

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

HIERARCHICAL ACTIVE LEARNING (HAL) APPLICATION TO MITOCHONDRIAL DISEASE PROTEIN DATASET

James Duin

University of Nebraska–Lincoln
Master's Thesis

Spring 2017

jamesdduin@gmail.com

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- Identify the source of mutations which give rise to mitochondrial disease
- Leigh Syndrome, Lebers Hereditary Optic Neuropathy
- Hierarchically labeled according to location in mitochondria
- Coarse-grained: learning labels near the root of the tree
- Fine-grained: learning labels towards the leaf nodes
- Learn mitochondrion concept (coarse) by combining classifiers for each target compartment (fine)

- **Active learning:** copious unlabeled data, cost associated with acquiring labels, yields best classifier for a given cost, or best for minimal cost
- Previous work in text classification and rich media indexing use hierarchies of labels to improve fine-level classification (McCallum et al. 1998, Jiang et al. 2013)
- Previous work in named entity recognition to target fine-grained entity categories (Fleischman et al. 2002)
- **First, investigation of active learning in a hierarchical setting, approach shown to find best classifier for a budget regardless of varying label acquisition cost**

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

Outline:

- Active Machine Learning
- Hierarchical Protein Dataset
- Coarse-grained vs Fine-grained Trade Off
- Active over-labeling algorithms
- Application to Protein Dataset
- Experimental Results

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

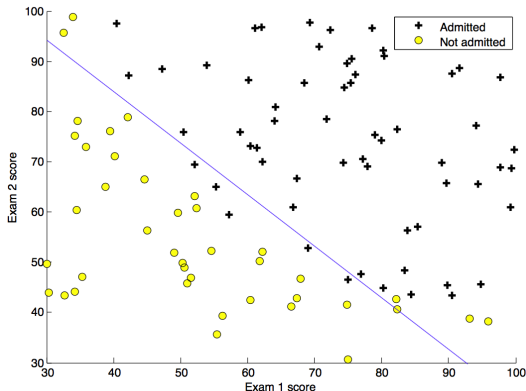
BANDIT
Results

Conclusions

INPUT: labeled data

OUTPUT: learned hypothesis used to predict new instances

$h_{\theta}(x)$, for fixed θ_0 and θ_1 line coefficients



Machine Learning

Cost Function

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

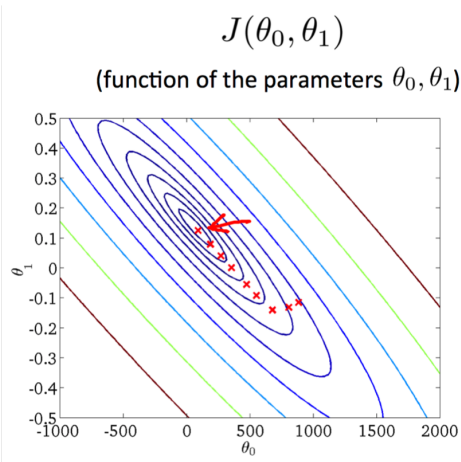
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

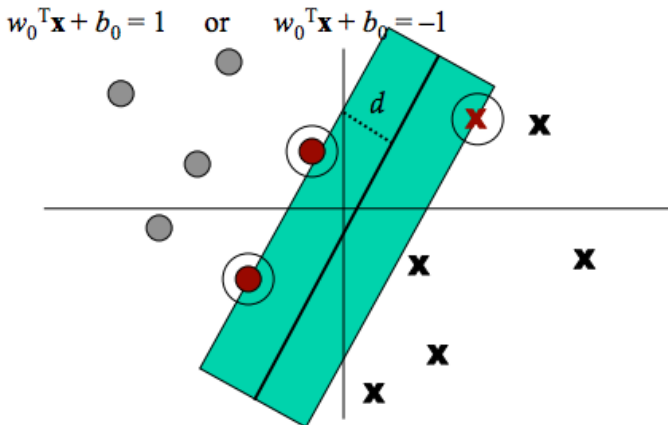
Conclusions



Machine Learning

Support Vector Machine (SVM)

