

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

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

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- Identify the source of mutations which give rise to mitochondrial disease.
- Leigh Syndrome, Lebers Hereditary Optic Neuropathy
- Hierarchically labeled according to location in mitochondria
- Learn mitochondrion concept (Coarse) by combining classifiers for each target compartment (Fine)

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- 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)
- This work is done in conjunction with Yugi Mo, Dr. Scott, and Dr. Downey
- First investigation of active learning in a hierarchical setting where label acquisition cost can vary

- Active Machine Learning
- Evaluating Classifier Performance
- Hierarchical Protein Dataset
- Coarse-grained vs Fine-grained Trade Off
- Active Over-Labeling
- Application to Protein Dataset

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- Machine learning (ML) algorithms are defined as computer programs that learn from experience E with respect to some class of tasks T and performance measure P , if their performance at tasks in T , as measured by P , improves with experience E - (Mitchell 1997).
- Support Vector Machine (SVM) - Uses support vectors and kernel functions
- Logistic Regression (Logit) - Uses logistic function

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- 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

Evaluating Classifier Performance

Confusion Matrix

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

Divide data into train and a test set. Analyze test set with the following values:

- True-Negatives (T_n): Correctly classified negatives
- False-Negatives (F_p): Incorrectly classified negatives
- False-Positives (F_n): Incorrectly classified positives
- True-Positives (T_p): Correctly classified positives

Example of a confusion matrix for a test set with 100 negatives and 50 positives:

conf (T_n/F_n)	conf (F_p/T_p)
90	10
20	30

Evaluating Classifier Performance

Precision and Recall

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

Precision is a measure of result relevancy:

$$P = \frac{T_p}{T_p + F_p} \quad (1)$$

Recall is a measure of how many truly relevant results are returned:

$$R = \frac{T_p}{T_p + F_n} \quad (2)$$

Evaluating Classifier Performance

F-Measure

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

The F-measure or F1-measure (F1) is the harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (3)$$

Evaluating Classifier Performance

ROC - PR curves

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

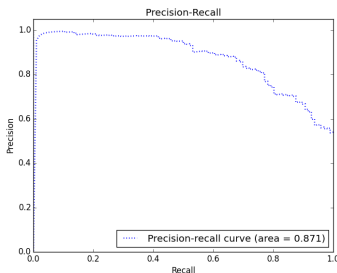
Act. vs Pass.

FFR Results

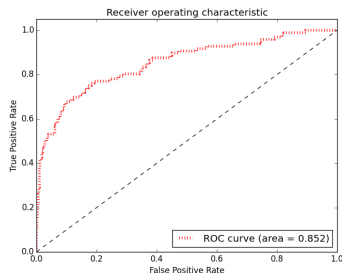
BANDIT
Results

Conclusions

Bibliography



(a) PR curve.



(b) ROC curve.

Figure: Examples of PR and ROC curves with their corresponding AUC values.

Hierarchical Bioinformatics Data Set

Feature Sources

Table: Features of the protein dataset along with their respective sources:

Type of Properties	Features	Sources
General sequence features	Amino acid composition, sequence length, etc.	Calculated by Kevin Chiang at UNL, etc.
Physico chemical properties	Hydrophobicity, polarity, etc.	Computed from Cui et al, etc.
Structural properties	Secondary structural content, shape, etc.	SSCP, etc.
Domains and motifs	Signal peptide, transmembrane domains, etc.	SignalP, NetOgly, etc.

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

Hierarchical Bioinformatics Data Set

Labeling Hierarchy

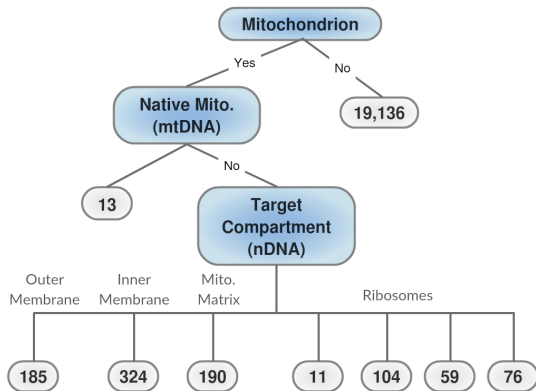


Figure: The protein dataset hierarchy of labels along with the instance count for each label.

