From Data to Knowledge

Michael C. Ferris
(joint work with Olvi Mangasarian, Edward Wild)

University of Wisconsin, Madison
NSF Grants IIS-0511905, DMI-0521953 & DMS-0427689

November 3, 2007

Knowledge is Summary of Data

Points, records, images			data
Zip file	compression	$\Longleftrightarrow$	knowledge
Visualization	summary	$\Longleftrightarrow$	knowledge
Progressive meshes/JPEG	inference	$\Longleftrightarrow$	knowledge
Animations	level of detail	$\Longleftrightarrow$	knowledge
(foreground vs background)			
Robot football	state	$\Longleftrightarrow$	data
(field, location of players)			
Where to pass	implications, action	$\Longleftrightarrow$	knowledge

Knowledge is often formulated as an implication

Medical Applications

Symptoms given to doctor [partial information]

What is ailment?	determine state	$\Longleftrightarrow$	knowledge
What is treatment?	action	$\Longleftrightarrow$	knowledge
What is prognosis?	prediction	$\Longleftrightarrow$	knowledge

Medical Applications

Symptoms given to doctor [partial information]

What is ailment?	determine state	$\iff$	knowledge
What is treatment?	action	$\iff$	knowledge
What is prognosis?	prediction	$\iff$	knowledge

- Clinical images (of different biological response) [data conflict]
- Uncertainty of tumor and organ location [data uncertainty]
- Required treatment plan: adaptive (feedback) $\iff$ knowledge

Medical Applications

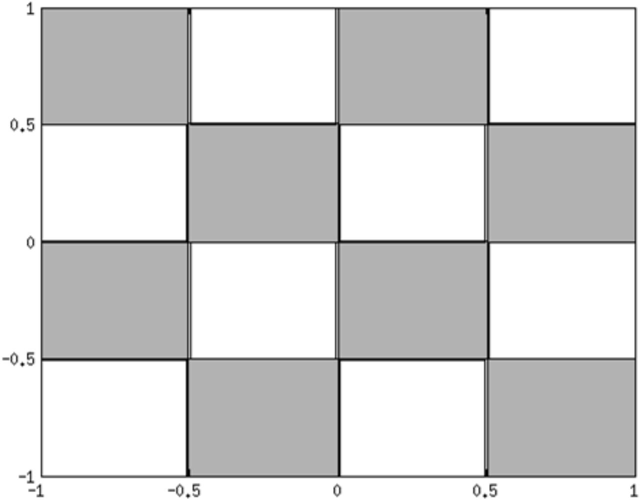
Symptoms given to doctor [partial information]

What is ailment?	determine state	$\Longleftrightarrow$	knowledge
What is treatment?	action	$\Longleftrightarrow$	knowledge
What is prognosis?	prediction	$\Longleftrightarrow$	knowledge

- Clinical images (of different biological response) [data conflict]
- Uncertainty of tumor and organ location [data uncertainty]
- Required treatment plan: adaptive (feedback) $\Longleftrightarrow$ knowledge

ITR funding facilitated domain expert interaction

Checkerboard Dataset Black and White Points in $\mathbf{R}^2$



Multiple representations of knowledge

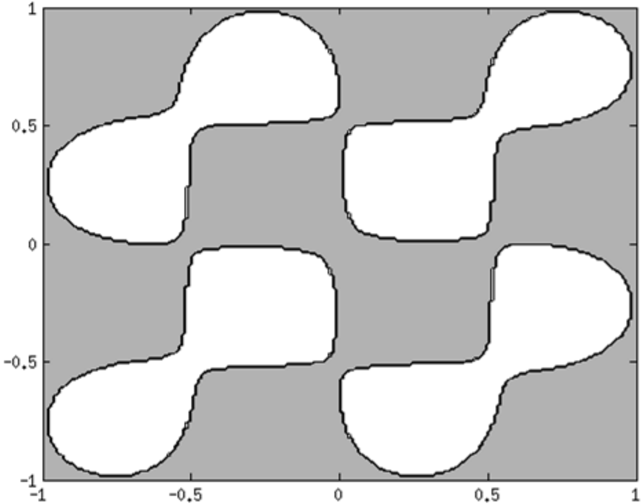
1. Bitmap image
2. Patches plus anchor points
3. Compression using replication
4. Nonlinear function "classifier" $f(x) = \sum_i K_i(x)u_i$

