Supervised Sequential Classification Under Budget Constraints

Kirill Trapeznikov and Venkatesh Saligrama
Boston University

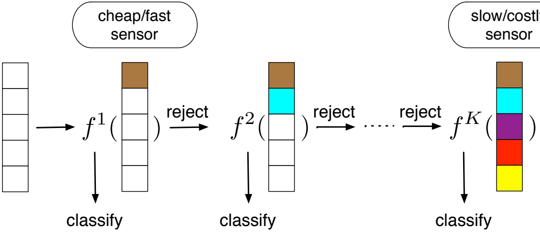
May 1st, 2013

Department of Electrical & Computer Engineering
The Center for Information and Systems Engineering
CISE

Overview

- Introduce sequential decision problem
- Myopic approach: relies on current uncertainty to make a decision
  - Consider synthetic examples
  - Why it does not always work
- Our approach: incorporate future uncertainty in current decision
  - Examine a two stage system
  - Reduce to supervised learning
- Experiment
- Extend to Multiple Stages
- Generalization Results

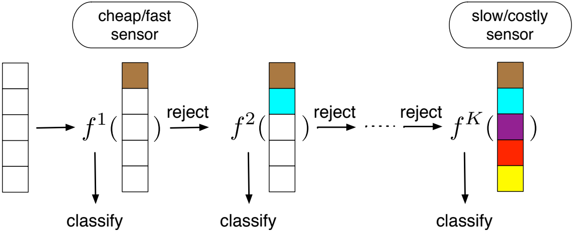
The Problem: Sequential Decision System



K **stage decision system**:

- Stage k can use sensor k for a cost c_k
- Measurements can be high dimensional
- Order of stages/sensors is fixed

The Problem: Sequential Decision System



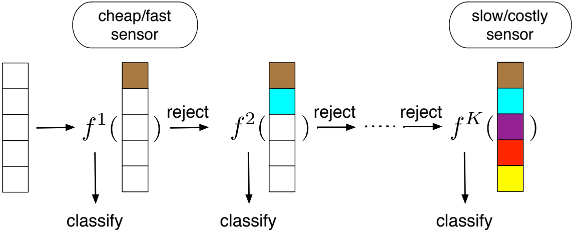
K **stage decision system**:

- Stage k can use sensor k for a cost c_k
- Measurements can be high dimensional
- Order of stages/sensors is fixed

Decision at each stage:

- classify using current measurements, or
- request (reject to) next sensor

The Problem: Sequential Decision System



K **stage decision system**:

- Stage k can use sensor k for a cost c_k
- Measurements can be high dimensional
- Order of stages/sensors is fixed

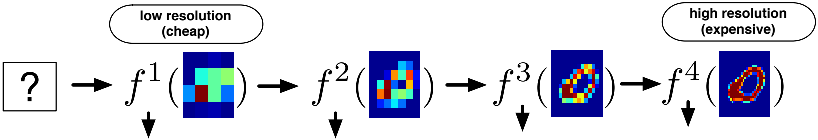
Decision at each stage:

- classify using current measurements, or
- request (reject to) next sensor

Goal: Find decisions: $F = \{f^1, f^2, \ldots, f^K\}$
trade-off error rate vs average acquisition cost

Sensors of Increasing Resolutions

classify handwritten digit images

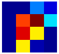


Do we need all sensors for every decision?

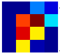
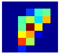
Difficult Decision

$$\boxed{?} \rightarrow f^1(\quad) \rightarrow f^2(\quad) \rightarrow f^3(\quad) \rightarrow f^4(\quad)$$

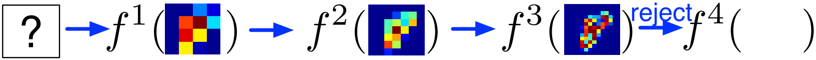
Difficult Decision

$\boxed{?} \rightarrow f^1($  $) \xrightarrow{\text{reject}} f^2(\qquad) \rightarrow f^3(\qquad) \rightarrow f^4(\qquad)$

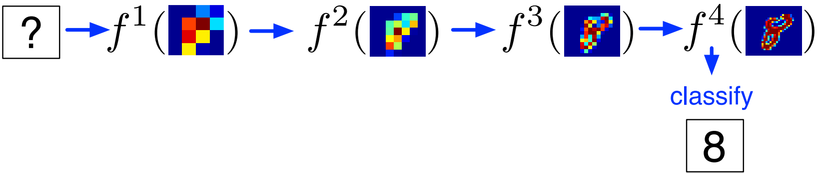
Difficult Decision

$\boxed{?} \rightarrow f^1($  $) \rightarrow f^2($  $) \xrightarrow{\text{reject}} f^3(\quad) \rightarrow f^4(\quad)$

Difficult Decision



Difficult Decision

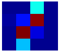


high acquisition cost: need full resolution to make a decision

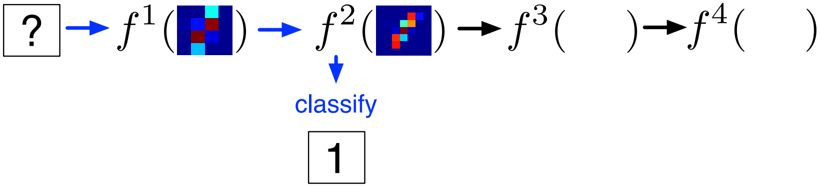
Easy Decision

$$\boxed{?} \rightarrow f^1(\quad) \rightarrow f^2(\quad) \rightarrow f^3(\quad) \rightarrow f^4(\quad)$$

