Improved Multi-Label Propagation for Small Data with Multi-Objective Optimization: Appendix

K.Musayeva and M.Binois

A Evaluation Measures

In what follows, we denote $h = (h_1, \ldots, h_C) : \mathcal{X} \to \mathcal{Y}$, and $h^s = (h_1^s, \ldots, h_C^s) : \mathcal{X} \to \mathbb{R}^C$.

The subset accuracy is the generalization of the traditional indicator loss function to the multi-label classification setting, and as such is the strictest evaluation measure since it penalizes the output of h if it does not exactly match the true labels:

$$l_{SA}(\mathbf{Y}, h(X)) = \mathbb{1}_{\{h(X) \neq \mathbf{Y}\}}.$$

Unlike the subset accuracy, the Hamming loss penalizes the misclassifications for each label independently:

$$l_{HL}(\mathbf{Y}, h(X)) = \frac{1}{C} \sum_{j=1}^{C} \mathbb{1}_{\{h_j(X) \neq Y_j\}}.$$

The AUC measure evaluates the capacity of multi-label classifier for each instance to score higher the relevant labels than the irrelevant ones:

$$l_{AUC}(\mathbf{Y}, h^s(X)) = \frac{S}{l^r l^{irr}},$$

where S is the number of pairs (r, irr) of all relevant and irrelevant labels for which $h_r^s(X) \geqslant h_{irr}^s(X)$, and l^r and l^{irr} are the number of relevant and irrelevant labels for X.

Primarily used in information retrieval problems, the $F1$-measure is the harmonic mean of precision, $\frac{\sum_{j=1}^{C} h_j(X) Y_j}{\sum_{j=1}^{C} h_j(X)}$, and recall, $\frac{\sum_{j=1}^{C} h_j(X) Y_j}{\sum_{j=1}^{C} Y_j(X)}$:

$$l_{F1}(\mathbf{Y}, h(X)) = \frac{2 \sum_{j=1}^{C} h_j(X) Y_j}{\sum_{j=1}^{C} Y_j + \sum_{j=1}^{C} h_j(X)}.$$

Due to the multi-dimensionality of outputs, on an n-sample, the $F1$-measure and AUC can be averaged differently with respect to labels and instances. These are called *macro* and *micro* averaging.

Macro-$F1$ is primarily used in class-imbalance setting to evaluate the performance of classifier on rare labels:

$$\textit{macro-}F1((\mathbf{Y}_i)_{i=1}^n, h(X_i)_{i=1}^n) = \frac{1}{C} \sum_{j=1}^{C} \frac{2 \sum_{i=1}^{n} h_j(X_i) Y_{ij}}{\sum_{i=1}^{n} Y_{ij} + \sum_{i=1}^{n} h_j(X_i)},$$

where Y_{ij} denotes the j-th element of the vector $\mathbf{Y}_i$. Micro-$F1$ measure, on the other hand, is not sensitive to rare labels and is defined as

$$\textit{micro-F1}((\mathbf{Y}_i)_{i=1}^n, h(X_i)_{i=1}^n) = \frac{2\sum_{i=1}^n \sum_{j=1}^C h_j(X_i) Y_{ij}}{\sum_{i=1}^n \sum_{j=1}^C Y_{ij} + \sum_{i=1}^n \sum_{j=1}^C h_j(X_i)}.$$

Macro- and micro-AUC measures are defined as follows:

$$\textit{macro-AUC}((\mathbf{Y}_i)_{i=1}^n, h^s(X_i)_{i=1}^n) = \frac{1}{C}\sum_{j=1}^C \frac{S^j}{n_j^r n_j^{irr}},$$

where n_j^r and n_j^{irr} denote the number of relevant and irrelevant instances for the given label j, and S^j is the number of all pairs (X_r, X_{irr}) of relevant and irrelevant instances for which $h_j^s(X_r) \geqslant h_j^s(X_{irr})$, and

$$\textit{micro-AUC}((\mathbf{Y}_i)_{i=1}^n, h^s(X_i)_{i=1}^n) = \frac{S}{n^r n^{irr}},$$

where n^r and n^{irr} denote the total number of relevant and irrelevant labels for all instances, and S is the number of all quadruples $(X_r(i), X_{irr}(j), i, j)$ for which $h_i^s(X_r(i)) \geqslant h_j^s(X_{irr}(j))$.

The average precision is the average fraction of labels ranked above a given label k in the set R of relevant labels:

$$l_{ap}((\mathbf{Y}_i)_{i=1}^n, h^s(X_i)_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n \frac{1}{|R|}\sum_{k\in R} \frac{|\{k' \in R : h_{k'}^s(X_i) \leqslant h_k^s(X_i)\}|}{h_k^s(X_i)},$$

where $|S|$ stands the cardinality of the set S.

