Optimal Ranking in Multi-label Classification Using Local Precision Rates

Ci-Ren Jiang[1], Chun-Chi Liu[2], Xianghong J. Zhou[3] and Haiyan Huang[4]

[1] *Academia Sinica*
[2] *National Chung Hsing University*
[3] *University of Southern California*
[4] *University of California at Berkeley*

Supplementary Material

Huang et al. (2010) developed a two-stage Bayesian probabilistic approach to performing automated disease diagnosis. In the first stage, the authors derived a Bayesian classifier for each disease concept independently. In the second stage, those individual predictions were integrated in a hierarchical Bayesian network model for collaborative error correction. Here are some details about this method.

S1 Stage I: Build Bayesian Classifiers for Individual UMLS Concepts

Let $Q_{k,x}$ denote the binary label of the query profile x at disease concept k. The Bayesian classifier is the posterior distribution of $Q_{k,x}$ given the similarity scores between x and the profiles in the standardized datasets and the phenotype annotations for the standardized datasets in our disease diagnosis database. Specifically, under the independence assumption among the standardized datasets, the target posterior probability is

$$\Pr(Q_{k,x}|E) \propto \left(\prod_{i=1}^{M} \Pr(\xi_{x,i}|Q_{k,x}, e_{i,k})\right) \Pr(Q_{k,x}|\mathbf{e}), \tag{S1.1}$$

where the "Evidence" E consists of the similarity scores $\{\xi_{x,i}, i = 1, \ldots, M\}$ and the phenotype annotations $\mathbf{e} = \{e_{i,k}, i = 1, \ldots, M\}$. The term M is the number of standardized datasets. Since without the similarity scores $\mathbf{e}$ does not provide helpful information in inferring $Q_{k,x}$, the authors assumed $\Pr(Q_{k,x}|\mathbf{e}) = \Pr(Q_{k,x})$. With this further assumption, $\Pr(Q_{k,x} = 1|E) = 1/(1 + \Lambda)$, where

$$\Lambda = \frac{\Pr(Q_{k,x} = 0)}{\Pr(Q_{k,x} = 1)} \left[\prod_{i=1}^{M} \frac{\Pr(\xi_{x,i}|Q_{k,x} = 0, e_{i,k})}{\Pr(\xi_{x,i}|Q_{k,x} = 1, e_{i,k})}\right].$$

To estimate Λ, the authors introduced a new latent variable corresponding to the phenotype $\mathbf{e}$ to take possible text-mining errors into consideration. They then employed

logistic regression to estimate Λ. Given the estimated Λ, we can infer $Q_{k,x}$ by

$$\hat{Q}_{k,x} = \begin{cases} 1, & \Pr(Q_{k,x}|E) \geq \lambda; \\ 0, & \text{otherwise.} \end{cases}$$

S2 Stage II: A Bayesian Network Model for Collaborative Error Correction

Since in Stage I the predictions $Q_{k,x}$ for different disease concepts are made independently from their own classifiers, the set of $Q_{k,x}$ across all disease concepts might be inconsistent. In order to find the most probable set of consistent $Q_{k,x}$, a Bayesian network model based on the UMLS hierarchy is developed to adjust the predictions of individual disease concepts made in Stage I. Specifically, the target is to find

$$(Q^*_{1,x}, \ldots, Q^*_{K,x}) = \underset{Q_{1,x},\ldots,Q_{K,x}}{\arg\max} \Pr(Q_{1,x}, \ldots, Q_{K,x}|\hat{Q}_{1,x}, \ldots, \hat{Q}_{K,x}), \tag{S2.2}$$

where $\Pr(Q_{1,x}, \ldots, Q_{K,x}|\hat{Q}_{1,x}, \ldots, \hat{Q}_{K,x})$ is characterized by a Bayesian network model. Under the conditional independence assumptions listed below, model (S2.2) can be represented as

$$(Q^*_{1,x}, \ldots, Q^*_{K,x}) = \prod_{j=1}^{K} \Big(\Pr(\hat{Q}_{j,x}|Q_{j,x}) \Pr(Q_{j,x}|\text{children}(Q_{j,x})) \Big) \ .$$

An exact inference algorithm is applied to find the maximum a posteriori.

Conditional Independence Assumptions:

1. The $Q_{k,x}$ nodes are conditioned on their child nodes and independent of all other nodes given children($Q_{k,x}$).
2. The predicted label $\hat{Q}_{k,x}$ is independent of all other classifier outputs $\hat{Q}_{j,x}$ and true labels $Q_{j,x}$ $(j \neq k)$ given true $Q_{k,x}$.

References

Huang, H., Liu, J. C.-C. and Zhou, X. J. (2010). Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc Natl Acad Sci. USA* **107**, 6823–6828.