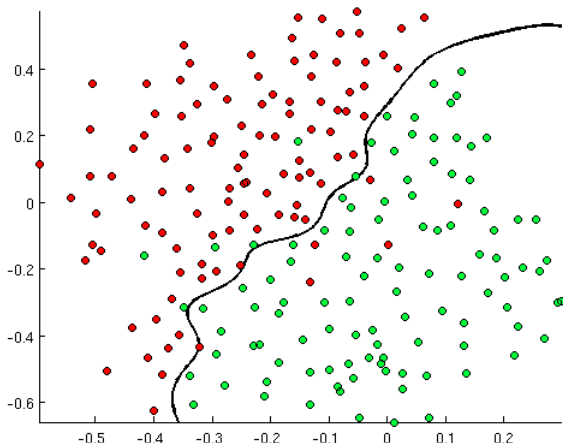
And SVM constructs a line, plane or hyperplane that separates the features with the greatest margin.



Machine Learning

Support Vector Machine (SVM)

The greater the functional margin the lower the generalization error of the classifier



Machine Learning

Support Vector Machine (SVM)

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

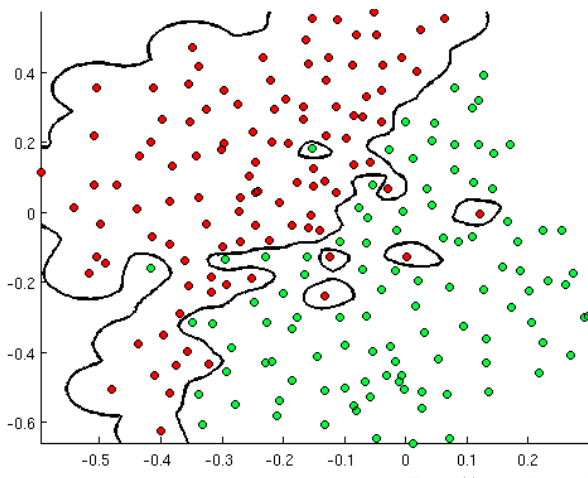
Act. vs Pass.

HAL Results

BANDIT
Results

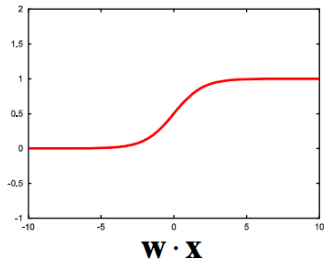
Conclusions

Kernel functions implicitly map inputs into high-dimensional feature spaces



Logistic Regression (Logit) estimates the probability of a binary response, learns coefficients \mathbf{w} of the input vector \mathbf{x} and passes dot product through sigmoid function. (Maximum likelihood learning)

$$g(\mathbf{x}, \mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x})}$$



- The learner queries an **oracle** or **supervisor** which labels the data at a certain cost
- Active learning solicits new instances that can maximally improve performance of the learned classifier
- Learns the best performing classifier for the minimal amount of labeling cost, or for a given purchase budget
- Acquires labels for each level of the hierarchy at a certain cost, spends according to a purchase budget

Hierarchical Bioinformatics Data Set

Feature Sources

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

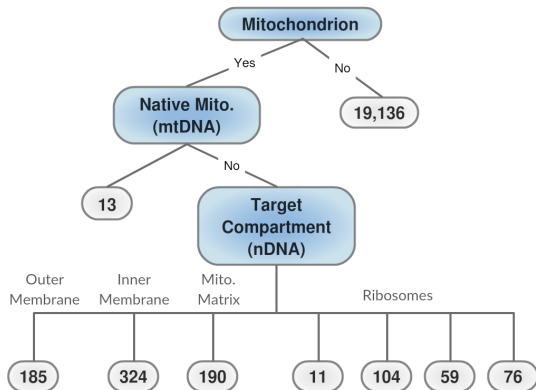
Conclusions

- Mitoproteome: database of human mitochondrial proteins
- SwissProt: database of experimentally validated human proteins

