Supplementary Material - An Efficient Multi-Class Selective Sampling Algorithm on Graphs

Peng Yang[‡] and **Peilin Zhao**[‡] and **Zhen Hai** [‡] and **Wei Liu**[†] and **Xiao-Li Li**[‡] and **Steven C.H. Hoi** [♯]
[‡] Institute for Infocomm Research, Singapore, 138632. Email: {yangp,zhaop,haiz,xlli}@i2r.a-star.edu.sg
[†] Didi Research, Beijing China, 100085. Email: wliu@ee.columbia.edu
[♯] Singapore Management University, Singapore, 178902. Email: chhoi@smu.edu.sg

I. Supplementary

□

The proof of Lemma 1

Proof. From

$$G(W)=\sum_{t=1}^{T}\|W^\top\mathbf{m}_t-\mathbf{y}_t\|_2^2+\gamma tr\left(W^\top W\right)$$

$$\begin{aligned}
=&\sum_{t=1}^{T}\left[\|\mathbf{y}_t\|_2^2-2tr\left(W^\top\mathbf{m}_t\mathbf{y}_t^\top\right)+tr\left(W^\top\mathbf{m}_t\mathbf{m}_t^\top W\right)\right]\\
&+\gamma tr\left(W^\top W\right)\\
=&tr\left(W^\top\left(\gamma I+\sum_{t=1}^{T}\mathbf{m}_t\mathbf{m}_t^\top\right)W\right)+\sum_{t=1}^{T}\|\mathbf{y}_t\|^2\\
&-2tr\left(W^\top\left(\sum_{t=1}^{T}\mathbf{m}_t\mathbf{y}_t^\top\right)\right)\\
\overset{(6)}{=}&\sum_{t=1}^{T}\|\mathbf{y}_t\|^2-2tr\left(W^\top B_T\right)+tr\left(W^\top A_T W\right),
\end{aligned}$$

it follows that $\nabla G(W)=2A_TW-2B_T$, $\nabla^2G(W)=2A_T$. Thus $G(W)$ is convex and it is minimal if $\nabla G(W)=A_TW-B_T=0$ with $W=A_T^{-1}B_T$. This shows that with $W_T=A_T^{-1}B_T$, we obtain

$$G(W_T)=G\left(A_T^{-1}B_T\right)=\sum_{t=1}^{T}\|\mathbf{y}_t\|^2-tr\left(B_T^\top A_T^{-1}B_T\right).$$

□

Corollary 1. *Given that $A_t=\gamma I+\sum_t\mathbf{m}_t\mathbf{m}_t^\top$, for all $t\geq 1$:*

$$\begin{aligned}
&A_{t-1}^{-1}-A_t^{-1}-A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}\\
=&(\mathbf{m}_t^\top A_{t-1}^{-1}\mathbf{m}_t)A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}.
\end{aligned}\tag{1}$$

Proof. From the quality $A_t-A_{t-1}=\mathbf{m}_t\mathbf{m}_t^\top$,

$$A_{t-1}^{-1}\left(A_t-A_{t-1}\right)A_t^{-1}=A_{t-1}^{-1}\left(\mathbf{m}_t\mathbf{m}_t^\top\right)A_t^{-1},$$

we get $A_{t-1}^{-1}-A_t^{-1}=A_{t-1}^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}$. Similar, $A_{t-1}^{-1}-A_t^{-1}=A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_{t-1}^{-1}$. Thus,

$$\begin{aligned}
&A_{t-1}^{-1}-A_t^{-1}-A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}\\
=&\left(A_{t-1}^{-1}-A_t^{-1}\right)\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}=\left(\mathbf{m}_t^\top A_{t-1}^{-1}\mathbf{m}_t\right)A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}.
\end{aligned}$$