Coarse-grained vs Fine-grained Trade Off

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

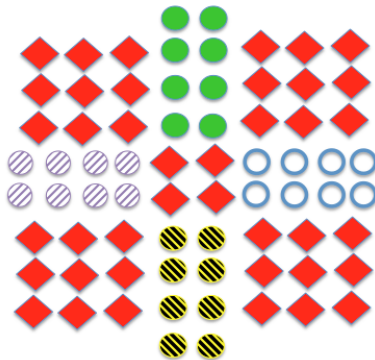


Figure: Demonstration of a dataset that would benefit from multiple fine-grained learners for each circle type, from Mo et al.

Active Over-Labeling

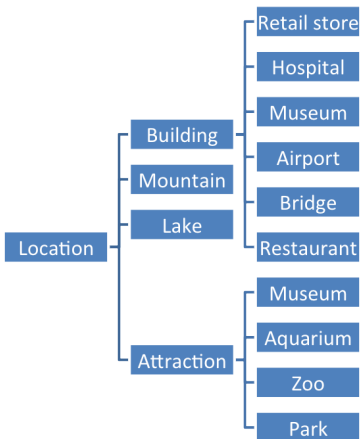


Figure: A labeling tree based on the text categorization dataset RCV1, from Mo et al.

Hierarchical Active Learning

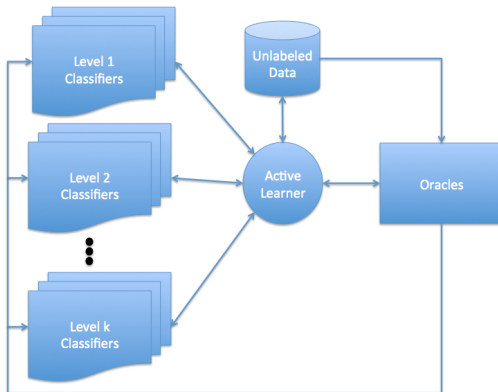


Figure: Diagram of HAL approach

Dynamically Adapting Purchase Proportions

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- HAL is a fixed-fine ratio (FFR) methodology
- Input is a purchase proportion vector p , which allocates budget to purchase labels at a given level in the hierarchy
- The task of choosing the level of granularity to purchase labels is solved using Auer et al.'s ϵ -greedy bandit algorithm
- With probability $1 - \epsilon_n$ play arm with highest current average reward for round n , otherwise explore

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

Analysis and evaluation follow Mo et al.'s work.

- Fine outperforms Coarse in PR-AUC
- Active outperforms Passive in PR-AUC
- HAL ran with variable cost, fine proportions and budget
- BANDIT approach shown to be robust to changes in cost and budget

Training and Testing Coarse-Grain and Fine-Grain Classifiers

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

Number of proteins in each class:

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

Training and Testing Coarse-Grain and Fine-Grain Classifiers

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

Table: Number of proteins in each partition:

Folds	All	0	1	2	3	4	5	6	7	8
1	2010	1914	1	19	32	19	1	11	6	7
2	2010	1914	1	19	32	19	1	11	6	7
3	2010	1914	1	19	32	19	1	11	5	8
4	2010	1914	1	19	32	19	1	10	6	8
5	2010	1914	1	18	33	19	1	10	6	8
6	2010	1914	1	18	33	19	1	10	6	8
7	2010	1913	2	18	33	19	1	10	6	8
8	2010	1913	2	18	33	19	1	10	6	8
9	2009	1913	2	18	32	19	2	10	6	7
10	2009	1913	1	19	32	19	1	11	6	7
Total	20098	19136	13	185	324	190	11	104	59	76

Training and Testing Coarse-Grain and Fine-Grain Classifiers

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

The following variables were varied for both SVM and Logit classifiers:

- Preprocessing Scaling Methods
- Preprocessing Feature Selection
- Class Weight
- SVM Kernel, Cost, and Gamma parameters
- Logit Cost, Fine class weights, Tolerance

SVM and Logit Classifier Performance

Conventional ML

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

Table: Logit results after parameter tuning:

Title	PR	ROC	Acc	F1	conf (tn/fn)	conf (fp/tp)
coarse	0.870	0.871	0.787	0.268	(1503.2 / 17.8)	(410.4 / 78.3)
fine	0.875	0.871	0.913	0.403	(1776.5 / 37.3)	(137.1 / 58.8)

Table: SVM results after parameter tuning:

Title	PR	ROC	Acc	F1	conf (tn/fn)	conf (fp/tp)
coarse	0.892	0.880	0.866	0.347	(1669.5 / 24.8)	(244.1 / 71.3)
fine	0.898	0.882	0.942	0.485	(1839.0 / 41.5)	(74.6 / 54.6)

SVM and Logit Classifier Performance

F-measure Analysis

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

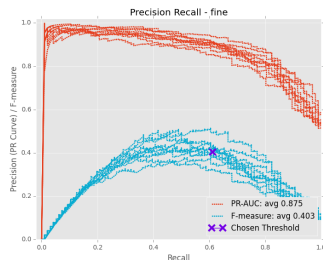
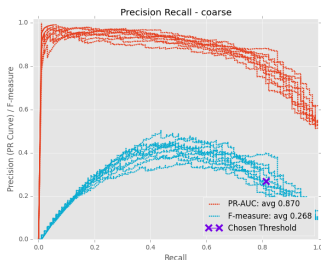
Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography



(a) Log Reg Pr Curves - Coarse

(b) Log Reg Pr Curves - Fine

Figure: The fine default threshold occurs at a point on the PR curve associated with a higher F-measure score compared to the coarse curves.

SVM and Logit Classifier Performance

F-measure Analysis

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

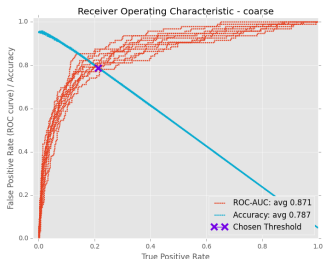
Act. vs Pass.

FFR Results

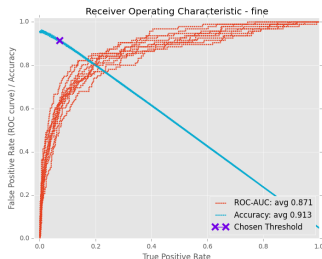
BANDIT
Results

Conclusions

Bibliography



(a) Log Reg ROC Curves - coarse



(b) Log Reg ROC Curves - fine

Figure: Fine has a higher accuracy than coarse at the default threshold for the Logit classifier.

Active vs. Passive Curve Analysis

Logit PR-AUC curves

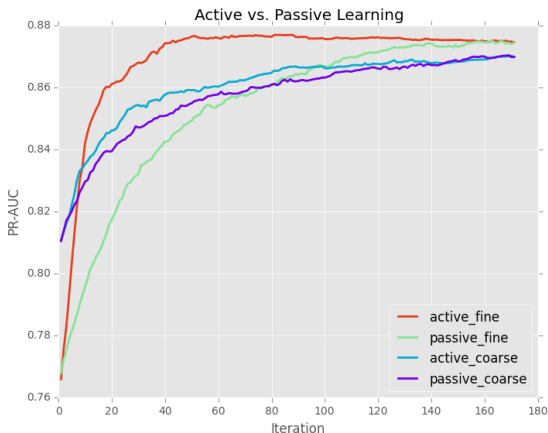


Figure: The PR-AUC curves for rounds with the Logit classifier conforms to expectations