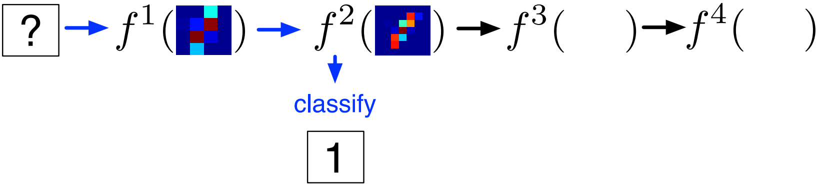
Easy Decision

$\boxed{?} \rightarrow f^1($  $) \xrightarrow{\text{reject}} f^2(\quad) \rightarrow f^3(\quad) \rightarrow f^4(\quad)$

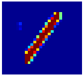
Easy Decision



Easy Decision

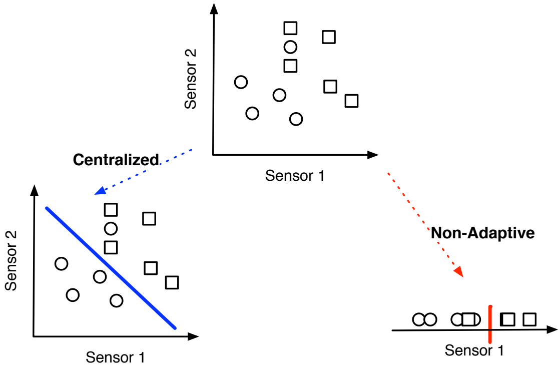


small acquisition cost: full resolution is unnecessary



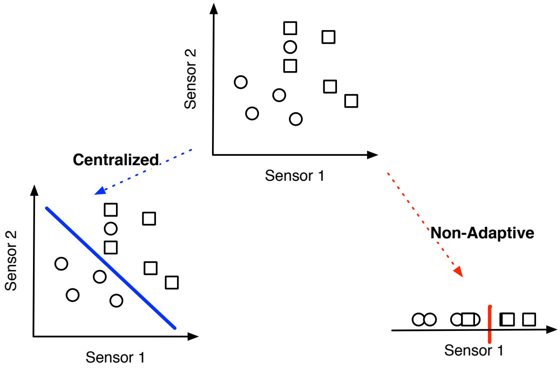
How to reduce sensor cost?

Sensor 1 is cheap, Sensor 2 is expensive



How to reduce sensor cost?

Sensor 1 is cheap, Sensor 2 is expensive

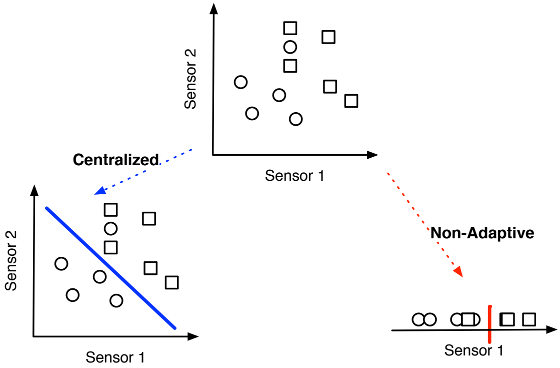


Centralized strategy:

- use both sensors
- high cost, low error

How to reduce sensor cost?

Sensor 1 is cheap, Sensor 2 is expensive



Centralized strategy:

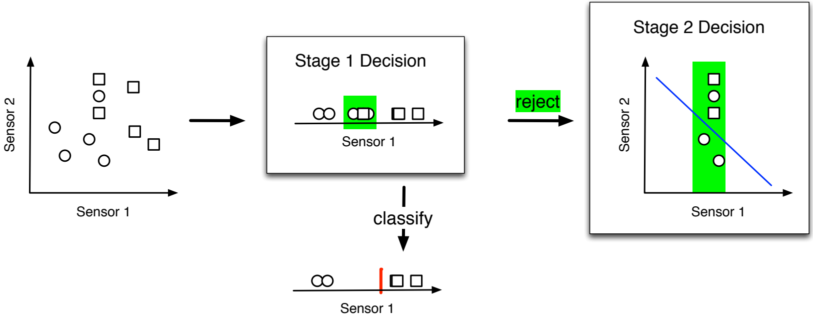
- use both sensors
- high cost, low error

Non-adaptive strategy:

- only use sensor 1
- low cost, high error

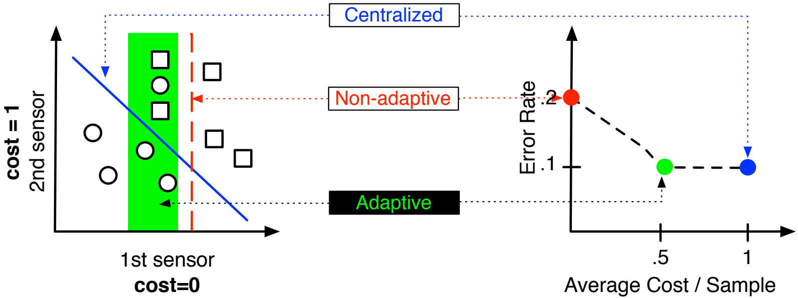
A better strategy: be adaptive

Only request 2nd sensor on difficult examples



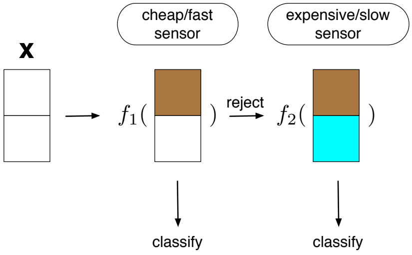
How does it compare?

Same error rate as centralized for half the cost



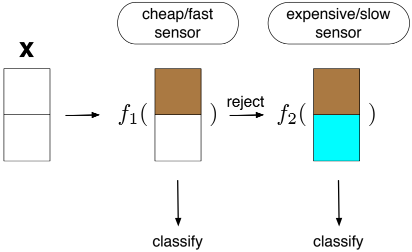
Deciding to reject

How to decide if to use the next sensor?



Deciding to reject

How to decide if to use the next sensor?



