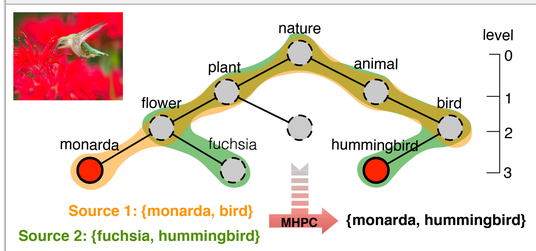
Multi-source Hierarchical Prediction Consolidation

Chenwei Zhang[1] Sihong Xie[2] Yaliang Li[3] Jing Gao[3] Wei Fan[4] Philip S. Yu[1,5]

[1]University of Illinois at Chicago [2]Lehigh University [3]SUNY Buffalo [4]Baidu Research Big Data Lab [5]Tsinghua University

1. Multi-source Hierarchical Prediction Consolidation Problem



Multiple information sources may provide labeling information on the same instance simultaneously.

Consolidation on the output-level: raw features are neglected or withheld due to storage limitation or privacy concerns.

Label hierarchies are prevalently observed. Traditionally, only "flat" labels are considered.

Problem Studied: incorporate the **label hierarchy** into the prediction consolidation process when **only the labeling information** from **multiple information sources** are available.

Challenges: label vagueness, label ambiguity, label sparsity.

2. Modeling Hierarchical Consensus

Minimizing the Consensus Cost

An optimization function to minimize the consensus cost among information sources

$$\min_{\hat{\mathbf{Y}}} \frac{1}{M}\sum_{m=1}^{M}\left\|\hat{\mathbf{Y}}-\mathbf{Y}_m\right\|_F^2 + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{W}_{ij}\left\|\hat{\mathbf{Y}}_{(i)}-\hat{\mathbf{Y}}_{(j)}\right\|_2^2$$

Favors the smoothness of the consolidation result from all the information sources

Regularizes the model to ensure hierarchical label constraints

A closed-form solution:

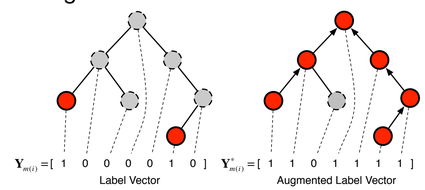
$$\hat{\mathbf{Y}} = \left(1+\lambda(\mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-\frac{1}{2}})\right)^{-1}\frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{Y}}_m = (1+\lambda\cdot\mathbf{L}))^{-1}\bar{\mathbf{Y}}$$

$\mathbf{L}$ regularizes the simple averaging result and guides the $\bar{\mathbf{Y}}$ towards a global consensus $\hat{\mathbf{Y}}$.

Estimating the Hierarchical Similarity

$$\mathbf{W}_{ij} = \exp\left(-\frac{1}{\sigma}\sqrt{\sum_{k=1}^{K}\vec{\mathbf{S}}_k(\hat{\mathbf{Y}}_{(i)}^k-\hat{\mathbf{Y}}_{(j)}^k)^2}\right)$$

Label augmentation based on the label Hierarchy:



The label occurrence in augmented label vectors:

$$\vec{\mathbf{C}}_k = \sum_{m=1}^{M}\sum_{i=1}^{N}\mathbf{Y}^{k*}_{m(i)} \qquad \vec{\mathbf{S}}_k = B\left(\frac{\alpha}{2};\vec{\mathbf{C}}_k, N-\vec{\mathbf{C}}_k+1\right)$$

Uncertainty modeled by the Beta distribution:

$$B\left(\frac{\alpha}{2};\vec{\mathbf{C}}_k, N-\vec{\mathbf{C}}_k+1\right) < \theta_k < B\left(1-\frac{\alpha}{2};\vec{\mathbf{C}}_k, N-\vec{\mathbf{C}}_k+1\right)$$

3. Consolidating Hierarchical Labels

The MHPC Algorithm: a totally **unsupervised**, **iterative** updating algorithm.

Minimizing the Consensus Cost

$$\left(\hat{\mathbf{Y}}\right)_{t_{x+1}} = \left(1+\lambda\cdot(\mathbf{L})_{t_x}\right)^{-1}\left(\hat{\mathbf{Y}}\right)_{t_x}$$

Initialize:

$$\left(\hat{\mathbf{Y}}\right)_{t_0} = \bar{\mathbf{Y}}$$

$$\hat{\mathbf{Y}} = (1+\lambda\cdot\mathbf{L}))^{-1}\bar{\mathbf{Y}}$$

$$\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-\frac{1}{2}}$$

accumulate the consolidation progresses from previous iterations.

Estimating the Hierarchical Similarity

$$\left(\mathbf{W}_{ij}\right)_{t_{x+1}} = \exp\left(-\frac{1}{\sigma}\sqrt{\sum_{k=1}^{K}\vec{\mathbf{S}}_k(\left(\hat{\mathbf{Y}}_{(i)}^k\right)_{t_x}-\left(\hat{\mathbf{Y}}_{(j)}^k\right)_{t_x})^2}\right)$$

Initialize:

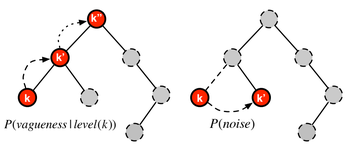
$$\left(\mathbf{W}_{ij}\right)_{t_0} = \exp\left(-\frac{1}{\sigma}\sqrt{\sum_{k=1}^{K}\vec{\mathbf{S}}_k(\bar{\mathbf{Y}}_{(i)}^k-\bar{\mathbf{Y}}_{(j)}^k)^2}\right)$$

The two-phase updating schema terminates when updates on the consolidation result $\left(\hat{\mathbf{Y}}\right)_{t_{x+1}}$ is no longer significant after an iteration.