B Decision Function/Thresholding Strategy

Assume h^s is defined as in the previous section. For any unlabelled point $\mathbf{x}_i$ and any $j \in \{1, 2, \ldots, C\}$, $z_j^i = h_j^s(\mathbf{x}_i)$ can be mapped to an element in $\mathcal{Y}$ using a decision function

$$dr_{t_j}(z_j^i) = \begin{cases} 1 & \text{if } z_j^i \geqslant t_j \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $t_j \geqslant 0$ is a threshold. In class-mass normalization method, t_j is computed as

$$t_j = \frac{1-p_j}{p_j} \cdot \frac{\sum_{p=l+1}^n z_j^p}{\sum_{p=l+1}^n (1-z_j^p)} \cdot (1 - z_j^i), \tag{9}$$

where $p_j = \sum_{i=1}^l y_{ij}/l$ is the fraction of class 1 for the label j. In [25], a single threshold is used for all labels which is found by minimizing the following difference:

$$t^* = \underset{t\in(0,1)}{\operatorname{argmin}} \left| \operatorname{lc}(D_l) - \frac{1}{n-l}\sum_{i=l+1}^n \sum_{j=1}^C \mathbb{1}_{\{z_j^i \geqslant t\}} \right|, \tag{10}$$

where the quantity

$$\mathrm{lc}(D_l) = \frac{1}{l}\sum_{i=1}^{l}\sum_{j=1}^{C} y_{ij} \tag{11}$$

is the label cardinality. Then, in (8), $t_j = t^*$ for all j.

C Comparison of CMN and LCO Thresholding Approaches for HF

Table 1: The KS solution for HF. The up/down arrow next to the name of an evaluation metric means the higher/lower the value the better the performance. Only the measures that requires a thresholding strategy are reported.

	Emotions		Fungi	
Measures	CMN	LCO	CMN	LCO
Hamming loss ↓	0.1856 ± 0.0049	0.1891 ± 0.0085	0.2049 ± 0.0078	0.2062 ± 0.0062
Subset accuracy ↑	0.3167 ± 0.0075	0.3305 ± 0.0151	0.1506 ± 0.0190	0.1423 ± 0.0163
F1 ↑	0.6475 ± 0.0100	0.6730 ± 0.0133	0.6662 ± 0.0149	0.6772 ± 0.0129
Macro-F1 ↑	0.6511 ± 0.0111	0.6792 ± 0.0146	0.5521 ± 0.0193	0.5789 ± 0.0239
Micro-F1 ↑	0.6792 ± 0.0099	0.6975 ± 0.0141	0.6727 ± 0.0117	0.6880 ± 0.0090

	Scene		Yeast	
Measures	CMN	LCO	CMN	LCO
Hamming loss ↓	0.0926 ± 0.0042	0.0986 ± 0.0045	0.1965 ± 0.0052	0.2024 ± 0.0043
Subset accuracy ↑	0.6892 ± 0.0132	0.6335 ± 0.0146	0.2056 ± 0.0094	0.2046 ± 0.0123
F1 ↑	0.7369 ± 0.0122	0.7409 ± 0.0122	0.6118 ± 0.0098	0.6505 ± 0.0067
Macro-F1 ↑	0.7448 ± 0.0114	0.7482 ± 0.0099	0.3967 ± 0.0138	0.4315 ± 0.0102
Micro-F1 ↑	0.7325 ± 0.0119	0.7328 ± 0.0114	0.6404 ± 0.0095	0.6672 ± 0.0066

D Comparison with Exact-F1-Plug-in Classifier

Table 2: Comparison of EFC with the KS solution of HF and Stacking HF (SHF). SHF uses the predictions of RAKEL method. EFC, being sensitive to class imbalance, performed poorly on Fungi dataset, and thus the latter is excluded from this report.

Emotions

Measures	EFC	HF	SHF
Hamming loss ↓	0.2345 ± 0.0038	0.1886 ± 0.0082	0.1846 ± 0.0065
F1 ↑	0.6754 ± 0.0070	0.6740 ± 0.0120	0.6784 ± 0.0095

Scene

Measures	EFC	HF	SHF
Hamming loss ↓	0.1163 ± 0.0031	0.0986 ± 0.0045	0.0808 ± 0.0029
F1 ↑	0.7497 ± 0.0052	0.7409 ± 0.0122	0.7817 ± 0.0089

Yeast

Measures	EFC	HF	SHF
Hamming loss ↓	0.2334 ± 0.0037	0.2016 ± 0.0041	0.1982 ± 0.0034
F1 ↑	0.6457 ± 0.0065	0.6519 ± 0.0073	0.6587 ± 0.0058

E Standard Deviations for Tables 3-6.

Table 3: Standard deviations for Emotions dataset.

