Two Major Themes in The Design of Sensor Networks: data integrity and sampling

Nabil Hajj Chehade, Sheela Nair, Andrew Parker, Mark Hansen, Greg Pottie

Introduction

- Sensor networks are useful for learning about natural phenomena.
- Learning consists of extracting information from data to understand the underlying phenomenon.
- Two major concerns: data integrity and sampling.
- The widespread of use of sensor networks is limited by the poor quality of sensor data, which are often compromised by various faults.

- Adaptive sampling algorithms are more efficient than passive sampling algorithms. They are derived by optimizing a certain criterion:
  - For mobile sensors, sample paths can be selected to decrease a model's uncertainty, in the shortest amount of time.
  - Jointly optimize model selection and sampling strategy

Data Integrity

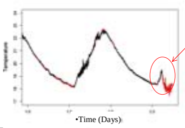
Signature-based fault detection

Determine major types of sensor faults, such as bias, drift, noise, clipping, stuck-at, outliers, etc. (following sensor fault taxonomy in Ni et. al)

Identify features that are effective in detecting these faults by examining their performance under various scenarios (including temporal and spatial structures)

Combine the effective features into "signatures" to model the common sensor faults

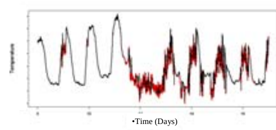
Signature-based fault detection evaluated on real data from three deployments
- Injected faults into clean sensor data and real faults flagged visually
- Algorithm able to detect real faults consistently and quickly while maintaining low (~6%) false alarm rate



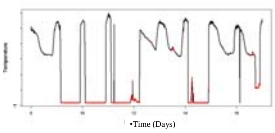
Battery failure causes drift fault in node 112

Error rate	D_{Te}	$D_j(112)$ $j = near$	far	$e_j(112)$ $near$	far
type I	0.046	0.055	0.066	0.056	0.065
type II	0.936	0.096	0.101	0.069	0.170

Noise faults



Stuck-At faults



On-line estimation and fault detection

On-line updating of estimates in the presence of fault detection algorithms often lead to biased estimates

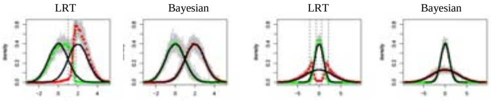
The problem is more difficult when the null distributions are themselves changing over time.

For parametric cases, the biases can be addressed in two ways: a) analytically compute the bias and adjust; or b) treat the flagged observation as missing and impute it from the conditional distribution.

For nonparametric estimation, the Bayesian classification rule does well in addressing the problem.

Frequency with which parameters estimates are updated is important. If updates are done too quickly, faults may be "learned" before they are detected

Performance of decision rules for on-line nonparametric density estimation

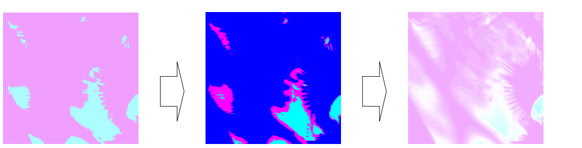


Sampling and Estimation of Latent Geometric Structures

Many Environmental Phenomena have Latent Geometric Structure

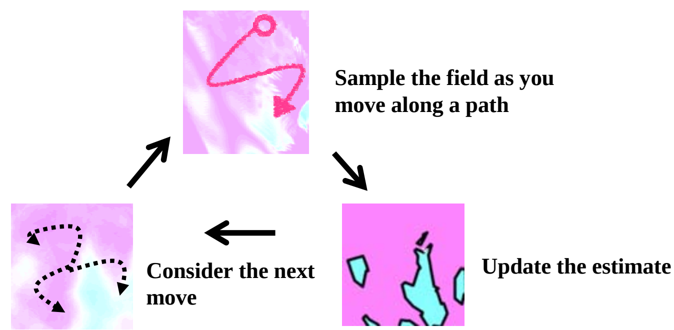
- **Sunflecks in the Forest**
- **Soil surface temperature**
- **Marine coastal nutrient and contaminant concentration**

The Geometric Structure is often part of a Hierarchical Model: A Sunfleck Model Shown Below



Geometric Shadow — Penumbral Fringe — Smooth Region Variation

Polygonal Random Fields ⇔ Continuous Path Samples



Adaptive Sampling for Model Selection

Algorithm

- Adaptive sampling algorithm to distinguish between n models.
- Idea: Find the points that result in the minimum probability of error.

- Example: A set of 2 regression models

$$H_1 : y_i = h_1(x_i, a) + e_i, \qquad i = 1,\dots,n$$
$$H_2 : y_i = h_2(x_i, b) + e_i, \qquad i = 1,\dots,n$$

1.Given a design w_N, where N is the number of observations, find :

$$\hat{a}_N = \arg\min_a \sum_{i=1}^{N} (y_i - h_1(x_i, a))^2$$
$$\hat{b}_N = \arg\min_b \sum_{i=1}^{N} (y_i - h_2(x_i, b))^2$$

2. Add to the design a point x_{N+1} such that :

$$x_{j+1} = \arg\max_{x \in Z} \left(h_1(x;\hat{a}) - h_2(x;\hat{b})\right)^2 - c(x - x_j)^2$$

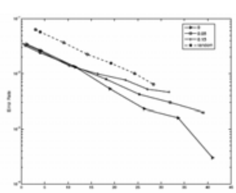
3.The $(N+1)th$ observation is taken at x_{N+1}

Update w : $w_{N+1} = (1-\alpha) * w_N + \alpha * \delta(x_{N+1})$

4.Go back to 1

Performance

$$h_1(x_i; a) = a_0 + a_1 x_i \qquad i = 1,\dots,m$$
$$h_2(x_i; b) = b_0 + b_1 e^{x_i} + b_2 e^{-x_i} \qquad i = 1,\dots,m$$

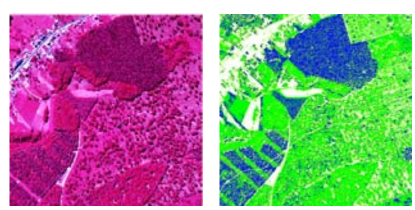


- Static sensors are used to find an initial estimate of the models.
- Adaptive sampling using a mobile sensor: Joint minimization of the probability of error and the mobility cost.
- This can be easily extended to multiple hypotheses

Preliminary Work

(Joint work with the ARIANA group at INRIA Sophia-Antipolis. And IFN)

- Non-parametric models: SVM using Gaussian kernel.
- Example: Tree type classification
- Training set selection is crucial. How can we design optimal training set selection algorithms?



© IFN