4. Experiments

Multi-source Yeast Genome Annotation

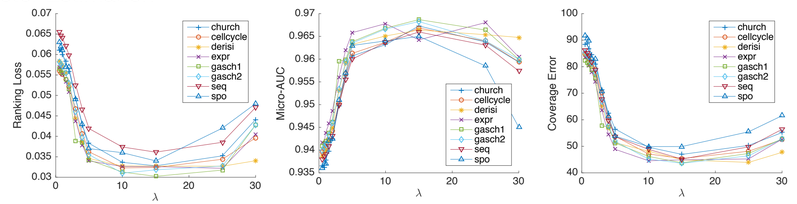
Noisy and ambiguous labels in each annotation:



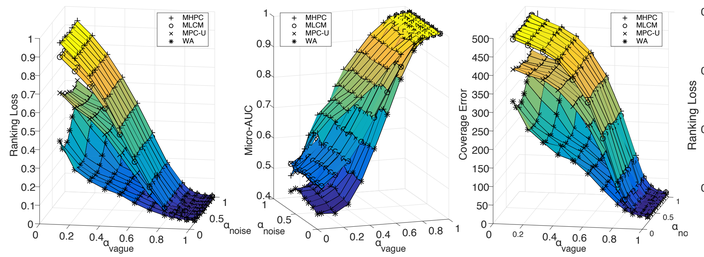
Information Source	Ranking Loss	Micro-AUC	Coverage Error
1	0.171	0.830	180.679
3	0.085	0.920	102.949
4	0.064	0.941	77.910
5	0.054	0.952	64.182
10	0.046	0.964	47.717
15	0.042	0.967	44.019
20	0.042	0.967	44.041
30	0.042	0.967	43.901
50	0.041	0.968	43.592

Table 5: Performance of the MHPC method with label predictions collected from a varying number of information sources.

Parameter Estimation



Sensitivity Analysis

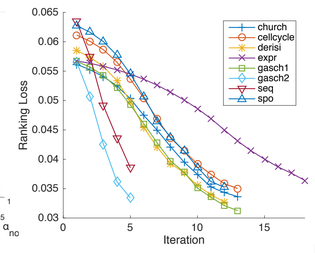


Crowdsourced Online Medical Consultation

Each instance is a set of symptoms that a user describes. For each set of symptoms, disease names are the labels we collected from different doctors.

Instance	Doctor_id	Labels	Ground Truth
sneezing, runny nose, sleepy	52****17	{common cold, allergic rhinitis}	{common cold, allergic rhinitis, rhinitis, sinusitis, antritis}
	46****35	{common cold, rhinitis}	
	53****11	{rhinitis, sinusitis}	

Convergence Analysis



Datasets	Methods	Evaluation Metrics		
		Ranking Loss	Micro-AUC	Coverage Error
MED.1	SA	0.520 (5)	0.620 (5)	213.175 (5)
	WA	0.518 (4)	0.621 (3)	212.844 (4)
	MPC-U	0.304 (3)	0.547 (5)	125.108 (3)
	MLCM	0.262 (2)	0.643 (2)	108.146 (1)
	MHPC	0.196 (1)	0.754 (1)	125.107 (2)
MED.2	SA	0.358 (5)	0.603 (4)	75.544 (5)
	WA	0.357 (4)	0.604 (3)	75.416 (4)
	MPC-U	0.200 (2)	0.539 (5)	42.630 (3)
	MLCM	0.268 (3)	0.680 (2)	31.221 (1)
	MHPC	0.166 (1)	0.715 (1)	35.336 (2)
MED.3	SA	0.067 (4)	0.706 (2)	3.400 (4)
	WA	0.065 (3)	0.704 (3)	3.314 (3)
	MPC-U	0.069 (5)	0.432 (5)	3.543 (5)
	MLCM	0.064 (2)	0.542 (4)	3.288 (2)
	MHPC	0.038 (1)	0.847 (1)	2.000 (1)
MED.4	SA	0.301 (5)	0.601 (4)	38.325 (5)
	WA	0.300 (4)	0.602 (3)	38.275 (4)
	MPC-U	0.173 (2)	0.460 (5)	22.242 (3)
	MLCM	0.225 (3)	0.617 (2)	19.325 (2)
	MHPC	0.118 (1)	0.707 (1)	15.233 (1)
MED.5	SA	0.355 (5)	0.563 (4)	67.534 (5)
	WA	0.353 (4)	0.564 (3)	67.170 (4)
	MPC-U	0.189 (1)	0.549 (5)	36.080 (2)
	MLCM	0.276 (3)	0.565 (2)	52.523 (3)
	MHPC	0.232 (2)	0.572 (1)	32.648 (1)
MED.6	SA	0.365 (5)	0.598 (3)	47.175 (5)
	WA	0.363 (4)	0.599 (2)	46.991 (4)
	MPC-U	0.236 (3)	0.491 (5)	30.632 (3)
	MLCM	0.190 (1)	0.632 (1)	24.798 (2)
	MHPC	0.203 (2)	0.594 (4)	22.633 (1)

Table 1: Performance on medical data sets.

5. Acknowledgements

This work is supported in part by NSF through grants III-1526499. We greatly appreciate **SIGIR** for providing the SIGIR Student Travel Grant.