Multilabel Classification through Random Graph Ensemble

Aalto University
School of Science

Hongyu Su, Juho Rousu
{hongyu.su, juho.rousu}@aalto.fi
Helsinki Institute for Information Technology, Aalto University, Finland

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

ABSTRACT

We present new methods for multilabel classification, relying on ensemble learning on a collection of random output graphs imposed on the multilabel and a kernel-based structured output learner as the base classifier. Diversity of base classifiers arises from the different random output structures, a different approach from boosting or bagging. In our experiments, the random graph ensembles are very competitive and robust, ranking first or second on most of the datasets.

ENSEMBLE ANALYSIS

We study the theoretical property of MAM ensemble by analyzing reconstruction error of compatibility score. Compatibility score for a fixed pair $(x, \mathbf{y})$ is

$$\psi(x,\mathbf{y}) = \sum_{e\in E} \psi_e(x,\mathbf{y}_e) = \sum_{j\in V} \psi_j(x,y_j).$$

Denote the $\psi^*(x,\mathbf{y})$ optimal compatibility score. Reconstruction error is given by the squared distance:

$$\Delta^R_{\mathsf{MAM}}(x,\mathbf{y}) = \left(\psi^*(x,\mathbf{y}) - \psi^{\mathsf{MAM}}(x,\mathbf{y})\right)^2$$
$$\Delta^R_I(x,\mathbf{y}) = \frac{1}{T}\sum_t \left(\psi^*(x,\mathbf{y}) - \psi^{(t)}(x,\mathbf{y})\right)^2.$$

THEOREM The reconstruction error of compatibility score distribution given by MAM ensemble $\Delta^R_{\mathsf{MAM}}(x,\mathbf{y})$ is guaranteed to be no greater than the average reconstruction error given by individual base learners $\Delta^R_I(x,\mathbf{y})$. In addition, the gap can be estimated as

$$\Delta^R_I(x,\mathbf{y}) - \Delta^R_{\mathsf{MAM}}(x,\mathbf{y}) = \mathbf{Var}_t\Big(\sum_{j\in V}\Psi_j(x,y_j)\Big) \geq 0.$$

The variance can be further expanded as

$$\mathbf{Var}\Big(\sum_{j\in V}\Psi_j(x,y_j)\Big) = \underbrace{\sum_{j\in V}\mathbf{Var}(\Psi_j(x,y_j))}_{diversity} + \underbrace{\sum_{\substack{p,q\in V,\\ p\neq q}}\mathbf{Cov}(\Psi_p(x,y_p),\Psi_q(x,y_q))}_{coherence}.$$

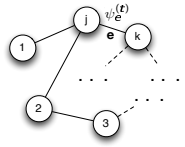
CONCLUSION

We have put forward new methods for multilabel classification, relying on ensemble learning on random output graphs. In our experiments, models thus created have favourable predictive performances on a heterogeneous collection of multilabel datasets. The theoretical analysis of the MAM ensemble highlights the covariance of the compatibility scores between the inputs and microlabels learned by the base learners as the quantity explaining the advantage of the ensemble prediction over the base learners. Our results indicate that structured output prediction methods can be successfully applied to problems where no prior known output structure exists, and thus widen the applicability of the structured output prediction.

MODELS

BASE LEARNER (MMCRF)

Can be seen to decompose into a set of "potential functions" $\Psi_E^{(t)}(x) = (\psi_e^{(t)}(x,\mathbf{u}_e))_{e\in E^{(t)},\mathbf{u}_e\in\mathcal{Y}_e}$



$\{\psi_e^{(t)}(x,++), \psi_e^{(t)}(x,+-), \psi_e^{(t)}(x,-+), \psi_e^{(t)}(x,--)\}$
Prediction is by $\hat{\mathbf{y}}(x) = \mathbf{argmax}_{\mathbf{y}\in\mathcal{Y}}\sum_e \psi_e(x,\mathbf{y}_e)$.

