluctuations for one or more climate variables, discover episodic events, or detect propagating structures with both temporal and spatial coherence. Existing methods include the multi-taper method (MTM), singular spectrum analysis (SSA), and wavelet decompositions. In the spatiotemporal domain, extensions of EOF that concatenate time-lagged examples at multiple locations (e.g., EEOF, MSSA) can create potentially enormous data matrices which would benefit from the capability to execute large-scale, parallel eigenvalue computations.

3. Detrending and Signal Separation

Long-term climate simulations may have linear or nonlinear trends due to model deficiencies, and these must be separated from the true signal before further analysis may be done. While linear trends are easily fitted and subtracted, nonlinear trends are less difficult to separate from signal. Furthermore, these trends may vary spatially, so procedures may need to operate on the spatial and temporal domains simultaneously. Blind Signal Separation (BSS) methods may be appropriate for these problems.

H. V. Storch and F. Zwiers. Statistical Analysis in Climate Research. Cambridge University Press, Cambridge, UK, 1999.

Ghil M., R. M. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. Robertson, A. Saunders, Y. Tian, F. Varadi, and P. Yiou, 2002: "Advanced spectral methods for climatic time series," Rev. Geophys., 40(1), pp. 3.1-3.41, 10.1029/2000RG000092.

Astrophysics

1. Feature Extraction & Parameterization of Supernova Spectra

Given a set of observed and/or synthetic spectra of Type Ia supernovae, find a low-dimensional representation that has physically interpretable coordinates, e.g. phase, luminosity. Currently, SNe Ia are classified according to a single-parameter family of peak luminosity, as a function of light-curve shape, but this parameter does not completely describe the diversity of observed SNe Ia. Finding a low-dimensional parameterization or projection of SN Ia spectra would be useful for classifying future observations into classes exhibiting more constrained magnitude-redshift relations, identifying stable spectral regions for the design of future cosmology probes. Current statistical methods do not go much further than PCA, which do not necessarily have meaningful physical interpretations. Other linear or nonlinear dimensionality reduction methods may be appropriate.

2. Analysis and Matching of Time-Varying Spectra

Time-varying synthetic spectra could be useful for learning the temporal dynamics of how a SN spectrum changes over time. However, observational sets of spectra from a given SN are incomplete, captured only at several time points.

The temporal evolution of synthetic spectra might be used to 1) match a series of sparsely observed spectra from the same SN to a time series of synthetic spectra, improving over current methods that match only individual spectra, 2) infer the trajectory through some feature space that the observed SN spectra takes; this space could in theory be the original multidimensional space of measured flux at multiple wavelengths, but more likely a lower-dimensional space parameterized by time.

J.B. James, T.M. Davis, B.P. Schmidt, A.G. Kim, Spectral diversity of Type Ia supernovae, Mon. Not. R. Astron. Soc. 370, 933-940 (2006).

K. Glazebrook, A. R. Offer, K. Deeley, Automatic Redshift Determination by use of Principal Component Analysis – I: Fundamentals,

Genomics

Note: I picked a few recent papers as representative examples of machine learning research applications in genomics domain. This is not an exhaustive compilation just what I managed to scour in the little time I had during this weekend and based on my knowledge of this domain. In each area, each paper's abstract and some salient features are highlighted. I wished I had more time to devote, however. (Krishna Palaniappan, November 20, 2006)

1. Gene ranking and significance analysis in high-throughput DNA microarray data

An important application of DNA microarray technologies in functional genomics is to classify samples according to their gene expression profiles, e.g., to classify cancer versus normal samples or to classify different types or subtypes of cancer. Selecting genes that are informative for the classification is one key issue for understanding the biology behind the classification and an important step toward discovering those genes responsible for the distinction. Many methods for classification and gene selection with microarray data have been developed. These methods usually give a ranking of genes. Evaluating the statistical significance of the gene ranking is important for understanding the results and for further biological investigations, but this question has not been well addressed for machine learning methods in existing works. .

In this paper, authors (Zhang et al 2006) address this problem by formulating it in the framework of hypothesis testing and propose a solution based on re-sampling. Existing hypothesis testing methods are not sufficient to handle high-dimensional multivariate analysis problems arising from current high-throughput genomic and proteomic studies. Many machine-learning-based gene selection and classification methods may achieve very good performance in solving the specific classification problems, but the results are usually of a "black-box" type and judging the significance of the features being used for the classification was usually not deemed important. This fact compromises their further contribution in helping biologists to understand the mechanisms underlying the investigated disease classification.