Type of Properties	Features	Sources
General sequence features	Amino acid composition, sequence length, etc.	Cui et al, PROFEAT
Physico chemical properties	Hydrophobicity, polarity, etc.	Cui et al, PROSO, Phoebus
Structural properties	Secondary structural content, shape, etc.	SSCP
Domains and motifs	Signal peptide, transmembrane domains, etc.	SignalP, TMB-Hunt, NetOgly, TatP

Hierarchical Bioinformatics Data Set

Labeling Hierarchy



Coarse-grained vs Fine-grained Trade Off

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

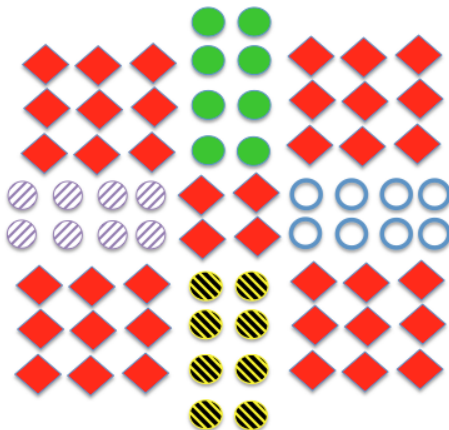
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Hierarchical Active Learning

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

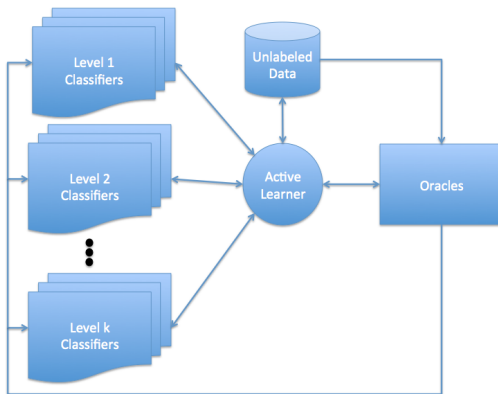
Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

INPUT: purchase proportion p



Dynamically Adapting Purchase Proportions

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

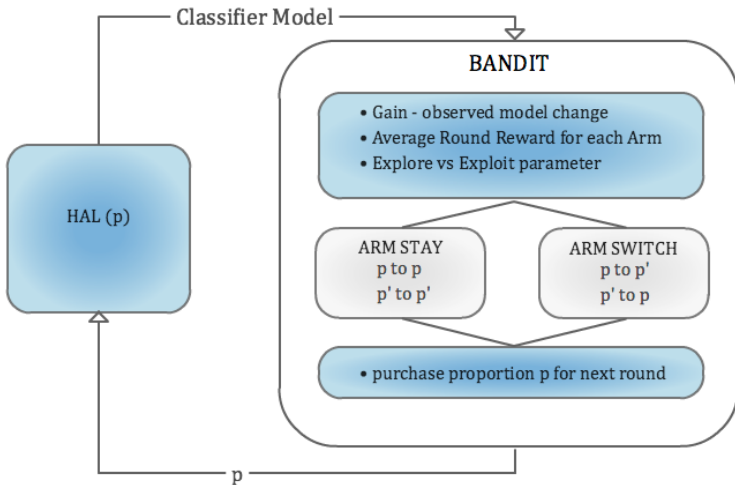
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Training and Testing Coarse-Grain and Fine-Grain Classifiers

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

Number of proteins in each class:

Classes	Count	Totals
Non Mito 0	19136	All: 20098
mtDNA 1	13	Coarse: 19136
nDNA 2	185	Fine: 962
nDNA 3	324	Features: 449
nDNA 4	190	
nDNA 5	11	
nDNA 6	104	
nDNA 7	59	
nDNA 8	76	

SVM and Logit Classifier Performance

F-measure Analysis

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

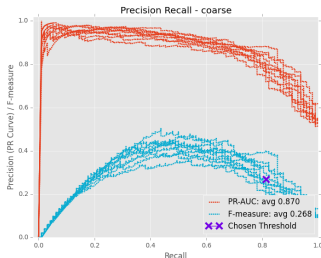
Conv. ML

Act. vs Pass.