Active vs. Passive Curve Analysis

Logit ROC-AUC curves

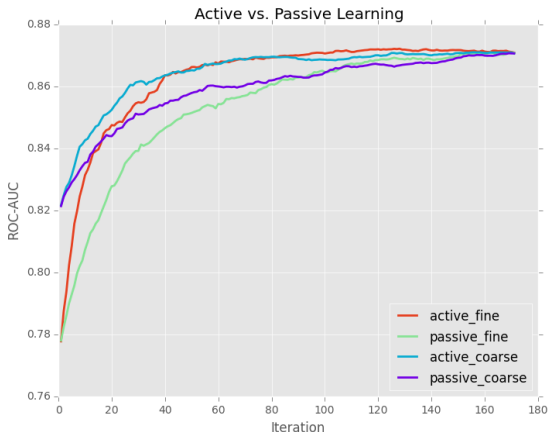


Figure: The ROC-AUC curves for rounds with the Logit classifier; active curves beat out the passive curves for both coarse and fine.

Active vs. Passive Curve Analysis

SVM PR-AUC curves

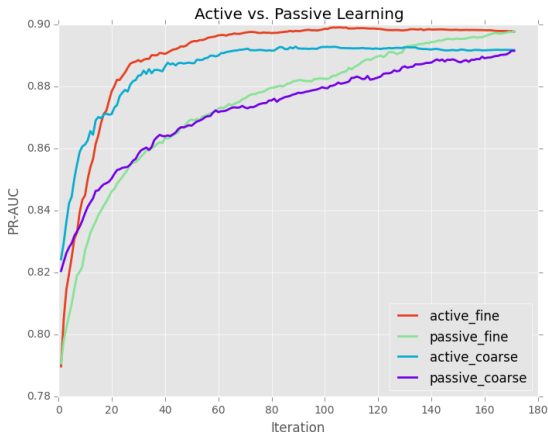


Figure: The PR AUC curves for SVM show a slight advantage for active fine, similar to the Logit results.

Active vs. Passive Curve Analysis

SVM ROC-AUC curves

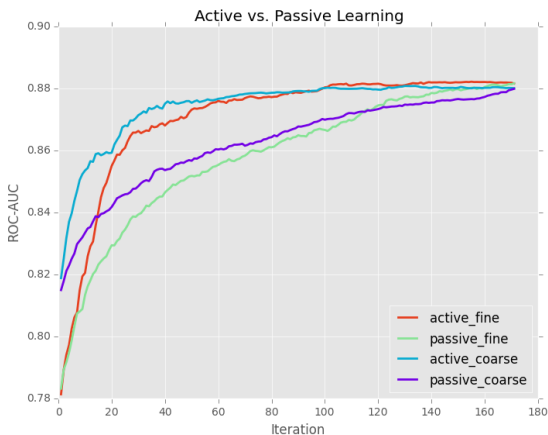


Figure: The ROC AUC curves for SVM match the Logit results, the convergence of active fine to active coarse takes slightly longer.

Plots for Fine Fixed Ratio Results

Fine Cost 1

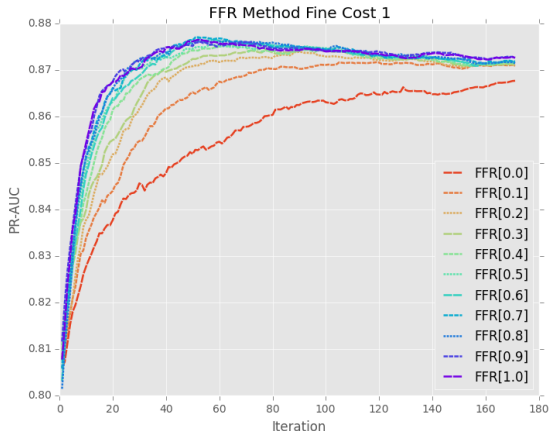


Figure: The fine and coarse grain labels both have a cost of 1.

