Diss. ETH No. 20918

Information Theoretic Modeling of Dynamical Systems: Estimation and Experimental Design

A dissertation submitted to
ETH Zurich

for the degree of
Doctor of Sciences

presented by
Alberto Giovanni Busetto
Dott. Mag. in Ingegneria Informatica,
Università degli Studi di Padova
born November 26, 1983 *in* Venice
citizen of the Italian Republic

accepted on the recommendation of
Prof. Dr. Joachim M. Buhmann, examiner
Prof. Dr. Manfred Morari, co-examiner
Prof. Dr. Jörg Stelling, co-examiner

2012

Abstract

Dynamical systems are mathematical models expressing cause-effect relations of time-varying phenomena. This thesis focuses on learning dynamical systems from empirical observations. Three settings are considered: unsupervised, supervised, and active learning. The unifying goal is to extract predictive information from data.

A method is introduced to cluster time-series and perform model validation. The method addresses order and model selection with the principle of Approximation Set Coding [Buhmann (2010)]. Experimental verification is performed in the context of relational clustering of temporal gene expression profiles. The results demonstrate wide applicability and consistency with the Bayesian Information Criterion. Then, discrete dynamic transitions are reconstructed from high-dimensional time-series with an unsupervised approach. The approach, based on Hidden Markov Models over Gaussian Mixtures, is applied to predict cell morphology classes from time-resolved microscopy data. Experimental validation with fluorescent markers and screening data demonstrates accurate identification of human cell phenotypes. Reported results highlight competitiveness and increased objectivity in comparison to supervised approaches based on user labeling.

In the supervised setting, clustering is employed to improve conventional particle filtering for generalized state estimation. Preventive clustering mitigates the inevitable divergence of resampling for sequential Monte Carlo methods. Supervised learning with dynamic

Bayesian Networks is employed to model human learning for effective treatment of learning disabilities. In dyslexia, the model predicts forgetting, focus and receptive states of the subject on the basis of input behavior. In dyscalculia, numerical cognition is enhanced through model-based adaptive training.

In the context of active learning, the thesis focuses on the near-optimal design of experiments for dynamical system modeling. An efficient method is introduced to select informative time points and measurable quantities. The design method is guaranteed to yield near-optimal informativeness with a polynomial number of evaluations of the objective function. The method builds on previous work on submodular active learning [Krause and Guestrin (2005)] and achieves the best possible constant approximation factor, unless P=NP [Feige (1998)]. Experimental design is applied to the reconstruction of cell signaling networks in systems biology.

The introduced contributions highlight fundamental analogies between learning and communication. In conclusion, the results demonstrate that predictive models can be built from efficient strategies of information transmission over a noisy channel. On the basis of statistical arguments, the presented results formalize and automate aspects of the hypothetico-deductive method of scientific inquiry.

Sommario

I sistemi dinamici sono modelli matematici che esprimono relazioni di causa-effetto riguardanti fenomeni soggetti a variazione temporale. Questa tesi si concentra sulla stima di sistemi dinamici sulla base di osservazioni empiriche. Si considerano tre scenari: stima senza supervisione, con supervisione e attiva. L'obiettivo unificante è l'estrazione di informazione predittiva dai dati.

Viene introdotto un metodo per il raggruppamento di serie temporali e la validazione statistica di modelli. Il metodo si propone di risolvere le questioni di selezione d'ordine e di modello con il principio della Codifica tramite Insiemi di Approssimazione [Buhmann (2010)]. La verifica sperimentale è perseguita nell'ambito di raggruppamento relazionale per profili temporali di espressione genetica. I risultati dimostrano vasta applicabilità e congruenza col Criterio d'Informazione Bayesiano. Inoltre, le transizioni dinamiche discrete sono ricostruite sulla base di serie temporali d'alta dimensionalità tramite un approccio privo di supervisione. L'approccio, che si basa su Modelli Markoviani Nascosti su Miscele Gaussiane, trova applicazione nella predizione di classi morfologiche da dati di microscopia che tengono conto del fattore tempo. La conferma sperimentale con marcatori fluorescenti e cernita dei campioni dimostra la capacità di accurata identificazione dei fenotipi cellulari umani. I risultati riportati evidenziano competitività e miglioramento dell'oggettività rispetto agli approcci supervisionati basati sulla catalogazione operata da parte dell'utente.

Nello scenario con supervisione, il processo di raggruppamento trova impiego nel perfezionamento del filtraggio convenzionale di campioni al fine di ottenere stime di stato generalizzate. Il raggruppamento preventivo mitiga l'inevitabile divergenza del ricampionamento impiegato nei metodi sequenziali di tipo Monte Carlo. La stima supervisionata con reti Bayesiane dinamiche è utilizzata nella modellistica dell'apprendimento umano con il fine di curare efficacemente alcune disabilità dell'apprendimento. Per la dislessia, il modello è in grado di predire rate di dimenticanza, livelli di concentrazione e inclinazione all'apprendimento del soggetto sulla base del comportamento misurato. Per la discalculia, la cognizione numerica è migliorata per mezzo di una terapia adattabile ottenuta sulla base del modello.

Nel contesto della stima attiva, la tesi si concentra sulla progettazione quasi ottima di esperimenti finalizzati alla modellistica di sistemi dinamici. Viene introdotto un metodo efficiente per selezionare quantità misurabili e istanti temporali particolarmente informativi. Si garantisce la quasi ottima informatività del metodo di progettazione, il quale richiede un numero polinomiale di valutazioni della funzione obiettivo. Il metodo si basa su lavoro precedente nell'ambito della stima attiva sotto-modulare [Krause and Guestrin (2005)] e ottiene il migliore fattore costante di approssimazione possibile, a meno che non valga P=NP [Feige (1998)]. La progettazione di esperimenti trova applicazione nella ricostruzione di reti cellulari segnalatrici nell'ambito della biologia dei sistemi.

I contributi presentati sottolineano le analogie fondamentali tra stima e comunicazione. In conclusione, i risultati dimostrano che modelli predittivi possono essere costruiti sulla base di efficienti strategie di trasmissione d'informazione tramite un canale rumoroso. Sulla base di argomentazioni di natura statistica, i risultati esibiti formalizzano e automatizzano aspetti del metodo ipotetico-deduttivo di investigazione scientifica.