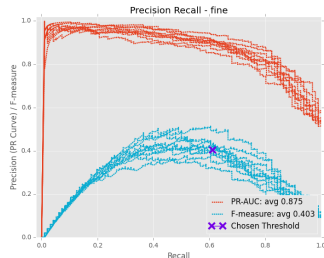
HAL Results

BANDIT
Results

Conclusions



(a) Log Reg Pr Curves - Coarse



(b) Log Reg Pr Curves - Fine

SVM and Logit Classifier Performance

F-measure Analysis

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

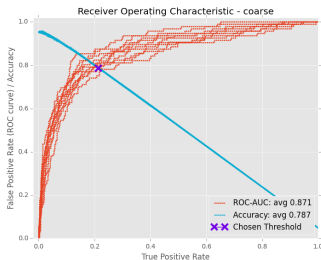
Conv. ML

Act. vs Pass.

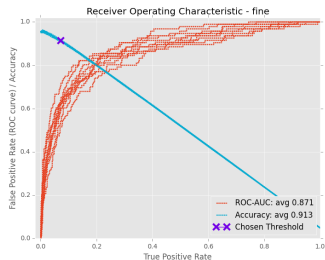
HAL Results

BANDIT
Results

Conclusions



(a) Log Reg ROC Curves - coarse



(b) Log Reg ROC Curves - fine

Active vs. Passive Curve Analysis

Logit PR-AUC curves

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

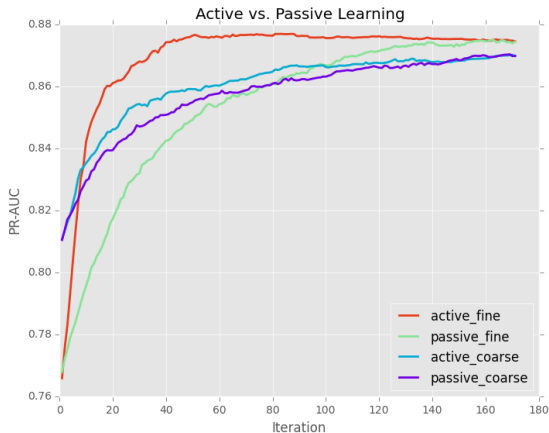
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Active vs. Passive Curve Analysis

Logit ROC-AUC curves

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

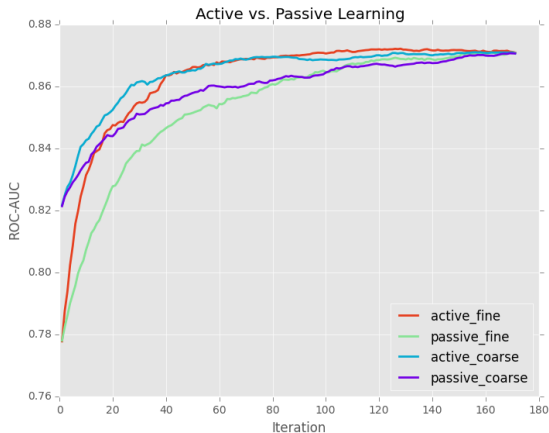
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Active vs. Passive Curve Analysis

SVM PR-AUC curves

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

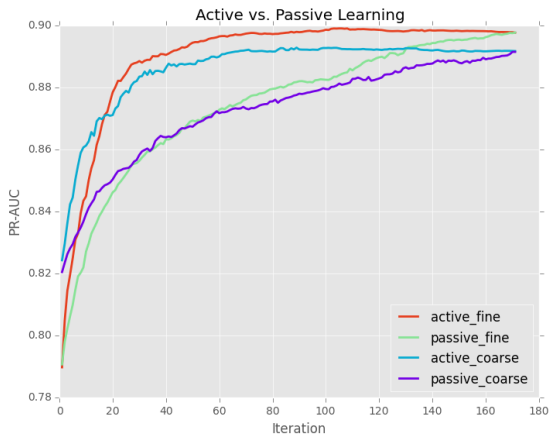
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Active vs. Passive Curve Analysis