Risk of a decision:

$$\min\ \Big[\ \underbrace{\text{current uncertainty}}_{\text{classify}},\ \ \underbrace{\alpha \times \text{cost} + \text{future uncertainty}}_{\text{reject to next stage}}\ \Big]$$

(uncertainty is in correct classification)

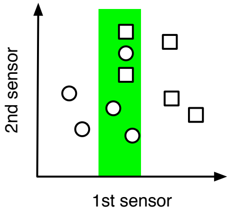
Acquisition cost justify the reduction in uncertainty?

Deciding to reject

$$\text{Risk} = \min \left[\underbrace{\text{current uncertainty}}_{\text{classify}}, \underbrace{\alpha \times \text{cost} + \text{future uncertainty}}_{\text{reject to next stage}} \right]$$

Difficulty: sensor output is not known since it has not been acquired

- How to determine future uncertainty?
- **Must base decision on collected measurements!**



Myopic Approach

Not clear how to determine uncertainty of the future:

$$\min \; [\; \underbrace{\text{current uncertainty},}_{\text{classify}} \;\; \underbrace{\alpha \times \text{cost} + \text{future uncertainty}}_{\text{reject to next stage}} \;]$$

Myopic Approach

Not clear how to determine uncertainty of the future:

$$\min \left[\underbrace{\text{current uncertainty}}_{\text{classify}}, \; \underbrace{\alpha \times \text{cost} + \text{future uncertainty}}_{\text{reject to next stage}} \right]$$

Ignore the future, and only use current uncertainty to make a decision:

$$\min \left[\underbrace{\text{current uncertainty}}_{\text{classify}}, \qquad \underbrace{\alpha \times \text{cost}}_{\text{reject to next stage}} \qquad \right]$$

Myopic Approach

Not clear how to determine uncertainty of the future:

$$\min \Big[\underbrace{\text{current uncertainty}}_{\text{classify}}, \ \underbrace{\alpha \times \text{cost} + \text{future uncertainty}}_{\text{reject to next stage}} \Big]$$

Ignore the future, and only use current uncertainty to make a decision:

$$\min \Big[\underbrace{\text{current uncertainty}}_{\text{classify}}, \ \underbrace{\alpha \times \text{cost}}_{\text{reject to next stage}} \Big]$$

Reduces to:

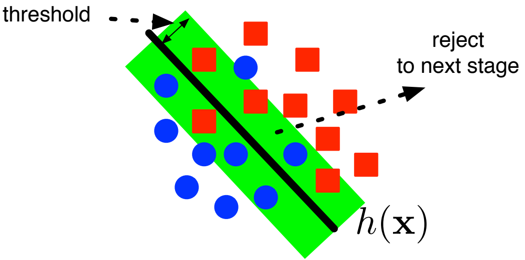
$$\text{decision} = \begin{cases} \text{classify}, & \text{uncertainly} < \text{threshold} \\ \text{reject}, & \text{uncertainty} \geq \text{threshold} \end{cases}$$

Myopic In Discriminative Setting

Train a classifier at a stage $h(\mathbf{x})$

Classifier uncertainty $\approx$ distance to decision boundary (margin)

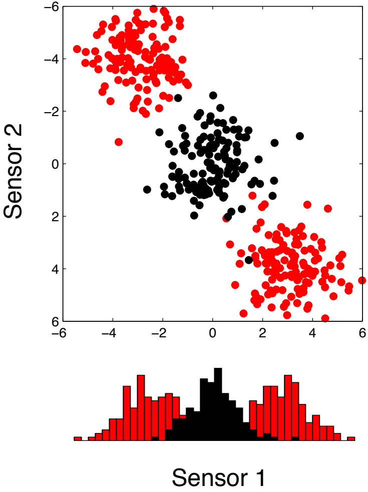
- Small distance $\rightarrow$ high uncertainty
- Large distance $\rightarrow$ low uncertainty



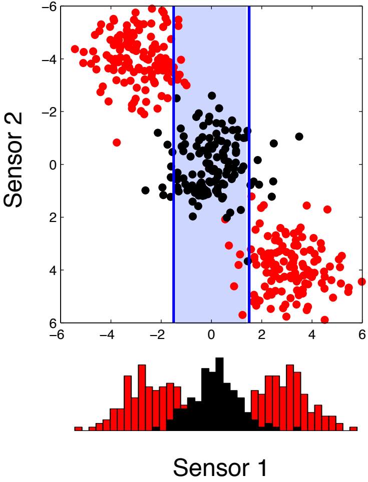
Related work: [Liu et al., 2008]

Example 1

Data:

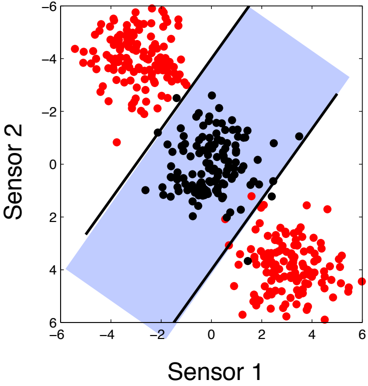


1st Stage Classifier: only utilizes Sensor 1

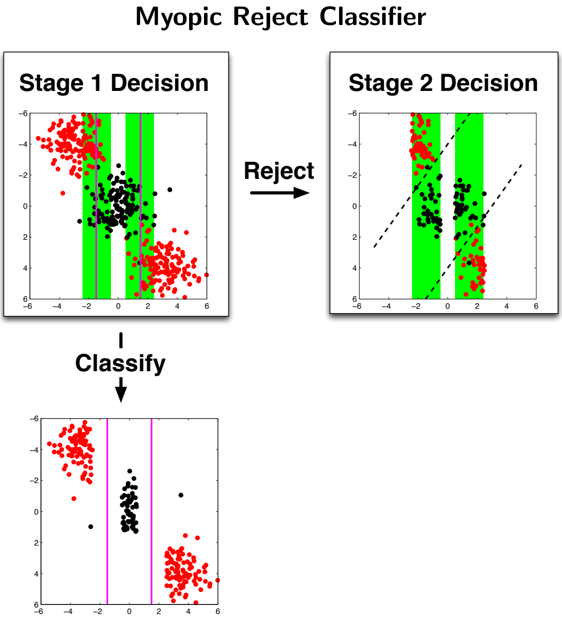


Example 1

2nd Stage Classifier: utilizes Sensors 1 and 2

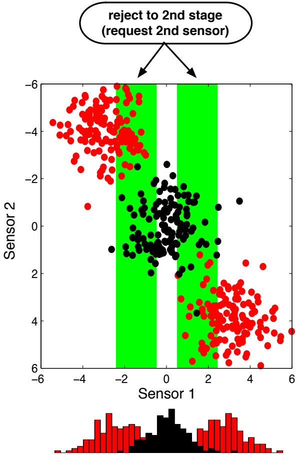


Example 1

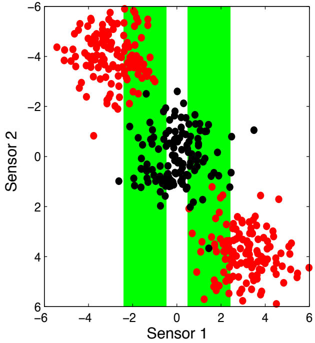

Myopic Reject Classifier
Stage 1 Decision
Stage 2 Decision
Reject
Classify
−6 −4 −2 0 2 4 6

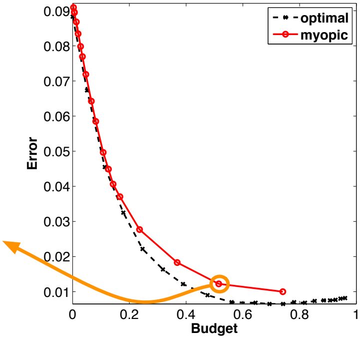
Myopic Reject Classifier

- Requests sensor 2 where sensor 1 is ambiguous
- Current uncertainty seems to be a good criteria to reject



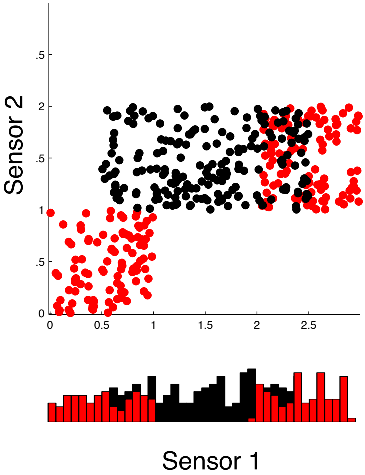
sweep threshold to generate different operating points



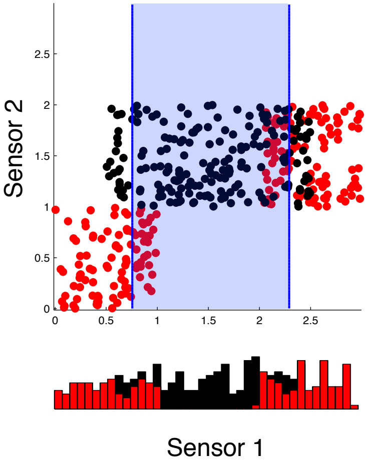


Good performance: close to optimal, seems to work

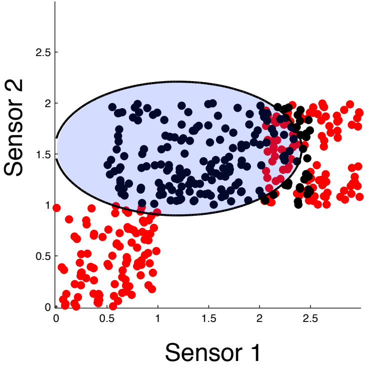
Example 2



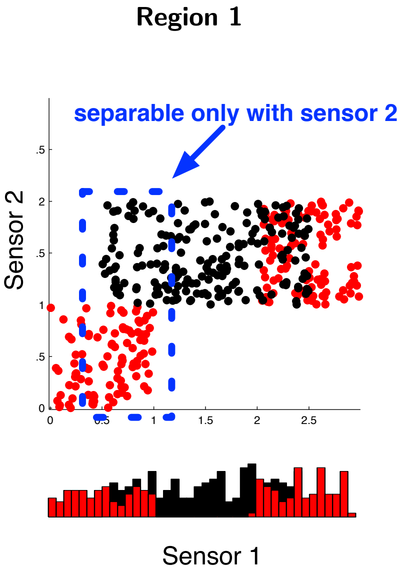
1st Stage Classifier: only utilizes Sensor 1