A) HF and LP/ML-KNN

Measures	HF	CM	LNP	TRAM	DLP	ML-KNN
Hamming loss	0.0085	0.0090	0.0076	0.0109	0.0060	0.0088
Subset-accuracy	0.0151	0.0250	0.0164	0.0176	0.0174	0.0258
F1	0.0133	0.0160	0.0141	0.0151	0.0082	0.0124
Macro-F1	0.0146	0.0120	0.0148	0.0167	0.0109	0.0151
Micro-F1	0.0141	0.0120	0.0142	0.0163	0.0089	0.0131
Macro-AUC	0.0087	0.0090	0.0088	0.0103	0.0119	0.0110
Micro-AUC	0.0081	0.0100	0.0084	0.0106	0.0113	0.0108
Average precision	0.0056	0.0090	0.0097	0.0097	0.0042	0.0077

B) Stacking: HF vs IBLR

Measures	BR	ECC	RAKEL	RAKEL+HF	IBLR	RAKEL+IBLR
Hamming loss	0.0077	0.0068	0.0069	0.0094	0.0082	0.0059
Subset accuracy	0.0219	0.0196	0.0138	0.0159	0.0165	0.0193
F1	0.0144	0.0135	0.0130	0.0144	0.0134	0.0098
Macro-F1	0.0147	0.0141	0.0119	0.0158	0.0155	0.0100
Micro-F1	0.0123	0.0126	0.0118	0.0155	0.0147	0.0098
Macro-AUC	0.0085	0.0111	0.0078	0.0083	0.0107	0.0093
Micro-AUC	0.0082	0.0097	0.0071	0.0091	0.0097	0.0082
Average precision	0.0133	0.0124	0.0075	0.0087	0.0099	0.0081

Table 4: Standard deviations for Fungi dataset.

A) HF and LP/ML-KNN

Measures	HF	CM	LNP	TRAM	DLP	ML-KNN
Hamming loss	0.0090	0.0082	0.0081	0.0087	0.0090	0.0122
Subset-accuracy	0.0203	0.0234	0.0153	0.0132	0.0213	0.0257
F1	0.0159	0.0150	0.0148	0.0105	0.0154	0.0255
Macro-F1	0.0141	0.0285	0.0226	0.0107	0.0244	0.0340
Micro-F1	0.0133	0.0170	0.0143	0.0107	0.0126	0.0181
Macro-AUC	0.0098	0.0162	0.0134	0.0118	0.0162	0.0098
Micro-AUC	0.0056	0.0096	0.0104	0.0063	0.0072	0.0078
Average precision	0.0101	0.0104	0.0090	0.0079	0.0084	0.0197

B) Stacking: HF vs IBLR

Measures	BR	ECC	RAKEL	ECC+HF	HF+HF	IBLR	ECC+IBLR
Hamming loss	0.0095	0.0102	0.0081	0.0101	0.0107	0.0077	0.0115
Subset-accuracy	0.0220	0.0264	0.0251	0.0263	0.0178	0.0166	0.0192
F1	0.0187	0.0162	0.0202	0.0194	0.0149	0.0149	0.0182
Macro-F1	0.0408	0.0240	0.0191	0.0220	0.0141	0.0165	0.0244
Micro-F1	0.0174	0.0177	0.0162	0.0164	0.0133	0.0156	0.0172
Macro-AUC	0.0179	0.0142	0.0130	0.0163	0.0140	0.0106	0.0102
Micro-AUC	0.0089	0.0092	0.0131	0.0132	0.0067	0.0060	0.0077
Average precision	0.0124	0.0167	0.0206	0.0101	0.0116	0.0178	0.0095

Table 5: Standard deviations for Scene dataset.

A) ILP and LP/ML-KNN

Measures	ILP	CM	LNP	TRAM	DLP	ML-KNN
Hamming loss	0.0030	0.0022	0.0042	0.0054	0.0032	0.0030
Subset-accuracy	0.0104	0.0074	0.0136	0.0159	0.0109	0.0146
F1	0.0094	0.0070	0.0117	0.0160	0.0100	0.0093
Macro-F1	0.0087	0.0065	0.0115	0.0155	0.0093	0.0083
Micro-F1	0.0077	0.0061	0.0101	0.0147	0.0084	0.0079
Macro-AUC	0.0026	0.0041	0.0066	0.0036	0.0037	0.0058
Micro-AUC	0.0036	0.0055	0.0066	0.0050	0.0044	0.0051
Average precision	0.0056	0.0050	0.0084	0.0096	0.0046	0.0066

B) Stacking: ILP vs IBLR

Measures	BR	ECC	RAKEL	RAKEL+ILP	IBLR	RAKEL+IBLR
Hamming loss	0.0051	0.0041	0.0029	0.0041	0.0038	0.0037
Subset-accuracy	0.0175	0.0163	0.0084	0.0110	0.0152	0.0112
F1	0.0142	0.0124	0.0082	0,0126	0.0119	0.0103
Macro-F1	0.0139	0.0118	0.0080	0.0121	0.0104	0.0105
Micro-F1	0.0139	0.0119	0.0074	0.0120	0.0096	0.0101
Macro-AUC	0.0044	0.0037	0.0054	0.0067	0.0033	0.0042
Micro-AUC	0.0044	0.0036	0.0056	0.0057	0.0035	0.0039
Average precision	0.0066	0.0064	0.0069	0.0079	0.0069	0.0069

Table 6: Standard deviations for Yeast dataset.