Basis pursuit, "sparse" L_1 (signal processing) models, reduced SVM

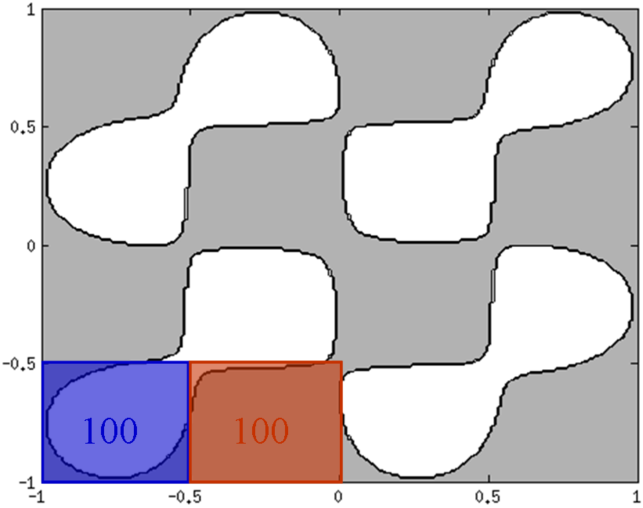
- Useful in different contexts
- Possibly problem domain specific
- Varying degrees of complexity

Models encode representations; Optimization is key to recovery

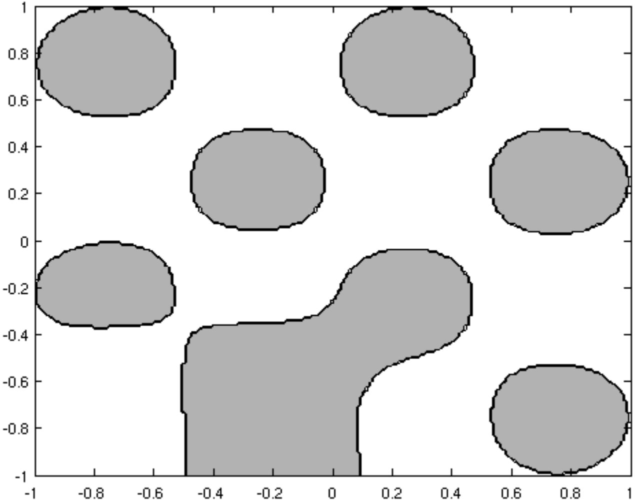
Checkerboard Classifier Using 16 Center Points



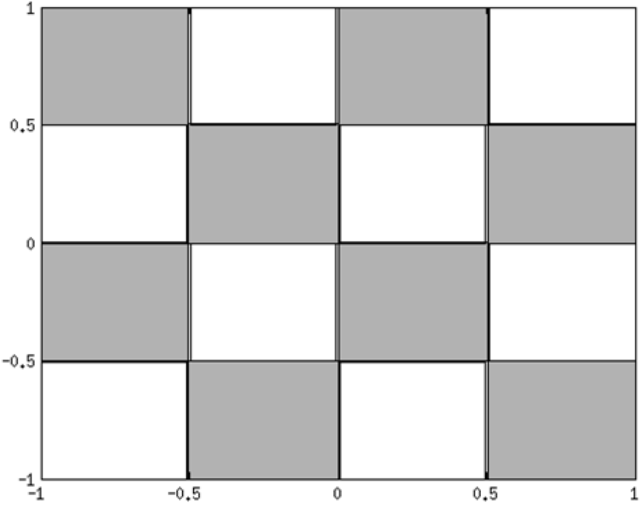
Prior Knowledge for 16-Point Checkerboard Classifier



Checkerboard Classifier with Knowledge as Extra Data



Checkerboard Classifier With Knowledge and 16 Points



Overview

Rigorous theory, good models, powerful computation

Use of knowledge

- Theoretical underpinnings
- Natural incorporation of knowledge
- Examples of application

Models based on “robust” and “generalized constraints”

- Use of powerful modeling tools coupled with cyber-infrastructure
- Training, and tuning via grid
- Large distributed datasets

Incorporating Knowledge

- Utilizing only given data may result in a poor classifier or approximation
  - Points may be noisy
  - Sampling may be costly
  - Overfitting/imbalanced data
- Use "summarization" of prior knowledge to improve the classifier or approximation
- Require "natural" imposition of knowledge understandable to humans

Rigorous Theory

Theorem (of the alternative, nonlinear)

Under appropriate assumptions, exactly one of the following two systems has a solution:

$$(I): \quad \exists x \in \Gamma : g(x) \leq 0, f(x) < \gamma(x)$$

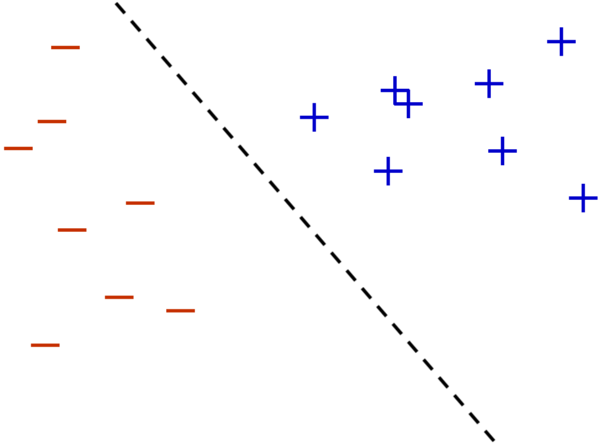
$$(II): \quad \exists v \geq 0 : v^T g(x) + f(x) - \gamma(x) \geq 0, \quad \forall x \in \Gamma$$