2nd Stage Classifier: utilizes Sensors 1 and 2

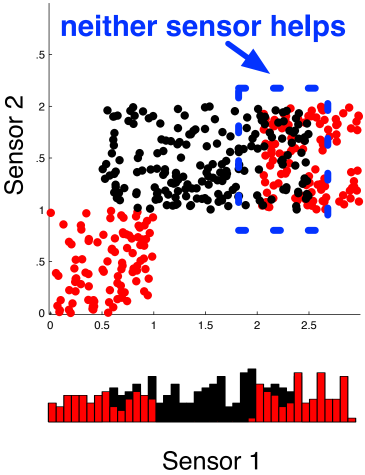


Example 2


Region 1
separable only with sensor 2
Sensor 2
Sensor 1
0
.5
1
1.5
2
2.5

Example 2

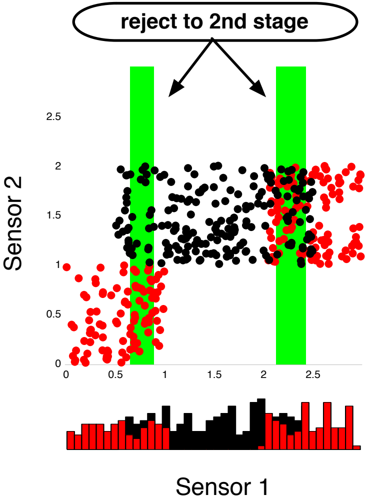
Region 2



Example 2

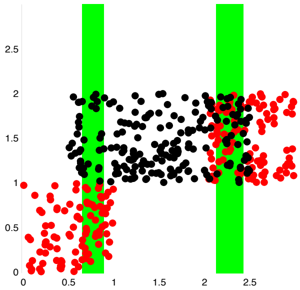
Myopic Reject Decision

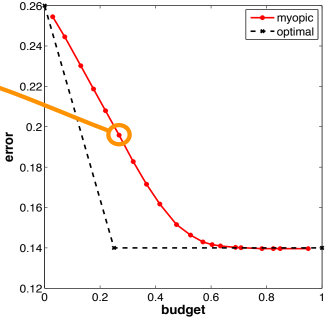
- Sensor 1 uncertainty is equally distributed between regions 1 and 2
- Uniformly rejects in both regions



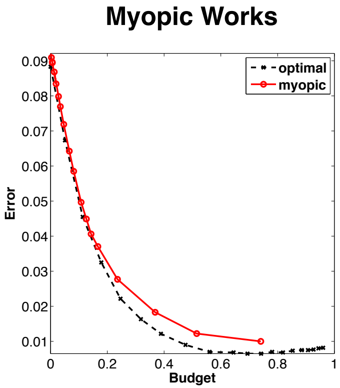
Myopic Reject Decision

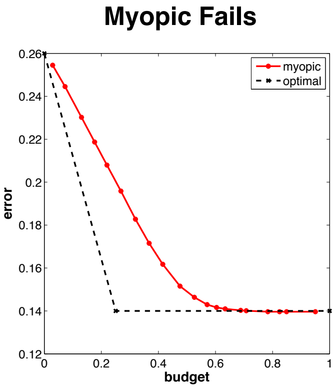
- Current uncertainty is equally distributed between regions 1 and 2
- Without future uncertainty cannot tell where sensor 2 is useful





Myopic


Myopic Works
optimal
myopic
Error
Budget
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0.2
0.4
0.6
0.8
1


Myopic Fails
myopic
optimal
error
budget
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0
0.2
0.4
0.6
0.8
1

Future Uncertainty is Important

Need to incorporate future uncertainty in the decision

$$\min \left[\underbrace{\text{current uncertainty}}_{\text{classify}},\ \underbrace{\alpha \times \text{cost} + \text{future uncertainty}}_{\text{reject to next stage}} \right]$$

Generative & Parametric Methods

Known model: partially observable Markov decision process (POMDP)

- Posterior Model: **P**(state | sensor measurements)
- Likelihood Model: **P**(sensor k | sensor j)

Method 1: Learn models and solve POMDP

- hard to learn models,
- cannot solve POMDP in general case

Previous Work:
[Ji and Carin, 2007, Kapoor and Horvitz, 2009, Zubek and Dietterich, 2002]

Method 2: Greedily maximize expected utility of a sensor

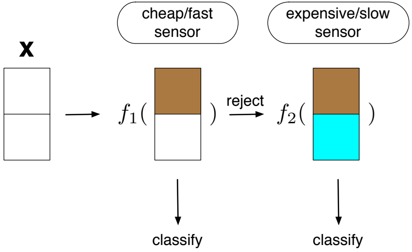
- One step look ahead approximation to POMDP, unclear how to choose utility
- Correlation across sensors: hard to learn likelihood (e.g. sensor output = image)

Previous Work: [Kanani and Melville, 2008, Koller and Gao, 2011]

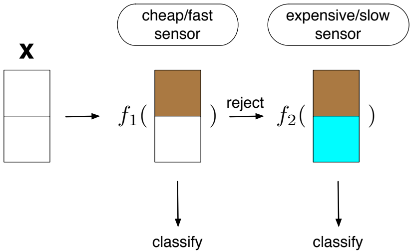
Our Approach

- Avoid estimating probability models
- Directly learn decision at each stage from training data
- Empirical Risk Minimization (ERM):
  incorporates uncertainty of future in the current decision

Two Stage System



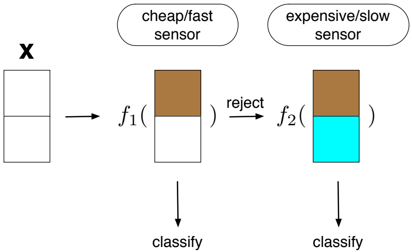
Stage Classifiers



Fix classifiers at each stage:

- $h_1(\mathbf{x})$ is standard classifier trained on sensor 1
- $h_2(\mathbf{x})$ is standard classifier trained on sensor 1 & 2

Decompose Reject Decision

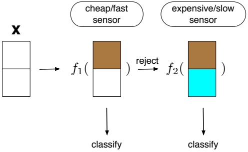


Decompose classification and rejection decisions:

- $g(\mathbf{x})$ is reject / not reject decision

$$f_1(\mathbf{x}) = \begin{cases} h_1(\mathbf{x}), & g(\mathbf{x}) = \text{not reject} \\ \text{reject}, & \text{else} \end{cases}$$

Risk Based Approach

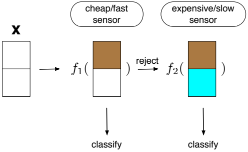


Risks of Each Stage:

$$\text{Current:} \quad R_{cu}(\mathbf{x}) = \mathbb{1}_{[h_1 \text{ misclassifies } \mathbf{x}]}$$

$$\text{Future:} \quad R_{fu}(\mathbf{x}) = \mathbb{1}_{[h_2 \text{ misclassifies } \mathbf{x}]} + \alpha \times \text{sensor 2 cost}$$

Risk Based Approach



Risks of Each Stage:

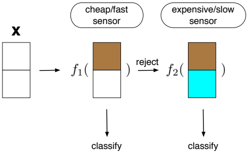
$$\text{Current:} \quad R_{cu}(\mathbf{x}) = \mathbb{1}_{[h_1 \text{ misclassifies } \mathbf{x}]}$$

$$\text{Future:} \quad R_{fu}(\mathbf{x}) = \mathbb{1}_{[h_2 \text{ misclassifies } \mathbf{x}]} + \alpha \times \text{sensor 2 cost}$$

Stage 1 reject decision $g(\mathbf{x})$:

$$g(\mathbf{x}) = \begin{cases} \text{classify at 1,} & R_{cu}(\mathbf{x}) < R_{fu}(\mathbf{x}) \\ \text{reject to 2nd sensor,} & R_{cu}(\mathbf{x}) \geq R_{fu}(\mathbf{x}) \end{cases}$$

Risk Based Approach



Risks of Each Stage:

$$\text{Current: } R_{cu}(\mathbf{x}) = \mathbb{1}_{[h_1 \text{ misclassifies } \mathbf{x}]}$$

$$\text{Future: } R_{fu}(\mathbf{x}) = \mathbb{1}_{[h_2 \text{ misclassifies } \mathbf{x}]} + \alpha \times \text{sensor 2 cost}$$

Stage 1 reject decision $g(\mathbf{x})$:

$$g(\mathbf{x}) = \begin{cases} \text{classify at 1}, & R_{cu}(\mathbf{x}) < R_{fu}(\mathbf{x}) \\ \text{reject to 2nd sensor}, & R_{cu}(\mathbf{x}) \geq R_{fu}(\mathbf{x}) \end{cases}$$

Difficulty: R_{cu}, R_{fu} require ground truth y and R_{fu} requires sensor 2

Empirical Risk Minimization

Use training data with full measurement,

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$$

Empirical Risk Minimization

Use training data with full measurement,

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$$

And system risk for a point $\mathbf{x}$ and decision $g(\mathbf{x})$

$$R(g, \mathbf{x}, y) = \begin{cases} R_{cu}(\mathbf{x}, y), & g(\mathbf{x}) = \text{not reject} \\ R_{fu}(\mathbf{x}, y), & g(\mathbf{x}) = \text{reject} \end{cases}$$

Empirical Risk Minimization

Use training data with full measurement,

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$$

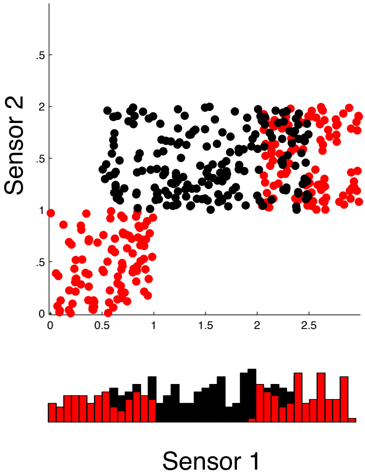
And system risk for a point $\mathbf{x}$ and decision $g(\mathbf{x})$

$$R(g, \mathbf{x}, y) = \begin{cases} R_{cu}(\mathbf{x}, y), & g(\mathbf{x}) = \text{not reject} \\ R_{fu}(\mathbf{x}, y), & g(\mathbf{x}) = \text{reject} \end{cases}$$

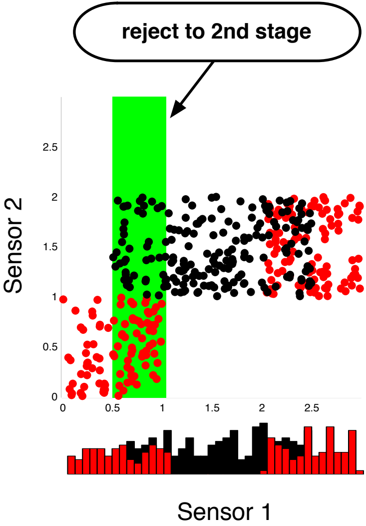
Minimize empirical risk,

$$\min_{g} \mathbf{E}_{\mathbf{x},y}[R(g, \mathbf{x}, y)] \approx \min_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} R(g, \mathbf{x}_i, y_i)$$

Back to Example 2

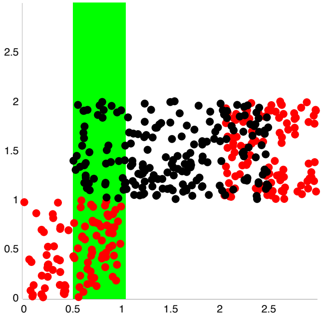


Example 2



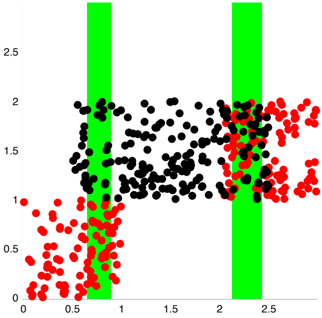
Smaller error for the same cost

Ours



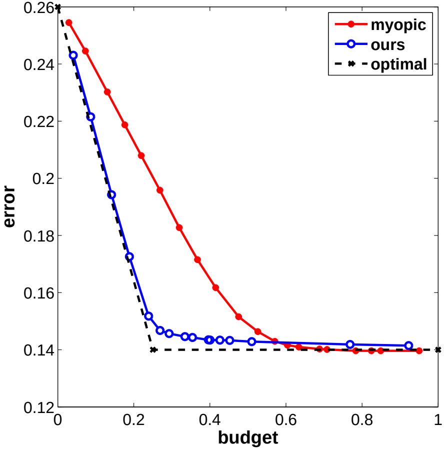
Error=14.8%

Myopic



Error=19%

Incorporating future uncertainty in current decision improves performance



Learning to Reject

How to learn reject decision $g(\mathbf{x})$ (green region)?