SVM ROC-AUC curves

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

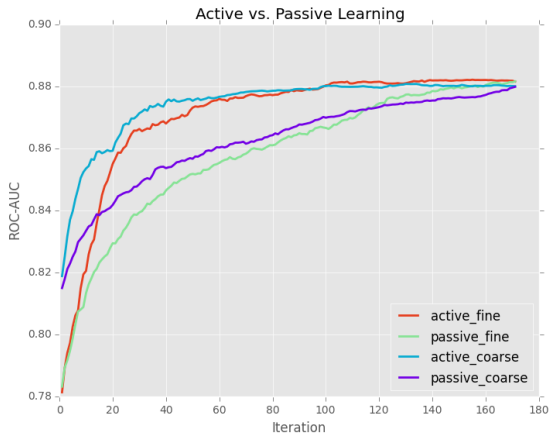
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Plots for Fine Fixed Ratio Results

Fine Cost 1

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

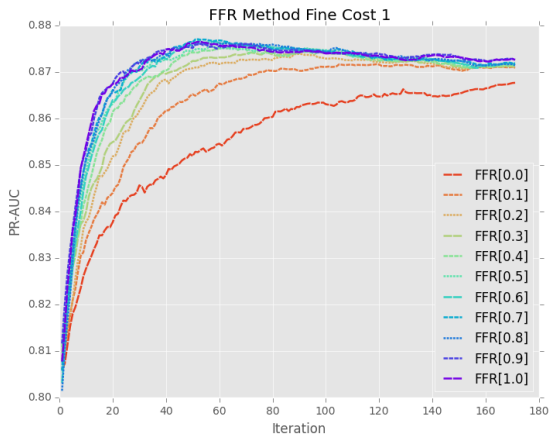
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Plots for Fine Fixed Ratio Results

Fine Cost 2

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

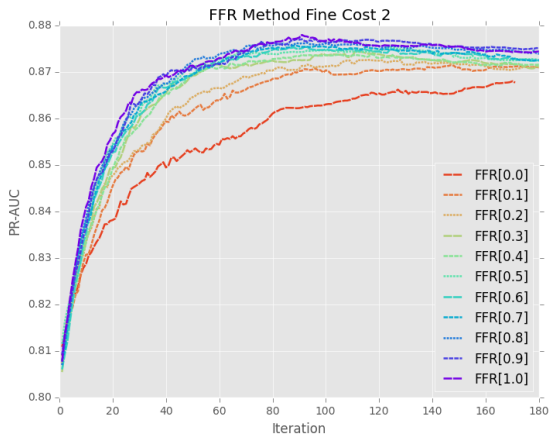
Conv. ML

Act. vs Pass.

HAL Results

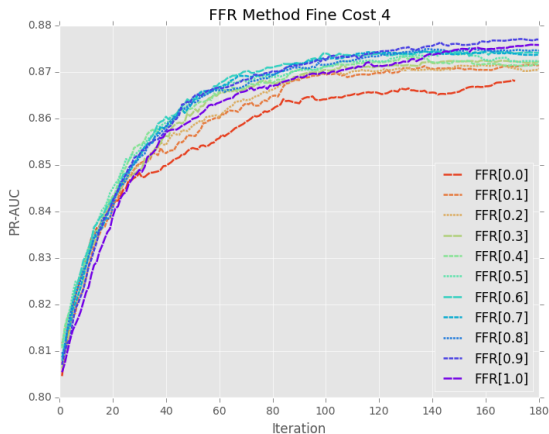
BANDIT
Results

Conclusions



Plots for Fine Fixed Ratio Results

Fine Cost 4



Plots for Fine Fixed Ratio Results

Fine Cost 8

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

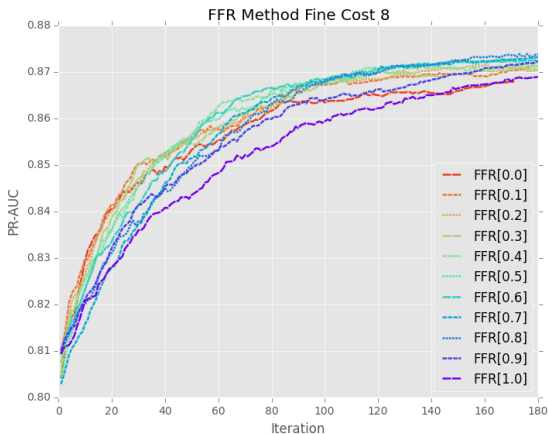
Conv. ML

Act. vs Pass.

