Fully Associative Ensemble Learning for Hierarchical Multi-Label Classification

Lingfeng Zhang
lzhang34@uh.edu
Shishir K. Shah
sshah@central.uh.edu
Ioannis A. Kakadiaris
ioannisk@uh.edu

Computational Biomedicine Lab
Department of Computer Science
University of Houston
Houston, TX, USA

In Hierarchical Multi-label Classification (HMC), rich hierarchical information is used to improve classification performance. Global approaches learn a single model for the whole class hierarchy [3, 6]. Local approaches introduce hierarchical information to the local prediction results of all the local classifiers to obtain the global prediction results for all the nodes [2, 5].

In this paper, we propose a novel local HMC framework, Fully Associative Ensemble Learning (FAEL). Specifically, a multi-variable regression model is built to minimize the empirical loss between the global predictions of all the training samples and their corresponding true label observations. Let X and Y represent local prediction matrix and label observation matrix, respectively. We define $W = \{w_{ij}\}$ as a weight matrix, where w_{ij} represents the weight of the i^{th} label's local prediction to the j^{th} label's global prediction. In the basic model, the objective function is:

$$\min_W \|Y - XW\|_F^2 + \lambda_1 \|W\|_F^2, \tag{1}$$

where the first term measures the empirical loss of the training set, the second term controls the generalization error, and λ_1 is a regularization parameter. The above function is known as ridge regression. We have:

$$W = \left(X^T X + \lambda_1 I_l\right)^{-1} X^T Y, \tag{2}$$

where I_l represents the $l \times l$ identity matrix.

To capture the complex correlation between global and local prediction, we can generalize the above basic model using the kernel trick. Let Φ represent the map applied to each example's local prediction vector $\mathbf{x}_i$. A kernel function is induced by $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$. By replacing the term X in (1), we obtain:

$$\min_{W_k} \|Y - \Phi W_k\|_F^2 + \lambda_1 \|W_k\|_F^2. \tag{3}$$

After several matrix manipulations [1], the solution of W_k becomes:

$$W_k = \left(\Phi^T \Phi + \lambda_1 I_l\right)^{-1} \Phi^T Y = \Phi^T \left(\Phi \Phi^T + \lambda_1 I_n\right)^{-1} Y, \tag{4}$$

where I_n represents the $n \times n$ identity matrix. For a given testing example s^t and its local prediction $\mathbf{x}^t$, the global prediction $\widehat{\mathbf{y}}^t$ is obtained by $\widehat{\mathbf{y}}^t = \mathbf{x}^t W$. For a kernel version, we obtain:

$$\widehat{\mathbf{y}}_k^t = K(\mathbf{x}^t, \mathbf{x}) \left(K(\mathbf{x}, \mathbf{x}) + \lambda_1 I_n\right)^{-1} Y. \tag{5}$$

To make full use of the hierarchical relationships between different nodes, we introduce a regularization term to the optimization function in (1). Let $\mathcal{R} = \{r_i(c_p, c_q)\}$ denote the binary constraint set of hierarchy $\mathcal{H}$. Each member $r_i(c_p, c_q)$ meets either $c_p =\uparrow c_q$ or $c_p =\Uparrow c_q$, where "$\uparrow$" and "$\Uparrow$" represent the "parent-child" constraint and the "ancestor-descendent" constraint, respectively. We introduce a weight restriction to each pair of nodes in $\mathcal{R}$. Define coefficient $m_{pq} \in \mathbb{R}^+$ for the i^{th} pair $r_i(c_p, c_q)$, so that:

$$w_{pk} = m_{pq} * w_{qk}. \tag{6}$$

For the global prediction of node k, the weight of node p is m_{pq} times the weight of node q. The value of m_{pq} is set by:

$$m_{pq} = \begin{cases} \mu & c_p =\uparrow c_q \\ \mu * (e_{pq} + 1) & c_p =\Uparrow c_q \end{cases}, \tag{7}$$

where μ is a positive constant and e_{pq} represents the number of nodes between p and q. All the restrictions over the hierarchy are summarized as:

$$\sum_{r_i(c_p,c_q)}^{|\mathcal{R}|} \sum_{k=1}^{l} \left(w_{pk} - m_{pq} * w_{qk}\right)^2. \tag{8}$$

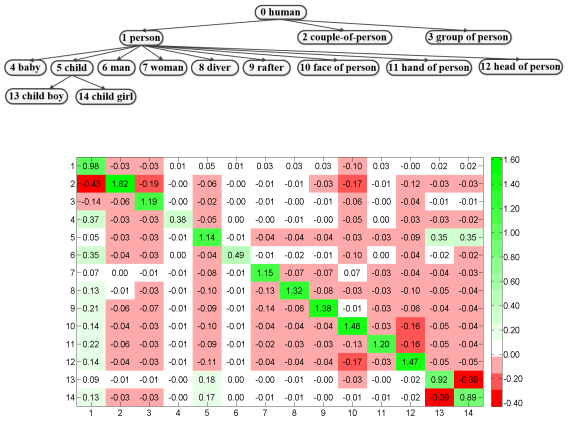

Figure 1: (Top) The "human" sub-hierarchy. (Bottom) The weight matrix W^* learned from B-FAEL.

We introduce a sparse matrix $M = [\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_{|R|}]^T$, in which the i^{th} row $\mathbf{m}_i$ corresponds to the i^{th} pair in $\mathcal{R}$. Each row in M has only two non-zero entries. The p^{th} entry is 1 and the q^{th} entry is $-m_{pq}$, all the other entries are zero. Thus, we obtain the regularization term of the binary constraint model:

$$\sum_{r_i(c_p,c_q)}^{|\mathcal{R}|} \sum_{k=1}^{l} \left(w_{pk} - m_{pq} * w_{qk}\right)^2 = \|MW_b\|_F^2. \tag{9}$$

Adding this term to (1), the optimization function becomes:

$$\min_W \|Y - XW_b\|_F^2 + \lambda_1 \|W_b\|_F^2 + \lambda_2 \|MW_b\|_F^2. \tag{10}$$

The analytical solution of the binary constraint model is given by:

$$W_b = \left(X^T X + \lambda_1 I_l + \lambda_2 M^T M\right)^{-1} X^T Y. \tag{11}$$

Take the "human" sub-hierarchy from the extended IAPR TC-12 image dataset [4] for example, Figure 1 depicts the merits of our model and shows the contribution of hierarchical and sibling nodes on each local prediction. The weight matrix computed indicates that each local node influences its own decision positively while nodes not directly connected in the hierarchy provide a negative influence.

[1] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, Minneapolis, MN, USA, 2007.

[2] Z. Barutcuoglu and C. DeCoro. Hierarchical shape classification using bayesian aggregation. In *Proc. IEEE International Conference on Shape Modeling and Applications*, Matsushima, Japan, 2006.

[3] I. Dimitrovski, D. Kocev, S. Loskovska, and S. Džeroski. Hierarchical annotation of medical images. *Pattern Recognition*, 44(10): 2436–2449, 2011.

[4] H. J. Escalante, C. A. Hernández, J. A. Gonzalez, A. López-López, M. Montes, E. F. Morales, L. E. Sucar, L. Villaseñor, and M. Grubinger. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428, 2010.

[5] G. Valentini. True path rule hierarchical ensembles for genome-wide gene function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):832–847, 2011.

[6] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.