Plots for Fine Fixed Ratio Results

Fine Cost 2

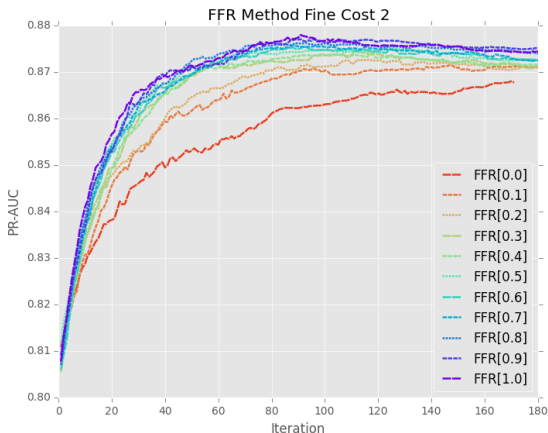


Figure: At fine cost 2, advantage of the higher FFR values decreases but the ordering of the curves remains unchanged.

Plots for Fine Fixed Ratio Results

Fine Cost 4

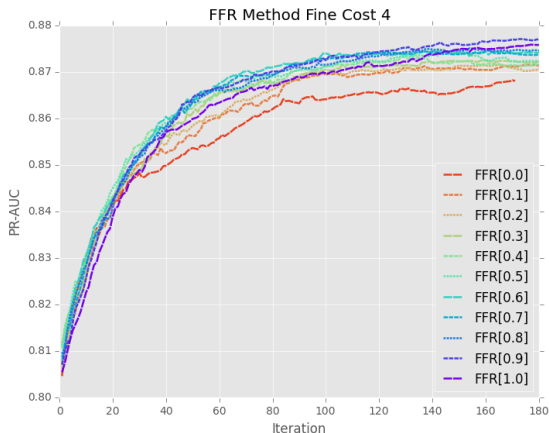


Figure: At fine cost 4, the highest FFR 1.0 is no longer preferred. Purchasing a greater number of coarse instances is a better strategy.

Plots for Fine Fixed Ratio Results

Fine Cost 8

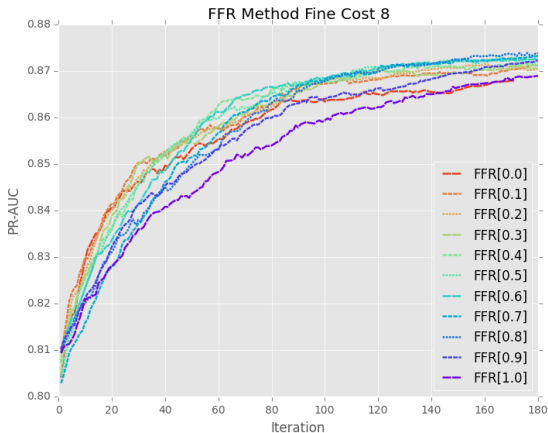


Figure: At fine cost 8 the middle FFR values outperform the extreme values for rounds 0 to 180.

Plots for Fine Fixed Ratio Results

Fine Cost 8 - Rnds to 500

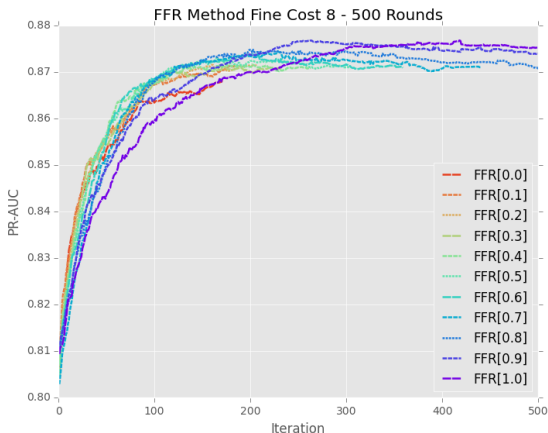


Figure: This shows the iterations continuing through round 500, the curves with the higher fine rates settle to the same end point.