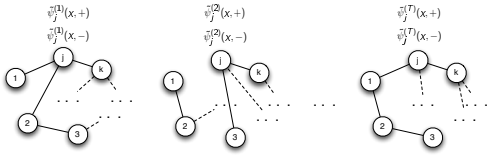
MAJORITY VOTING ENSEMBLE (MVE)

In MVE, the ensemble prediction or each microlabel is the most frequently appearing prediction among the base classifiers

$$F_j^{\mathsf{MVE}}(x) = \mathbf{argmax}_{y_j\in\mathcal{Y}_j}\left(\frac{1}{T}\sum_{i=1}^T \mathbf{1}_{\{F_j^{(t)}(x)=y_j\}}\right),$$

where $F^{(t)}(x) = (F_j^{(t)}(x))_{j=1}^k$ is the predicted multilabel in t'th base classifier.

AVERAGE OF MAX-MARGINALS (AMM)



Our goal is to infer for each microlabel u of each node j its *max-marginal*, that is, the maximum score of a multilabel that is consistent with $y_j = u_j$, $u_j \in \{+,-\}$

$$\tilde{\psi}_j(x,u_j) = \max_{\{\mathbf{y}\in\mathcal{Y}: y_j=u_j\}}\sum_e \psi_e(x,\mathbf{y}_e).$$

The ensemble prediction for each target is obtained by averaging the max-marginals of the base models and choosing the maximizing microlabel for the node:

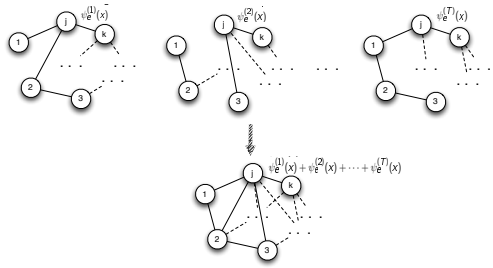
$$F_j^{\mathsf{AMM}}(x) = \mathbf{argmax}_{u_j\in\mathcal{Y}_j}\frac{1}{|T|}\sum_{t=1}^T \tilde{\psi}_{j,u_j}^{(t)}(x),$$

and the predicted multilabel is composed from the predicted microlabels

$$F^{\mathsf{AMM}}(x) = \left(F_j^{\mathsf{AMM}}(x)\right)_{j\in V}.$$

MAXIMUM AVERAGE MARGINALS (MAM)

Generate the **union graph** of the trees underlying the base models, with average edge labeling scores $\frac{1}{|T_e|}\sum_{t\in T(e)}\psi_{e,u}^{(t)}(x)$ (normalized by how many times an edge appears)



Inference on the union graph:

$$F^{\mathsf{MAM}}(x) = \mathbf{argmax}_{\mathbf{y}\in\mathcal{Y}}\sum_{e\in\cup_t E_t}\frac{1}{T}\sum_{t=1}^T \psi_e^{(t)}(x,\mathbf{y}_e)$$

Interpretation: ensemble prediction is the multilabel maximizing the average score over the base models.

EXPERIMENTAL RESULTS

Figure 1: Ensemble learning curve (microlabel accuracy) plotted as the size of ensemble. Average performance of base learner with random tree as output graph structure is denoted as horizontal dash line.

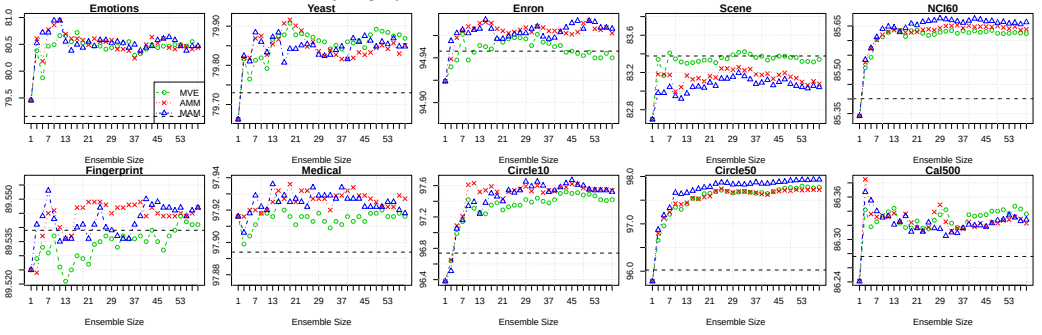


Table 1: Prediction performance by microlabel accuracy.

