Research Issues in Spatio-Temporal Learning

Thomas G. Dietterich
Department of Computer Science
College of Engineering
Oregon State University
Corvallis, Oregon, USA

tgd@cs.orst.edu
http://www.cs.orst.edu/~tgd

Funded by NSF grant IIS-0083292

The Problem

Given: Training examples of the form $\langle \mathbf{x}, \mathbf{y} \rangle$,
where both $\mathbf{x}$ and $\mathbf{y}$ are complex objects:
time-series, spatial (2- and 3D data), or even time-series of spatial data.

A loss function $L(\hat{\mathbf{y}}, \mathbf{y})$ defining the loss received when
the true output is $\mathbf{y}$ and the guessed output is $\hat{\mathbf{y}}$

Find: A function f: $\mathbf{y} = f(\mathbf{x})$ that minimizes expected loss on new points $\langle \mathbf{x}, \mathbf{y} \rangle$.

This is the standard supervised learning problem, except that the "class" $\mathbf{y}$ is a complex object.

The Solution: Some Form of Divide-and-Conquer

- **Divide**:

  Decompose $\mathbf{x}$ and $\mathbf{y}$ into small *windows*, $i = 1, \ldots, n$.

  Construct a feature vector $\mathbf{x}_i$ and a target class y_i for each window.

  Decompose the loss function $L(\hat{\mathbf{y}}, \mathbf{y})$ into a *window loss function* $\ell(\hat{y}_i, y_i)$.

- **Conquer**:

  Apply an ordinary (or perhaps, a cost-sensitive) learning algorithm to learn a *window mapping* $\hat{y}_i = f_i(\mathbf{x}_i)$ or a *window relation* $R(\mathbf{x}_i, \hat{y}_i)$.

- **Merge**:

  Apply some *merge method* to combine the $\hat{y}_i$'s for each window into a $\hat{\mathbf{y}}$ for the entire original object $\mathbf{x}$.

Example: NETTalk (Sejnowski & Rosenberg, 1987)

- **Input x:**

  English word: “`enough`”

- **Output y:**

  String of phonemes: “`In^-f-`”

  String of stresses: “`0>1<<<`”

- **Loss function:**

  If any phoneme or stress is incorrect, the entire word is considered incorrect.

NETtalk (2)

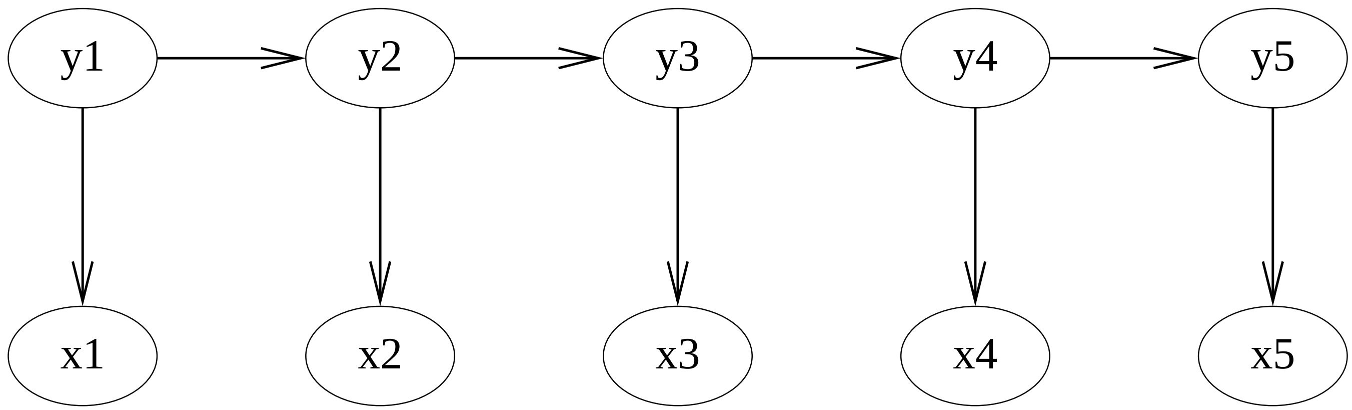
- **Divide**: $\mathbf{x}_i$: 7-letter window centered on a single letter; y_i: pair consisting of the phoneme and stress:

$\mathbf{x}_i$	y_i
`---enou`	(`I`, `0`)
`--enoug`	(`n`, `>`)
`-enough`	(`^`, `1`)
`enough-`	(`-`, `<`)
`nough--`	(`f`, `<`)
`ough---`	(`-`, `<`)

  Window loss function: simple 0-1 loss.

- **Conquer**: Backpropagation network with 80 hidden units.
- **Merge**: Simple concatenation of predicted phonemes and stresses.