Plots for Fine Fixed Ratio Results

Fine Cost 8 - Rnds 20 to 60

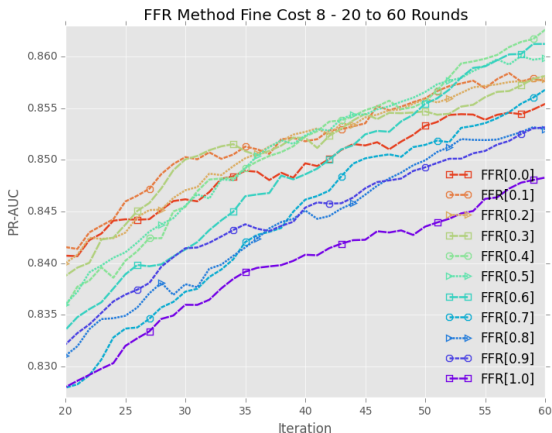


Figure: The fine cost 8 curves shown expanding the rounds 20-60. If a round budget of 40 occurs than the recommended FFR would be 0.2.

Plots for Fine Fixed Ratio Results

Fine Cost 16

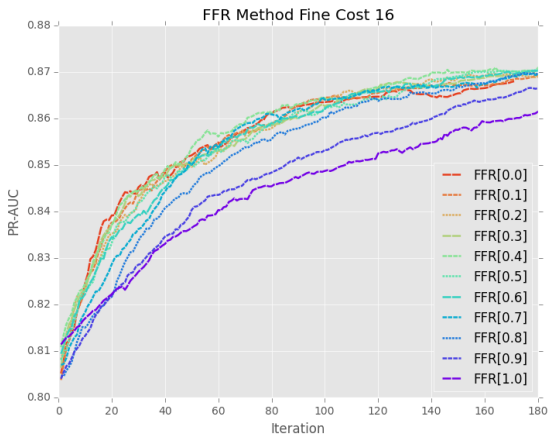


Figure: The fine cost is increased to 16. The fine cost is too high to offset the decreased number of instances purchased.

BANDIT Approach Results

Varying Cost Analysis

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

**BANDIT
Results**

Conclusions

Bibliography

- 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

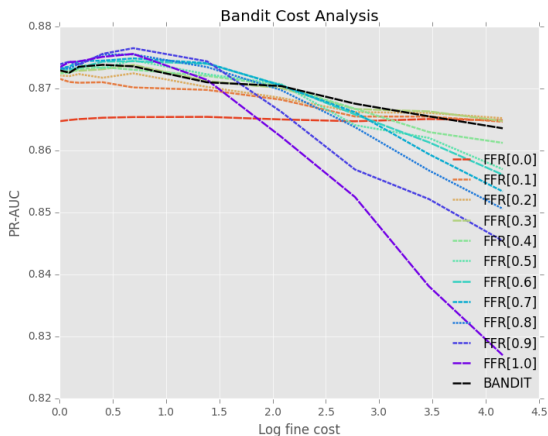


Figure: BANDIT log fine cost analysis with budget fixed.

BANDIT Approach Results

Varying Cost Analysis - Rank and Diff Metrics

Table: Aggregated PR AUC for the protein dataset

	diff				rank			
	min	max	mean	std	min	max	mean	std
algorithm								
BANDIT	0.000	0.003	<u>0.001</u>	0.001	0	8	4.8	2.315
FFR[0.0]	0.000	0.011	0.007	0.004	1	11	8.8	3.429
FFR[0.1]	0.001	0.006	0.003	0.002	3	10	8.0	2.793
FFR[0.2]	0.000	0.004	0.002	0.001	0	9	6.5	3.500
FFR[0.3]	0.000	0.003	0.001	0.001	0	8	5.1	2.663
FFR[0.4]	0.000	0.004	0.002	0.001	1	8	5.6	2.200
FFR[0.5]	0.000	0.008	0.002	0.002	0	8	4.6	2.200
FFR[0.6]	0.000	0.009	0.002	0.003	1	7	4.6	1.855
FFR[0.7]	0.000	0.012	0.002	0.004	0	8	<u>3.3</u>	2.571
FFR[0.8]	0.000	0.015	0.003	0.005	1	9	4.8	3.027
FFR[0.9]	0.000	0.020	0.005	0.007	0	10	4.3	4.605
FFR[1.0]	0.000	0.038	0.009	0.013	1	11	5.6	4.630