HAL Results

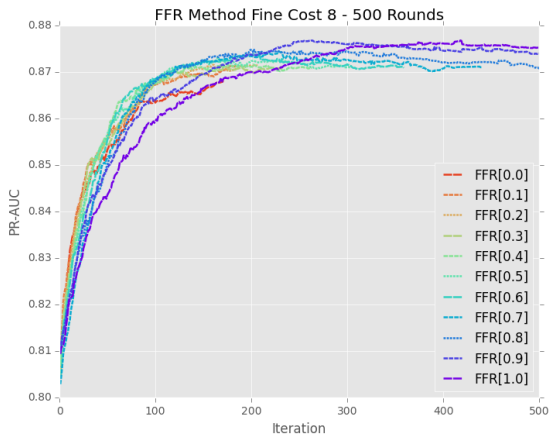
BANDIT
Results

Conclusions



Plots for Fine Fixed Ratio Results

Fine Cost 8 - Rnds to 500



Plots for Fine Fixed Ratio Results

Fine Cost 8 - Rnds 20 to 60

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

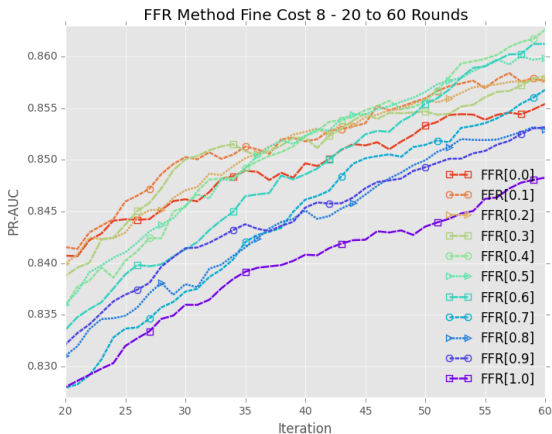
Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Plots for Fine Fixed Ratio Results

Fine Cost 16

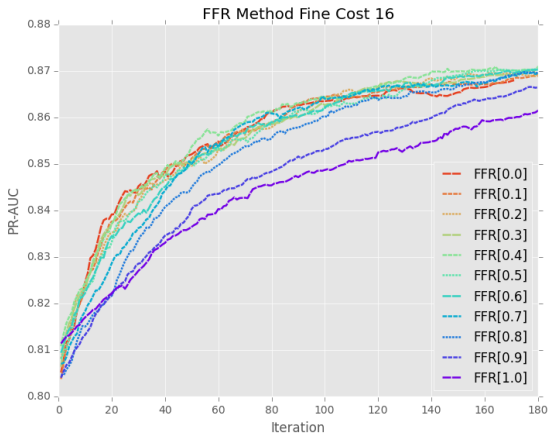


Figure: The fine cost is increased to 16. The fine cost is too high to offset the decreased number of instances purchased.

BANDIT Approach Results

Varying Cost Analysis

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

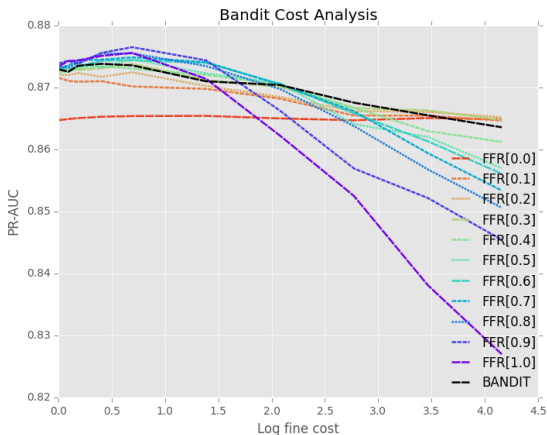
BANDIT
Results

Conclusions

- The BANDIT approach is compared to the previous FFR curves for the following fine-grain costs $\{1.0, 1.1, 1.2, 1.5, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0\}$
- Budget held fixed at round 120.
- The metric *diff* is the learner's absolute difference in PR-AUC from the top learner for a given cost.
- The metric *rank* is the learners 0 indexed ranking in terms of PR-AUC for a given cost.