A) HF and LP/ML-KNN

Measures	HF	CM	LNP	TRAM	DLP	ML-KNN
Hamming loss	0.0046	0.0037	0.0060	0.0047	0.0031	0.0046
Subset-accuracy	0.0097	0.0065	0.0054	0.0096	0.0127	0.0181
F1	0.0063	0.0108	0.0094	0.0082	0.0048	0.0080
Macro-F1	0.0084	0.0125	0.0097	0.0114	0.0067	0.0126
Micro-F1	0.0065	0.0084	0.0093	0.0081	0.0042	0.0074
Macro-AUC	0.0110	0.0091	0.0057	0.0120	0.0087	0.0054
Micro-AUC	0.0037	0.0049	0.0060	0.0043	0.0030	0.0029
iAverage precision	0.0061	0.0055	0.0109	0.0070	0.0038	0.0061

B) Stacking: HF vs IBLR

Measures	BR	ECC	RAKEL	RAKEL+HF	IBLR	RAKEL+IBLR
Hamming loss	0.0031	0.0047	0.0038	0.0041	0.0037	0.0031
Subset-accuracy	0.0096	0.0111	0.0146	0.0084	0.0146	0.0085
F1	0.0059	0.0091	0.0052	0.0070	0.0101	0.0083
Macro-F1	0.0084	0.0100	0.0094	0.0068	0.0145	0.0117
Micro-F1	0.0055	0.0080	0.0057	0.0065	0.0084	0.0073
Macro-AUC	0.0080	0.0073	0.0054	0.0095	0.0065	0.0092
Micro-AUC	0.0031	0.0053	0.0050	0.0030	0.0028	0.0033
Average precision	0.0055	0.0063	0.0080	0.0046	0.0078	0.0067

F Regularized Label Propagation

The regularized approach is based on the propagation matrix of the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{ll} & \mathbf{P}_{lu} \\ \mathbf{P}_{ul} & \mathbf{P}_{uu} \end{bmatrix}.$$

Let us now suppose that Λ in (1) is the same value λ for both the labelled and unlabelled points. Then, taking the derivative of (1) with respect to $\mathbf{f}$ we obtain $\mathbf{f} - \mathbf{P}\mathbf{f} + \lambda(\mathbf{f} - \mathbf{y}^{(n)}) = 0$, the solution of which is equivalent to $\mathbf{f} = \left(\mathbf{I} - \frac{1}{1+\lambda}\mathbf{P}\right)^{-1}\mathbf{y}^{(n)}$. Now, let $\mathbf{A} = \lambda'\mathbf{P}$ with $\lambda' = \frac{1}{1+\lambda}$. Since $\mathbf{f}$ can be expressed as $\mathbf{f} = \sum_{i=0}^{\infty}\mathbf{A}^{(i)}\mathbf{y}^{(n)}$, taking into account that the labels of the unlabelled points are initialized to zero, it follows that

$$\mathbf{f}_u = \mathbf{P}_{ul}(\lambda' \cdot \mathbf{y}_l^{(n)}) + (\mathbf{P}_{ul}\mathbf{P}_{ll} + \mathbf{P}_{uu}\mathbf{P}_{ul})((\lambda')^2 \cdot \mathbf{y}_l^{(n)}) + \dots. \tag{12}$$

Since $\lambda' \in (0, 1)$, the more the labels travel the less their contributions become.

G TRAM Method (Kong et al., 2011) is Harmonic Function

The optimization problem of TRAM method [17] is as follows:

$$\min_{\mathbf{f}} \sum_{i=l+1}^{n}\sum_{k=1}^{C}\left(f_k(i) - \sum_{j=1}^{n} P_{ij}f_k(j)\right)^2$$

where P_{ij} corresponds to the (i, j)-th entry in the transition matrix $\mathbf{P}$ (2), with the constraint that $f_k(i) = y_{ik}$. The solution follows by taking the derivative with respect to $\mathbf{f}$: $f_k(i) = \sum_{j=1}^{n} P_{ij}f_k(j)$ which in the matrix form gives the harmonic function (3). The fact that $f_k(i) \in [0, 1]$, $i \in \{l+1, \dots, n\}$ follows from the discrete maximum principle.