Zhang, C. Lu X, and Zhang X. (2006) Significance of Gene Ranking for Classification of Microarray Samples. IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 3, NO. 3, JULY-SEPTEMBER 2006

2. Comparative Genome Analysis based on orthologous genes and functional roles

Functions of human genes are often studied indirectly, by studying model organisms such as the mouse. An underlying assumption is that so-called orthologous genes, that is, genes with a common evolutionary origin, have similar functional roles in both species. Exploration of dependencies (regularities and irregularities) in functioning of orthologous genes helps in assessing to which extent this assumption holds. In practice, gene pairs are defined as putative orthologs based on sequence similarity, and we seek for regularities and irregularities in their expression by associative clustering. An

exceptional level of functional conservation of an orthologous gene group may indicate important physiological similarities, whereas differentiation of function may be due to significant evolutionary changes

Authors of this paper (Kaski et al 2006), introduced a new approach for a relatively little studied machine learning or data mining problem: From data sets of co-occurring samples, find what is in common. They have formulated the problem probabilistically, extending earlier mutual information-based approaches. The new solution is better-justified for finite (relatively small) data sets. The introduced method, coined associative clustering (AC), summarizes dependencies between data sets as clusters of similar samples having similar dependencies. They claim this method is particularly suitable for mining functional genomics data where measurements are available about different aspects of the same set of functioning genes. Then, a key challenge here is to find commonalities between the measurements and the answer should reveal characteristics of the genes, not only characteristics of the measurement setups.

Kaski S, Nikkila J, Sinkkonen J, Lahti L, Knuuttila J E A, and Roos C (2006) Associative Clustering for Exploring Dependencies between Functional Genomics Data Sets IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 2, NO. 3, JULY-SEPTEMBER 2005

3. Detection of 'hidden' signals of biological relevance in genomic sequences.

Signals in genomic sequences refer to specific sites or small sequence segments that are directly related to transcription and translation processes or to their regulation. This paper by Rajapakse et al (2006) deals with computational techniques that identify signals in genomic sequences. Knowledge of the presence of signals in genomic sequences gives insight into transcription and translation processes and the location and annotation of genes, which are vital to the investigations of novel and effective drugs having minimal side effects. They address two important problems in signal detection, how to automatically identify the transcription start sites (TSS) and recognize the translation initiation sites (TIS) in eukaryotic DNA sequences, and another important problem in gene annotation: the determination of intron and exon boundaries or splice sites (SS). Combining the probabilistic and neural network approaches in a sensible way would lead to more efficient approaches in modeling and detecting signals. This paper introduces the Markov/neural hybrid approach to signal detection by introducing a novel encoding scheme for inputs to neural networks, using lower-order Markov chains. The lower-order Markov models incorporate biological knowledge differentiating the compositional properties of the regions surrounding the signals; the neural networks combine the outputs from Markov chains, which we refer to as Markov encoding of inputs, to derive long range and complex interactions among nucleotides that improve the detection of signals.

Rajapakse J C and Ho L S. (2006). Markov Encoding for Detecting Signals in Genomic Sequences. IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 2, NO. 2, APRIL-JUNE 2005

4. In-silico Classification and Prediction methods for Phylotype (taxonomic) assignment of Metagenome DNA Sequences

Motivation for this study stems from the current need to adequately “bin” or classify metagenome DNA sequences with respect to their species origin. Metagenome DNA sequences are directly sampled from a close knit community of microbes of mixed species population. A species is defined in terms of its taxonomic classification in the tree of life. An identified species is usually classified into its taxonomic hierarchical lineage of ranks starting with the root as: domain, phylum, class, order, family, genus and species, strain as the leaf most node. There exist several challenges unique to metagenome dataset. (1) the microbial community of the metagenome dataset is usually made up of heterogeneous species population (2) often, the sampled DNA sequences are short fragments (700 – 800 bp length) containing weak phylogenetic (ie. Taxonomic) signals which can not be assembled into larger contigs of belonging to the same organism (3) depending on the complexity of the community being sequenced, usually dominant or more abundant members tend to have higher coverage of DNA sequences whereas rare or less dominant members will have fewer low coverage DNA sequences. These low abundance members’ DNA sequences cannot usually be assembled into larger contigs, and they may remain as singletons or (so called “shrapnels”). Unlike, single genome sequencing where the sample origin is clonal and it’s species (phylotype) assignment is known a priori, conventional methods for genome assembly, gene prediction and annotation methods and eventual genome analysis do not perform well for metagenome dataset for reasons mentioned above. Thus the field of metagenome analysis is at its early stages requiring new and innovative tools and analysis methods to deal with large amount of sequence data that are often sparse, noisy and whose species origin is unknown. This study addresses the problem of classifying the DNA sequences arising from metagenome samples to their species origin.

(Excerpt from my project proposal for my class in Decision Support Systems, Spring 2006, Krishna Palaniappan, November 20, 2006)

Combustion Modeling

1. Flame Extinction

Numerical simulations of premixed turbulent flames give rise to a variety of analysis problems. One such problem is to understand the relationships between regions of local flame extinction and the variables describing the dynamically changing chemical compositions associated with these extinction regions. Given a flame simulation exhibiting extinction pockets, a time sequence of fuel consumption can be constructed by retracing the space-time trajectories (streamlines of the instantaneous velocity field) passing through both flame extinction regions and strongly burning regions. These multivariate time series may then be analyzed to predict or quantify the formation of extinction pockets.

The difficulties of this problem include several data management issues: 1) detecting regions of interest in large, distributed data sets, i.e., hundreds of instantaneous snapshots, each 20-80 GB, each grid point containing dozens of 3D scalar or vector field components (fluid velocity, molecular species concentrations, temperature, density, etc.), and 2) recomputing variables along the trajectories that were not saved from the initial simulation.