BANDIT Approach Results

Varying Cost Analysis - Plot



HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

**BANDIT
Results**

Conclusions

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡ 🔍 ↺

BANDIT Approach Results

Varying Budget Analysis - Mixed Cost

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

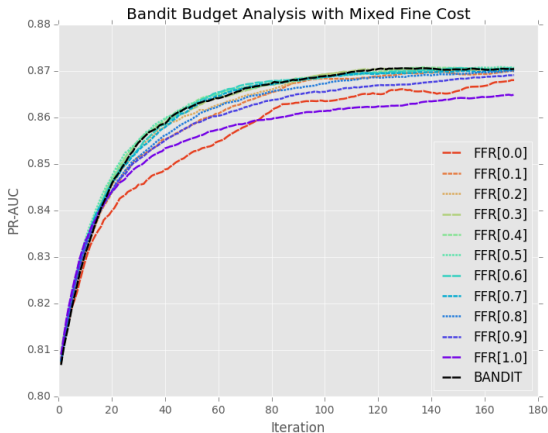
Conv. ML

Act. vs Pass.

HAL Results

**BANDIT
Results**

Conclusions



BANDIT Approach Results

BANDIT - Rnds 20 to 60

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

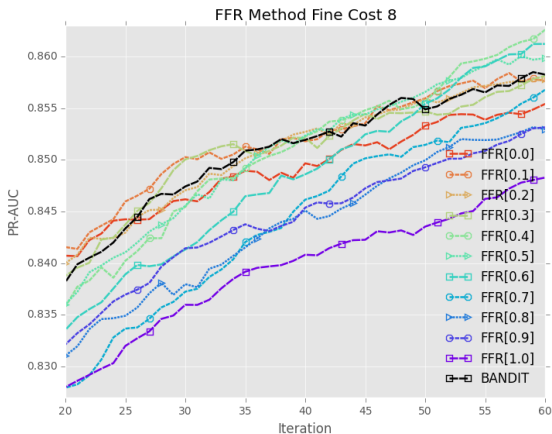
Conv. ML

Act. vs Pass.

HAL Results

**BANDIT
Results**

Conclusions



HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- Demonstrated fine-grained labels can be used to improve a coarse-grained classifier for the protein dataset
- Demonstrated a prominent advantage for active fine with the Logit classifier
- HAL is implemented and applied to the protein dataset for various FFR proportions and fine label costs
- The BANDIT approach is shown to be robust to both labeling cost and budget

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- Future work is to apply the active over-labeling approach to other datasets with more complex hierarchical label trees; datasets derived from Gene Ontology research could be investigated

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- I would like to thank my advisor Dr. Stephen Scott, Yugi Mo and Dr. Douglas Downey for continued guidance. I would like to thank Dr. Juan Cui and Dr. Ashok Samal for serving on my committee. Additionally, I would like to thank Jiang Shu and Kevin Chiang for their assistance accessing and understanding the protein dataset.

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions



Active vs. Passive Curve Analysis

Logit Accuracy

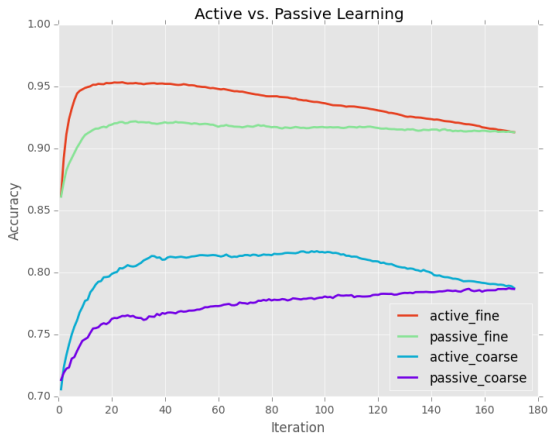


Figure: The accuracy of the classifiers stays at roughly the same rate throughout the rounds; this is due to an effective weighting scheme.