Example: HMMs for Speech



The HMM models relationships between y_i and y_{i+1} and also relationships between $\mathbf{x}_i$ and y_i

- **Temporal Relationships:** $P(y_{i+1}|y_i)$
- **Input-Output Relationships:** $P(\mathbf{x}_i|y_i)$

HMMs viewed as Divide-and-Conquer

- **Divide**. Input and output scale are 1.
- **Conquer**. There are *two* learning problems:
  - **Learn** $\mathbf{P}(\mathbf{x}_i|y_i)$ This is the standard conquer step.
  - **Learn** $\mathbf{P}(y_{i+1}|y_i)$ This is the temporal learning problem.
- **Merge**. The Viterbi algorithm finds the most likely $\hat{\mathbf{y}}$ given the observed $\mathbf{x}$.

Design Decisions

- **Output Scale:** Size of the y_i's.
  What is the size of a good predictable unit? Many temporal and spatial problems exhibit regularities at multiple scales.
- **Input Scale:** Size of the $\mathbf{x}_i$'s.
  How much context information is needed to predict y_i?
- **Alignment:** In some applications (e.g., speech), the y_i values are not pre-aligned with the $\mathbf{x}_i$'s.
- **Merge Method:** How can the $\hat{y}_i$'s be combined to obtain a predicted output $\hat{\mathbf{y}}$? The merging process should exploit temporal and spatial correlations.
- **Decomposition of Loss Function:** How can we decompose the global loss function into a local loss function such that minimizing local loss also minimizes global loss?

Research Issues

- **Temporal and Spatial Correlations**

  Accurate predictions require capturing the temporal and spatial correlations rather than ignoring them.

- **Long-Distance Effects**

  Because of temporal or spatial correlations, information from relatively-distant x values may be needed to predict y.

- **Bias-Variance Tradeoff**

  Good learning methods automatically adjust hypothesis complexity to the amount and complexity of the training data to avoid under- or over-fitting.

- **Ease of Deployment**

  We want methods that are easy for non-researchers to deploy.

State of the Art and Practice I: Sliding Window Methods

- **Ignore the spatial or temporal correlations and pretend the $\langle \mathbf{x}_i, y_i \rangle$ pairs are iid.** Typically requires larger input window to compensate for failure to model spatial or temporal correlations among the y_i's. This in turn gives high variance and requires ad hoc methods to deal with multiple detections and global loss functions.

- **Long-Distance Effects.** Sliding window methods can be implemented with discriminative learning algorithms (SVMs, neural networks, decision trees, etc.). These make it easy to employ large input windows to capture long-distance interactions.

- **Bias/Variance Tradeoff.** Discriminative methods handle the bias/variance tradeoff well.

- **Ease of Deployment.** These methods are relatively easy to deploy, although there are no generic packages available.

State of the Art and Practice II: HMM and MRF methods

- **Explicitly model spatial or temporal correlations.** However, models are usually not correct.
- **Long-Distance Effects**. Can be difficult to incorporate. A long-distance interaction must be "remembered" by the hidden state.
- **Bias/Variance Tradeoff**. These probabilistic models typically are incorrect, and hence, have high bias. Few automated tools are available to manage the bias/variance tradeoff. This is typically handled manually by tied distributions and design of the hidden finite-state machine.
- **Difficult to Deploy**. The hidden finite-state machine must usually be carefully designed. This can require extensive experimentation.

Open Questions

- **Can we find a sound way of applying discriminative methods in spatio-temporal learning problems?**

  How can we capture spatio-temporal correlations explicitly? No sound merge method is known.

- **Can new classes of probabilistic models be developed that are easier to deploy, that can handle long-distance effects well, and can correctly tune themselves to the problem complexity?**

- **Can methods be developed that can incorporate global loss functions in a sound way?**

- **Are there well-founded ways of addressing the alignment problem?**

Morning Schedule

9:00- 10:00	**15-minute talks on spatio-temporal applications**
	Foster Provost: Event Monitoring for Fraud
	Attilio Giordana: Mining of FTP Logs, Commercial Transactions
	Simon Perkins: Spatio-spectral Pixel Classification in Satellite Images
	Rene Quiniou: EKG Monitoring and Alarming in ICU's
10:00-10:15	**COFFEE BREAK**
10:15-11:15	**15-minute talks on spatio-temporal applications**
	Alan Fern: Learning Visual Event Definitions from Video
	Mehmet Kayaal/Greg Cooper: Predicting Survival Outcomes from ICU Time Series
	Ronny Kohavi: Temporal Problems in E-Commerce Analysis
	Cesare Furlanello & S. Merler: Finding Unexploded Bombs from WWII
11:15-12:00	**Survey of HMM methods**, Padhraic Smyth, UC Irvine
12:00-13:30	**Lunch**

Afternoon Schedule

13:45-14:30	**Spatial Methods in Image Analysis**, Mike Turmon, JPL
14:30-15:15	**Graph Transformer Networks and OCR**, Leon Bottou, AT&T Research
15:30-15:45	**COFFEE BREAK**
15:45-16:30	**Exponential Models for Sequential Data**, John Lafferty, CMU
16:30-17:15	**Rare Event Modeling and Validation Through Time: The case of corporate credit analysis**, Roger Stein, Moody's
17:15-17:45	**Final Panel: Assembling a Research Agenda**, Smyth, Provost, Dietterich