Thus, if (II) has solution then:

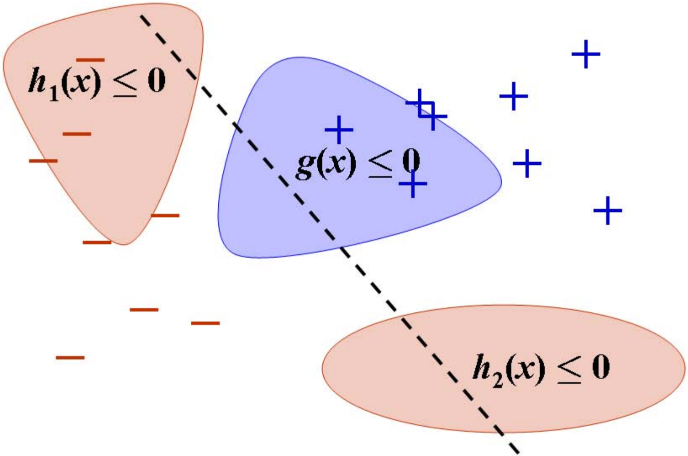
$$g(x) \leq 0 \implies f(x) \geq \gamma(x), \quad \forall x \in \Gamma$$

(Previous work involved "kernelizing" knowledge - effective but not transparent).

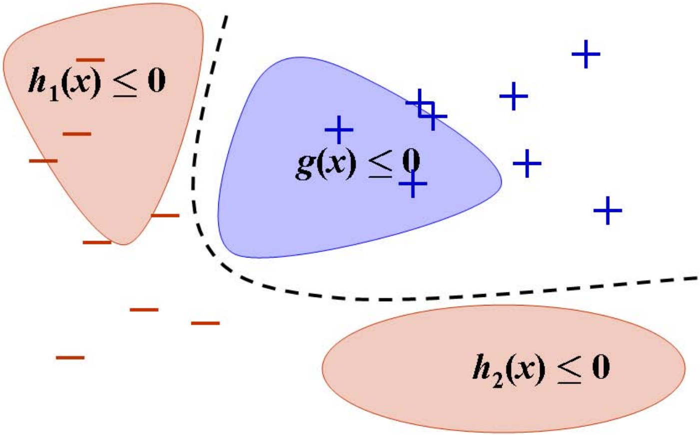
Knowledge Incorporation Example



Knowledge Incorporation Example



Knowledge Incorporation Example



Model Incorporating Knowledge

Linear semi-infinite program

$$
\begin{aligned}
\min_{(u,s,y\in Y,v\geq 0)} \quad & \|u\|_1 + \nu \|s\|_1 \\
\text{s.t.} \quad & -s_j \leq f(A_{j\cdot}) - y_j \leq s_j, \quad \forall j \\
& v^T g(x) + f(x) - \gamma(x) \geq 0, \quad \forall x \in \Gamma.
\end{aligned}
$$

Model Incorporating Knowledge

Linear semi-infinite program

$$\begin{array}{ll} \min\limits_{(u,s,y\in Y,v\geq 0)} & \|u\|_1 + \nu \|s\|_1 \\ \text{s.t.} & -s_j \leq f(A_{j\cdot}) - y_j \leq s_j, \;\; \forall j \\ & v^T g(x) + f(x) - \gamma(x) \geq 0, \;\; \forall x \in \Gamma. \end{array}$$

Discretize (to get LP) and add trade-off parameter

$$\begin{array}{ll} \min\limits_{(u,s,y\in Y,v\geq 0,z\geq 0)} & \|u\|_1 + \nu \|s\|_1 + \sigma \|z\|_1 \\ \text{s.t.} & -s_j \leq f(A_{j\cdot}) - y_j \leq s_j, \;\; \forall j \\ & v^T g(x^i) + f(x^i) - \gamma(x^i) + z_i \geq 0, \;\; \forall i. \end{array}$$

Global versus local

- Local $\equiv$ "training data"
- Global $\equiv$ "generalization"

Classification-Based Global Optimization

- Build a "classifier" to predict level sets of objective based on data
- Classifier is "cheap" to evaluate
- Use classifier to target evaluations in regions that are promising
- Component of WISOPT software
- Application to simulation calibration, e.g. Wisconsin Breast Cancer Epidemiology