Reduce reject option to learning a binary decision

Define a weighted supervised learning problem:

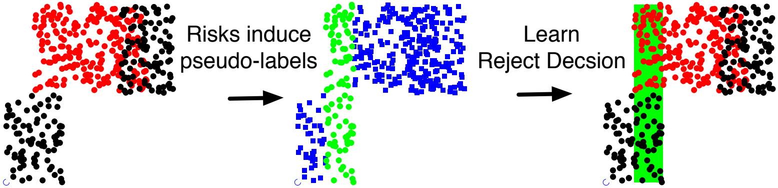
- risk difference induces pseudo labels on training data,

$$\text{pseudo label of } \mathbf{x}_i = \begin{cases} \textit{reject}, & R_{cu}(\mathbf{x}_i) > R_{fu}(\mathbf{x}_i) \\ \text{not reject}, & R_{cu}(\mathbf{x}_i) \leq R_{fu}(\mathbf{x}_i) \end{cases}$$

- importance weights, risk difference = penalty for misclassifying

$$\text{weight of } \mathbf{x}_i = |R_{cu}(\mathbf{x}_i) - R_{fu}(\mathbf{x}_i)|$$

Learning to Reject



$$\text{pseudo label of } \mathbf{x}_i = \begin{cases} \boxed{\textit{reject}}, & R_{cu}(\mathbf{x}_i) > R_{fu}(\mathbf{x}_i) \\ \text{not reject}, & R_{cu}(\mathbf{x}_i) \leq R_{fu}(\mathbf{x}_i) \end{cases}$$

$$\text{weight of } \mathbf{x}_i = |R_{cu}(\mathbf{x}_i) - R_{fu}(\mathbf{x}_i)|$$

Reduction to supervised learning

Theorem:
Empirical risk minimization simplifies to weighted supervised learning:

$$\arg\min_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^{N} R(g, \mathbf{x}_i, y_i) =$$

$$\arg\min_{g \in \mathcal{G}} \sum_{i=1}^{N} \mathbb{1}\left[g(\mathbf{x}_i) \neq \text{pseudo label of } \mathbf{x}_i\right] \times \text{weight of } \mathbf{x}_i$$

Reduction to supervised learning

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{N} \mathbb{1}\left[g(\mathbf{x}_i) \neq \text{pseudo label of } \mathbf{x}_i\right] \times \text{weight of } \mathbf{x}_i$$

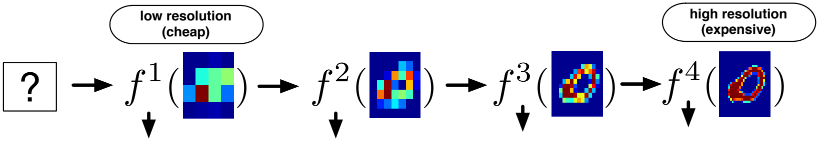
Can be solved with existing supervised learning tools

- pick a surrogate loss $\mathcal{L}[z] \geq \mathbb{1}_{[z>0]}$ (e.g. logistic)
- pick a classifier family $\mathcal{G}$ (e.g. linear)

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{N} \mathcal{L}\left[g(\mathbf{x}_i) \times \text{ pseudo label of } \mathbf{x}_i\right] \times \text{weight of } \mathbf{x}_i$$

Sensors Varying Resolutions

classify handwritten digit images (mnist)

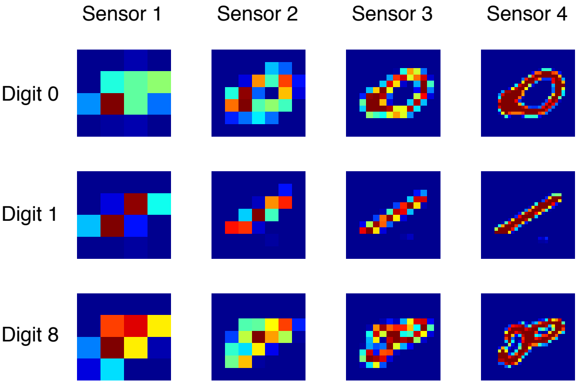


x handwritten digit image
$y \in \{0, 1, \dots, 9\}$ label

Stage	1	2	3	4
Sensor	4x4	8x8	16x16	32x32
Cost	1	2	3	4

Base Learner: logistic regression with linear classifiers

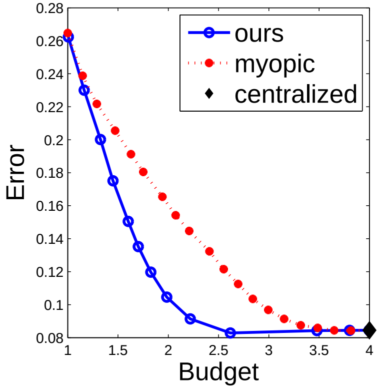
Example

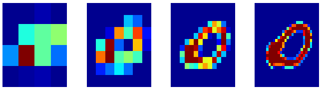


Sensor selection depends on example

Handwritten Digit Dataset

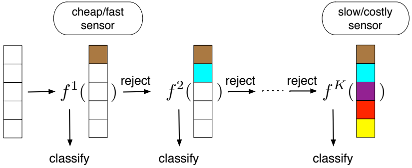
Same performance as centralized (best) with much lower budget





Generalize to Multiple Stages

Measurement $\mathbf{x} = [x_1, \ldots, x_K]$ and true label y



Seek decisions at each stage $F = f_1, f_2, \ldots f_k$

$$f_k(\mathbf{x}) = \begin{cases} h_k(\mathbf{x}), & g_k(\mathbf{x}) = \text{not reject} \\ \text{reject}, & \text{else} \end{cases}$$

h_k is a standard classifier trained on sensors $1, \ldots, k$

Stage-wise Decomposition

System Risk: $R(F, \mathbf{x}, y) = \text{Loss}(F(\mathbf{x}), y) + \alpha \text{Cost}(F, \mathbf{x})$

Stage-wise Decomposition

System Risk: $R(F, \mathbf{x}, y) = \text{Loss}(F(\mathbf{x}), y) + \alpha \text{Cost}(F, \mathbf{x})$