Dataset	Microlabel Accuracy					
	SVM	Bagging	AdaBoost	MTL	MMCRF	MAM
Emotions	77.3±1.9	74.1±1.8	76.8±1.6	*79.8±1.8*	79.2±0.9	**80.5±1.4**
Yeast	**80.0±0.6**	78.4±0.7	74.8±0.3	79.3±0.2	79.7±0.3	*79.9±0.4*
Scene	**90.2±0.3**	87.8±0.8	84.3±0.4	*88.4±0.6*	83.4±0.2	83.0±0.2
Enron	93.6±0.2	93.7±0.1	86.2±0.2	93.5±0.1	*94.9±0.1*	**95.0±0.2**
Cal500	**86.3±0.3**	86.0±0.2	74.9±0.4	86.2±0.2	**86.3±0.2**	**86.3±0.3**
FP	**89.7±0.2**	85.0±0.7	84.1±0.5	82.7±0.3	*89.5±0.3*	*89.5±0.8*
NCI60	84.7±0.7	79.5±0.8	79.3±1.0	84.0±1.1	*85.4±0.9*	**85.7±0.7**
Medical	97.4±0.1	97.4±0.1	91.4±0.3	97.4±0.1	**97.9±0.1**	**97.9±0.1**
Circle10	94.8±0.9	92.9±0.9	**98.0±0.4**	93.7±1.4	96.7±0.7	*97.5±0.3*
Circle50	94.1±0.3	91.7±0.3	*96.6±0.2*	93.8±0.7	96.0±0.1	**97.9±0.2**
@Top2	4	0	2	2	*5*	**9**

Table 2: Prediction performance by multilabel accuracy.

Dataset	Multilabel Accuracy					
	SVM	Bagging	AdaBoost	MTL	MMCRF	MAM
Emotions	21.2±3.4	20.9±2.6	23.8±2.3	25.5±3.5	*26.5±3.1*	**30.4±4.2**
Yeast	*14.0±1.8*	13.1±1.2	7.5±1.3	11.3±2.8	13.8±1.5	**14.0±0.6**
Scene	**52.8±1.0**	*46.5±2.5*	34.7±1.8	44.8±3.0	12.6±0.7	5.4±0.5
Enron	0.4±0.1	0.1±0.2	0.0±0.0	0.4±0.3	*11.7±1.2*	**12.1±1.0**
Cal500	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
FP	**1.0±1.0**	0.0±0.0	0.0±0.0	0.0±0.0	*0.4±0.9*	*0.4±0.5*
NCI60	*43.1±1.3*	21.1±1.3	2.5±0.6	**47.0±1.4**	36.9±0.8	40.0±1.0
Medical	8.2±2.3	8.2±1.6	5.1±1.0	8.2±1.2	*35.9±2.1*	**36.9±4.6**
Circle10	69.1±4.0	64.8±3.2	**86.0±2.0**	66.8±3.4	75.2±5.6	*82.3±2.2*
Circle50	29.7±2.5	21.7±2.6	28.9±3.6	27.7±3.4	*30.8±1.9*	**53.8±2.2**
@Top2	5	2	2	2	*6*	**8**

ACKNOWLEDGEMENTS

The work was financially supported by Helsinki Doctoral Programme in Computer Science (Hecse), Academy of Finland grant 118653 (ALGODAN), and in part by the IST Programme of the European Community under the PASCAL2 Network of Excellence, ICT-2007-216886. This work only reflects the authors' views.