Powerful Computation

- Global models incorporating (domain) knowledge (reflect more accurately our underlying knowledge of the system)
- Use of grid computation (eg. GAMS/grid, Condor, Sun Grid engine) to enhance global search
- Build a tractable "non-convex" model

$$\min_{i=1,\ldots,N} Q_i(x)$$

  to approximate a non-convex objective
- Use high performance computing to evaluate model minimizers in parallel
- Application to protein docking

Predicting Breast Cancer Recurrence Within 24 Months

Wisconsin Prognostic Breast Cancer (WPBC) dataset

- 155 patients monitored for recurrence within 24 months
- 30 cytological features
- 2 histological features: number of metastasized lymph nodes and tumor size

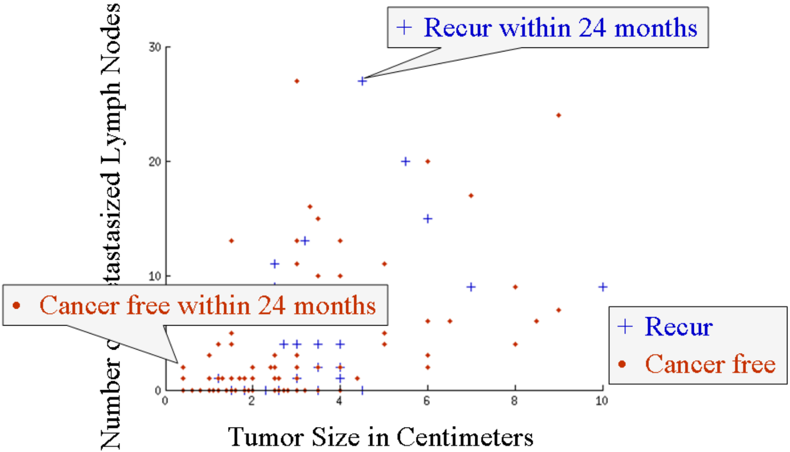
Predicting Breast Cancer Recurrence Within 24 Months

Wisconsin Prognostic Breast Cancer (WPBC) dataset

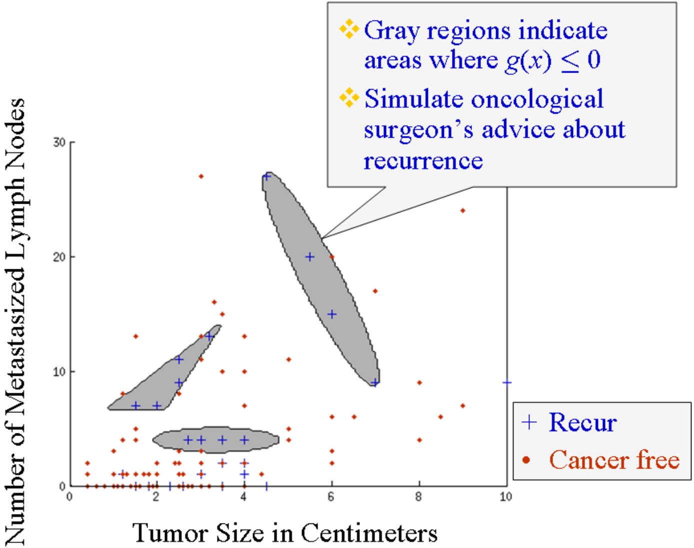
- 155 patients monitored for recurrence within 24 months
- 30 cytological features
- 2 histological features: number of metastasized lymph nodes and tumor size

- Predict whether or not a patient remains cancer free after 24 months
- 82% of patients remain disease free
- 86% accuracy (Bennett, 1992) best previously attained
- Prior knowledge allows us to incorporate additional information to improve accuracy

Generating WPBC Prior Knowledge



Generating WPBC Prior Knowledge



WPBC Results

Classifier	Misclassification Rate
Without Knowledge	18.1%
With Knowledge	9.0%

49.7% improvement due to knowledge

35.7% improvement over best previous predictor

Conclusion

Utilize nonlinear prior knowledge in classification/approximation

- Implemented as linear inequalities in a linear programming problem
- Knowledge appears transparently

Demonstrated effectiveness

- Real world problem from breast cancer prognosis

Future work

- Prior knowledge with more general implications
- Generate prior knowledge for real-world datasets
- See www.cs.wisc.edu/∼olvi/nsf04

Possible Funding Opportunities

- Privacy notions: infer a summary without identifying subjects
  - mental health
  - finance
  - health-care (features and individuals)
- Interdisciplinary (knowledge) incorporation
  - biological, imaging, ultrasound, transaction
- Imprecise, inaccurate, uncertain data
- Novel modeling paradigms
- Cyber-infrastructure resources applied in application domains