Active vs. Passive Curve Analysis

Logit F-measure

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

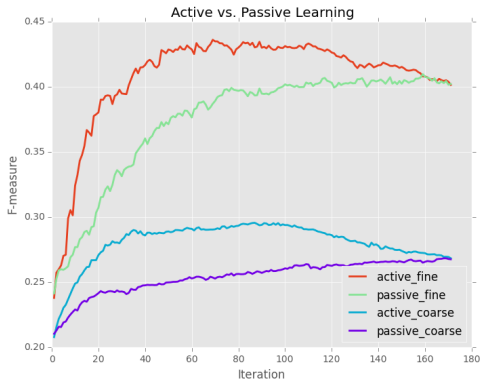


Figure: Both curves show a dominance of fine over coarse and Active over Passive.

Dynamically Adapting Purchase Proportions p or p'

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- For round n , calculate gain g in terms of observed model change
- Calculate average round reward for each arm
- Calculate $\varepsilon_n = \min \left\{ 1, \frac{2}{n} \right\}$
- With probability $1 - \varepsilon_n$ play arm with highest current average reward for round n , otherwise explore
- After playing arm, run HAL with chosen p or p'

ARM STAY	ARM SWITCH
$r(n) = 0$	$r(n) = \begin{cases} -g(n)/ g(n) & \text{if } p \rightarrow p' \\ g(n)/ g(n) & \text{if } p' \rightarrow p \\ 0 & \text{if } p \rightarrow p \text{ or } p' \rightarrow p' \end{cases}$

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- Y. Mo, S. D. Scott, and D. Downey, Learning hierarchically decomposable concepts with active over-labeling, in 2016 IEEE 16th International Conference on Data Mining (ICDM), Dec 2016, pp. 340349.
- J. Z. Juan Cui, Kevin Chiang, Prediction of nuclear and locally encoded mitochondrion. Lincoln, NE: Nebraska Gateway to Nutrigenomics 6th Annual Retreat, June 9 2014. [Online]. Available: <http://cehs.unl.edu/nutrigenomics/nebraska-gateway-nutrigenomics-6th-annual-retreat/>
- T. M. Mitchell, Machine Learning, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, API design for machine learning software: experiences from the scikit-learn project, in ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108122
- D. Cotter, P. Guda, E. Fahy, and S. Subramaniam, Mitoproteome: mitochondrial protein sequence database and annotation system, Nucleic Acids Research, vol. 32, no. suppl1, p. D463, 2004. [Online]. Available: +<http://dx.doi.org/10.1093/nar/gkh048>

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- J. Cui, L. Y. Han, H. Li, C. Y. Ung, Z. Q. Tang, C. J. Zheng, Z. W. Cao, and Y. Z. Chen, Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties, Molecular Immunology, vol. 44, no. 4, pp. 514 520, 2007. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S016158900>
- A. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng, Improving text classification by shrinkage in a hierarchy of classes, in Proceedings of the Fifteenth International Conference on Machine Learning, ser. ICML 98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998, pp. 359367. [Online]. Available: <http://dl.acm.org/citation.cfm?id=645527.657461>

HAL - Protein

James Duin

Introduction

Background

Exp. Setup

Conv. ML

Act. vs Pass.

HAL Results

BANDIT
Results

Conclusions

- W. Jiang and Z. W. Ras, Multi-label automatic indexing of music by cascade classifiers, Web Intelli. and Agent Sys., vol. 11, no. 2, pp. 149170, Apr. 2013. [Online]. Available:
<http://dl.acm.org/citation.cfm?id=2590084.2590088>
- etc.