Multi-source Hierarchical Prediction Consolidation

Chenwei Zhang¹  Sihong Xie² Yaliang Li³ Jing Gao³ Wei Fan⁴ Philip S. Yu¹,⁵
¹University of Illinois at Chicago  ²Lehigh University  ³SUNY Buffalo  ⁴Baidu Research Big Data Lab  ⁵Tsinghua 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

\[
\min \frac{1}{M} \sum_{i=1}^{N} \left| \hat{Y} - Y_i \right|_2^2 + \lambda \sum_{i=1}^{N} W_i \left| \hat{Y}_i - Y_i \right|_2^2
\]

Estimating the Hierarchical Similarity

\[
W_i = \exp \left( -\frac{1}{2} \sqrt{\sum_{k=1}^{K} S_k \left( \hat{Y}_i - Y_i \right)_k^2} \right)
\]

Label augmentation based on the label Hierarchy:

\[
\hat{Y} = \left( 1 + \lambda \cdot L \right) \hat{Y}
\]

The label occurrence in augmented label vectors:

\[
\hat{C}_k = \sum_{i=1}^{N} Y_i^k \quad \hat{S}_k = \hat{B} \left( \frac{\alpha}{2} \hat{C}_k, \hat{N} - \hat{C}_k + 1 \right)
\]

Uncertainty modeled by the Beta distribution:

\[
\hat{B} \left( \alpha \hat{C}_k, \hat{N} - \hat{C}_k + 1 \right) < \theta_i < \hat{B} \left( \frac{\alpha}{2} \hat{C}_k, \hat{N} - \hat{C}_k + 1 \right)
\]

3. Consolidating Hierarchical Labels

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

Minimizing the Consensus Cost

\[
\hat{Y}_{i+1} = \left( 1 + \lambda \cdot L \right) \hat{Y}_i
\]

Estimating the Hierarchical Similarity

\[
W_i = \exp \left( -\frac{1}{2} \sqrt{\sum_{k=1}^{K} S_k \left( \hat{Y}_i - Y_i \right)_k^2} \right)
\]

Initialize:

\[
\hat{Y}_0 = Y
\]

accumulate the consolidation progresses from previous iterations.

The two-phase updating schema terminates when updates on the consolidation result \( \hat{Y} \) is no longer significant after an iteration.

4. Experiments

Multi-source Yeast Genome Annotation

Noisy and ambiguous labels in each annotation:

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.

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.