BANDIT Approach Results

Varying Budget Analysis - Mixed Cost

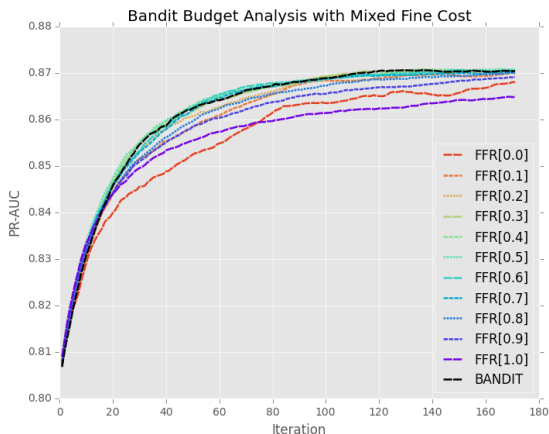


Figure: BANDIT mixed fine cost plot.

BANDIT Approach Results

BANDIT - Rnds 20 to 60

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

**BANDIT
Results**

Conclusions

Bibliography

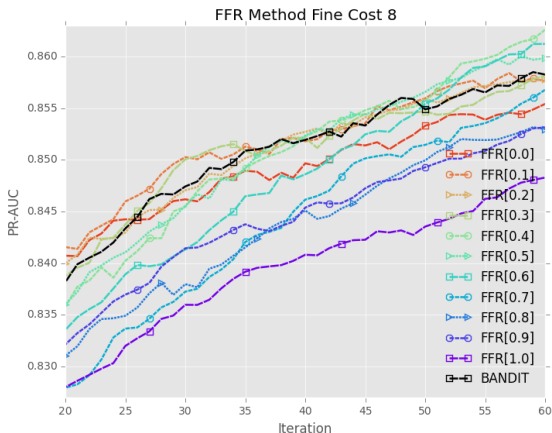


Figure: The fine cost 8 curves shown expanding the rounds 20-60. With the BANDIT approach plotted.

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- 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

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- 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

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography



HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- 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.

- 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

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- 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

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

- 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.

Active vs. Passive Curve Analysis

Logit Accuracy

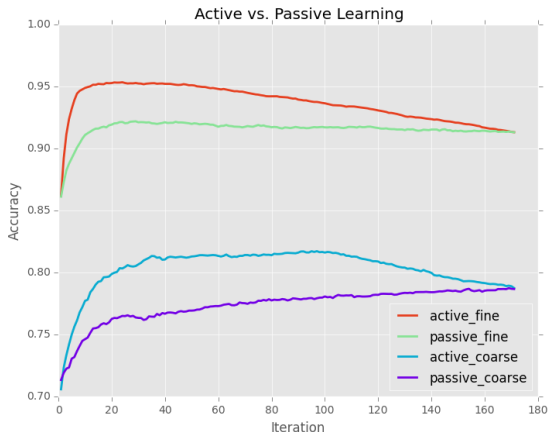


Figure: The accuracy of the classifiers stays at roughly the same rate throughout the rounds; this is due to an effective weighting scheme.

Active vs. Passive Curve Analysis

Logit F-measure

HAL - Protein

James Duin

Introduction

Background

Related Work

Exp. Setup

Conv. ML

Act. vs Pass.

FFR Results

BANDIT
Results

Conclusions

Bibliography

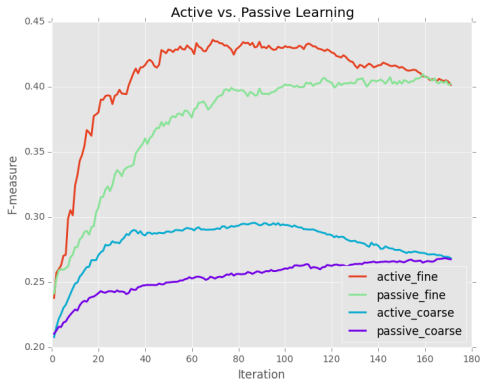


Figure: Both curves show a dominance of fine over coarse and Active over Passive.