□

The proof of Lemma 2

Proof. Given that

$$\begin{aligned}
&\|\mathbf{y}_t-\mathbf{f}_t\|_2^2+\inf_U G_{t-1}(U)-\inf_U G_t(U)\\
\overset{(7)}{=}&-2\mathbf{y}_t\cdot\mathbf{f}_t+\|\mathbf{f}_t\|_2^2-tr\left(B_{t-1}^\top A_{t-1}^{-1}B_{t-1}\right)+tr\left(B_t^\top A_t^{-1}B_t\right)\\
\overset{(6)}{=}&tr\left(B_{t-1}^\top A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}B_{t-1}\right)-tr\left(B_{t-1}^\top A_{t-1}^{-1}B_{t-1}\right)\\
&+tr\left(\left(B_{t-1}+\mathbf{m}_t\mathbf{y}_t^\top\right)^\top A_t^{-1}\left(B_{t-1}+\mathbf{m}_t\mathbf{y}_t^\top\right)\right)\\
&-2\mathbf{y}_t\cdot B_{t-1}^\top A_t^{-1}\mathbf{m}_t\\
=&tr\left(B_{t-1}^\top\left(A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}-A_{t-1}^{-1}+A_t^{-1}\right)B_{t-1}\right)\\
&+tr\left(\mathbf{y}_t\mathbf{m}_t^\top A_t^{-1}\mathbf{m}_t\mathbf{y}_t^\top\right)\\
\overset{(1)}{=}&-tr\left(B_{t-1}^\top(\mathbf{m}_t^\top A_{t-1}^{-1}\mathbf{m}_t)A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}B_{t-1}\right)\\
&+tr\left(\frac{r_t}{1+r_t}\mathbf{y}_t\mathbf{y}_t^\top\right)\\
\overset{(7)}{=}&\frac{r_t}{1+r_t}tr\left(\mathbf{y}_t\mathbf{y}_t^\top\right)-r_t tr\left(B_{t-1}^\top A_t^{-1}\mathbf{m}_t\mathbf{m}_t^\top A_t^{-1}B_{t-1}\right)\\
=&\frac{2r_t}{1+r_t}-r_t\|\mathbf{f}_t\|_2^2
\end{aligned}$$

□

Online Learning in Binary and Multi-class Setting

Binary-class setting reduces a multi-class problem into many binary-class sub-problems (i.e., given a dataset of N classes, N binary-class problems are generated via 1-vs-rest schema: for each binary-class problem, we assign the label of one class samples with $+1$ and other $N-1$ class samples with -1), while CMOG in this paper is directly a multi-class setting. We compare the two settings in terms of loss function, margin, model-update and performance evaluation.

Generally, a binary classification model is to differentiate binary-class samples with label "± 1". And binary classifier (f) predicts the sample label with a boundary 0, that is, given a sample x, if predicted value $f(x)>0$, x is predicted to class $+1$; If $f(x) < 0$, it is predicted to class -1. We update a binary classifier based on a hinge loss function in binary setting, $L(x) = [1 - y \cdot f(x)]_+$, where $y \in \{\pm 1\}$, $[.]_+ = \max\{., 0\}$. In addition, we define absolute value $|f(x)|$ as "margin": the higher the "margin" (distance to boundary 0), the more confident the predicted result is. However, binary classifier under 1-vs-rest schema can only answer whether a sample belongs to one class or not. Given three classes (a, b, c) where we set class "a" as label $+1$ and the other two classes "b" and "c" as -1, then the model $f(x)$ trained on above binary labels can tell whether x belongs to "a" or not. If not "a", the model cannot identify whether x belongs to "b" or "c". To address the above issue, we present a multi-class setting, where we give each class a linear model, i.e., $f_a(x)$, $f_b(x)$ and $f_c(x)$. We predict the label of x via $\arg\max_{i \in \{a,b,c\}} f_i(x)$. And the loss function for multi-class setting is $L(x) = \big[1 - \big(f_y(x) - \max_{i \in \{a,b,c\}/\{y\}} f_i(x)\big)\big]_+$, where y is the true class of x and $\max_{i \in \{a,b,c\}/\{y\}} f_i(x)$ is the highest score among "wrong" classes, e.g., if x belongs to "a", then $L(x) = [1 - (f_a(x) - \max\{f_b(x), f_c(x)\})]_+$. Different from "margin" ($|f(x)|$) in binary setting, "margin" in multi-class schema is $f_{\hat{y}_t}(x) - f_{y_t^{''}}(x)$, defined as δ in def 2. In addition, when updating the model, multi-class model can update two linear models simultaneously at each around, since the **y** is a vector with true class coordinate y_t to +1 and a wrong class with the highest score $y_t^{''}$ to -1. In binary setting, the y is only a binary variable (i.e., ± 1), thus only one linear model is updated.

For the evaluation metrics, although the cumulative error rate and number of queried labels are applied into both binary-class and multi-class setting, the two groups of results are unable to be compared. Given a dataset with N classes, the binary-setting (1-vs-rest schema) generates a set of N independent binary-class classifiers while each classifier is built on all data samples. After running N times of experiments independently to evaluate the N binary-classifiers, the error rate and queried number are averaged over the outputs of N experiments. Note that the average result can only tell the model effectiveness in binary classification. On the other hand, in the multiple-class setting, we run only 1 time of experiment to train the multiple class models simultaneously. This performance can tell the learner accuracy for multi-class classification. Due to the different experimental setting, a few binary-setting algorithms unable to be adapted into multi-class setting would not be included in baselines of this work, while the GPA adapted into multi-class setting would achieve a different result from binary-setting and the two groups of results are incomparable.