Stage-wise recursion:

$$R(F, \mathbf{x}, y) = R_0(F, \mathbf{x}, y)$$

$$R_k(\mathbf{x}, y, f_k) = \begin{cases} \alpha c_{k+1} + R_{k+1}(\cdot), & \text{reject to next stage} \\ 1, & \text{error \& not reject} \\ 0, & \text{correct \& not reject} \end{cases}$$

Stage-wise Decompostion

$$R_k(\mathbf{x}, y, f_k) = \begin{cases} \alpha c_{k+1} + R_{k+1}(\cdot), & \text{reject to next stage} \\ 1, & \text{error \& not reject} \\ 0, & \text{correct \& not reject} \end{cases}$$

Key Observation:
Given the past: $f_1, \dots, f_{k-1}$, and the future: $f_{k+1}, \dots, f_K$

- Find current decision, f_k, from single stage risk R_k
- Equivalent to a two stage problem:

$$R_{cu} = \mathbb{1}_{[h_k \text{ misclassifies } \mathbf{x}]}$$
$$R_{fu} = R_{k+1}(\mathbf{x}, \dots)$$

Algorithm

For every training example $\mathbf{x}_i$:

- $R_{k+1}(\mathbf{x}_i, \ldots)$: **cost-to-go**, empirical risk of future stages,
- **state**$_k(\mathbf{x}_i)$**:** indicates if example is still active at stage k

Algorithm

For every training example $\mathbf{x}_i$:

- $R_{k+1}(\mathbf{x}_i, \ldots)$: **cost-to-go**, empirical risk of future stages,
- $\mathbf{state}_k(\mathbf{x}_i)$**:** indicates if example is still active at stage k

Algorithm: alternatively minimize one stage at a time

For every stage k:

1: Learn decision f_k:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \mathbf{state}_k(\mathbf{x}_i)\ R_k[f, \mathbf{x}_i, y_i, R_{k+1}(\cdot)]$$

2: Update $\mathbf{state}_j(\mathbf{x}_i)$ for future stages $j > k$
3: Update $\mathbf{cost\text{-}to\text{-}go}(\mathbf{x}_i)$ for past stages $j < k$

Repeat until convergence

Other Experiments

Achieve target error rate with fraction of max budget

Dataset	Stages	Sensors	Target Error	Myopic	**Ours**
synthetic	2		.147	52%	**28%**
pima	3	weight, age, blood tests	.245	41%	**15%**
threat	3	ir,pmmw,ammw	.16	89%	**71%**
covertype	3	soils, wild. areas, elev, aspect	.285	79%	**40%**
letter	3	pixel counts, moments, edge feat's	.25	81%	**51%**
mnist	4	res. levels	.085	90%	**52%**
landsat	4	hyperspectral bands	.17	56%	**31%**
mam	2	CAD feat's, expert rating	.173	65 %	**25%**

Generalization Results

How well does it perform on unseen data, $\mathbf{E}_{\mathbf{x},y}\left[F(\mathbf{x}) \neq y\right]$?

Generalization Results

How well does it perform on unseen data, $\mathbf{E}_{\mathbf{x},y}\left[F(\mathbf{x}) \neq y\right]$?

Standard VC dimension test error bound:

For a classifier $F(\mathbf{x})$ in a family $\mathcal{F}$ with VC dimension $= h$, w.p. $1-\delta$,

$$\text{Test Error} \leq \text{Train Error} + \sqrt{\frac{h\log(\frac{2N}{h}+1)+\log\frac{4}{\delta}}{N}}$$

Smaller VC dimension $\rightarrow$ better generalization

Generalization Results

System VCD does not explode!

Theorem: VCD of a K stage sequential decision:

$$\leq O(K \log K) \max_k \{\mathrm{VCD}(\mathcal{F}_k)\}$$

$\mathrm{VCD}(\mathcal{F}_k)$ is VC dimension of kth stage

Complexity grows only as $K \log K$ times the most complex stage

Conclusion

- Introduced sequential decision problem
- Myopic approach: relies on current uncertainty to make a decision
  - Considered synthetic examples
  - Current uncertainty is not always enough
- Our approach: incorporate future uncertainty in current decision
  - Examined a two stage system
  - Reduced to supervised learning
- Experiment
- Extend to Multiple Stages
- Generalization Results

Future Work

Optimization Improvement

- Currently, cyclical local optimization of each stage
- Need a convex formulation of system risk, achieve global optimum and better performance

More general architecture

- Option to skip a sensor if unnecessary
- Arbitrary sensing order, intractable even with full models, need approximations

Read our paper:

K. Trapeznikov, V. Saligrama,
Supervised Sequential Classification Under Budget Constraints,
AISTATS, 2013

we are organizing:

Workshop on Learning with Test Time Budgets
International Conference on Machine Learning, Atlanta, June 21-22

website: https://sites.google.com/site/budgetedlearning2013/

Thanks for Listening!

[Ji and Carin, 2007] Ji, S. and Carin, L. (2007).
Cost-sensitive feature acquisition and classification.
In *Pattern Recognition*.

[Kanani and Melville, 2008] Kanani, P. and Melville, P. (2008).
Prediction-time active feature-value acquisition for cost-effective customer targeting.
In *NIPS*.

[Kapoor and Horvitz, 2009] Kapoor, A. and Horvitz, E. (2009).
Breaking boundaries: Active information acquisition across learning and diagnosis.
In *NIPS*.

[Koller and Gao, 2011] Koller and Gao (NIPS 2011).
Active value.

[Liu et al., 2008] Liu, L.-P., Yu, Y., Jiang, Y., and Zhou, Z.-H. (2008).
Tefe: A time-efficient approach to feature extraction.
In *ICDM*.

[Zubek and Dietterich, 2002] Zubek, V. B. and Dietterich, T. G. (2002).
Pruning improves heuristic search for cost-sensitive learning.
In *ICML*.