Self-Paced Co-training

Fan Ma [1] **Deyu Meng** [* 1] **Qi Xie** [1] **Zina Li** [1] **Xuanyi Dong** [2]

In this supplementary material, we present the condition of ϵ-expanding with respect to the proposed serial co-training process, and give the proof that SPaCo is an efficient PAC learning algorithm if such condition is satisfied.

Notation and Definition: We assume that examples are drawn from some distributions D over an instance space $X = X_1 \times X_2$, where X_1 and X_2 correspond to two different "views" of examples. Let c denote the target function, and let X^+ and X^- (for simplicity we assume we are doing binary classification) denote the positive and negative regions of X, respectively . For $i \in 1, 2$, let $X_i^+ = \{x_j \in X_i : c_i(x_j) = 1\}$, so we can think of X^+ as $X_1^+ \times X_2^+$, and let $X_i^- = X_i - X_i^+$. Let D^+ and D^- denote the marginal distribution of D over X^+ and X^-, respectively.

For $S_1 \subseteq X_1$ and $S_2 \subseteq X_2$, let boldface $\mathbf{S}_i$ denote the event that an example $\langle x_1, x_2 \rangle$ has $x_i \in S_i$. The $P(\mathbf{S}_i^n)$ denotes the possibility mass on example for which we are confident under i^{th} view in the n^{th} training round. Below we give the definition of ϵ-expanding affixing marks of training round.

Definition 1 *(Balcan et al., 2004) Let X^+ denote the positive region and D^+ denote the distribution over X^+, and $X_i (i = 1, 2)$ is the training data set in the i^{th} view. For $S_1 \subseteq X_1$ and $S_2 \subseteq X_2$, the D^+ is ϵ-expanding if the following inequality holds:*

$$P(\boldsymbol{S}_1 \oplus \boldsymbol{S}_2) \geq \epsilon\, min(P(\boldsymbol{S}_1 \wedge \boldsymbol{S}_2), P(\bar{\boldsymbol{S}_1} \wedge \bar{\boldsymbol{S}_2})), \quad (1)$$

where $P(\boldsymbol{S}_1 \wedge \boldsymbol{S}_2)$ denotes the probability of examples for being confident in both views, and $P(\boldsymbol{S}_1 \oplus \boldsymbol{S}_2)$ denotes the probability of examples for being confident in only one view.

To present training order of classifier under each view, we add superscript for distinguishing the order of iteration. The reivsed definition is:

[1]Xi'an Jiaotong University, Xi'an, China [2]University of Technology Sydney, Sydney, Australia. Correspondence to: Deyu Meng <dymeng@xjtu.edu.cn>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Definition 2 *D^+ is ϵ-expanding in the serial training process if*

$$P(\mathbf{S}_i^n \oplus \mathbf{S}_{3-i}^{n-1}) \geq \epsilon \min(P(\mathbf{S}_{3-i}^{n-1} \wedge \mathbf{S}_i^n), P(\overline{\mathbf{S}_{3-i}^{n-1}} \wedge \overline{\mathbf{S}_i^n})) \quad (2)$$

This ϵ-expanding definition is the same as that defined in (Balcan et al., 2004) except for the round mark in each view. When D^+ satisfies ϵ-expanding in every training round and there are sufficient unlabeled instances, classifiers under each view can acquire arbitrary accuracy with probability $1 - \delta$ after enough training rounds as described in Theorem 1.

Theorem 1 *Let ϵ_{fin} and δ_{fin} be the desired accuracy and confidence parameters. Suppose that serial ϵ-expanding condition is satisfied in each training round, then we can achieve error rate ϵ_{fin} with probability $1-\delta_{fin}$ by running the SPaCo for $N = O(\frac{1}{\epsilon} log \frac{1}{\epsilon_{fin}} + \frac{1}{\epsilon} \cdot \frac{1}{p_{init}})$ rounds, each time running algorithm A_1 and algorithm A_2 with accuracy and confidence parameters set to $\frac{\epsilon \cdot \epsilon_{fin}}{8}$ and $\frac{\delta_{fin}}{2N}$ respectively.*

Similar to proof in (Balcan et al., 2004), we begin by stating two lemmas that will be useful for the analysis. For both lemmas, let $S_i^n \subseteq X_i^+$, and all probabilities are with the respect to D^+.

Lemma1 *Suppose $P(\mathbf{S}_{3-i}^n \wedge \mathbf{S}_i^{n-1}) \leq P(\overline{\mathbf{S}_{3-i}^n} \wedge \overline{\mathbf{S}_i^{n-1}})$, $P(\mathbf{S}_{3-i}^n | \mathbf{S}_{3-i}^n \vee \mathbf{S}_i^{n-1}) \geq 1 - \frac{\epsilon}{8}$ and $P(\mathbf{S}_i^{n+1} | \mathbf{S}_{3-i}^n \vee \mathbf{S}_i^{n-1}) \geq 1 - \frac{\epsilon}{8}$, then $P(\mathbf{S}_i^{n+1} \wedge \mathbf{S}_{3-i}^n) \geq (1 + \frac{\epsilon}{2}) P(\mathbf{S}_{3-i}^n \wedge \mathbf{S}_i^{n-1})$*

Proof

$$\begin{aligned}
&P(\mathbf{S}_i^{n+1} \wedge \mathbf{S}_{3-i}^n) \\
&\geq P(\mathbf{S}_i^{n+1}, \mathbf{S}_{3-i}^n \vee \mathbf{S}_i^{n-1}) + P(\mathbf{S}_{3-i}^n, \mathbf{S}_{3-i}^n \vee \mathbf{S}_i^{n-1}) \\
&\quad - P(\mathbf{S}_{3-i}^n \vee \mathbf{S}_i^{n-1}) \\
&\geq (1 - \frac{\epsilon}{4})(1 + \epsilon) P(\mathbf{S}_{3-i}^n \wedge \mathbf{S}_i^{n-1}) \\
&\geq (1 + \frac{\epsilon}{2}) P(\mathbf{S}_{3-i}^n \wedge \mathbf{S}_i^{n-1})
\end{aligned} \quad (3)$$

Lemma2 *Suppose $P(\mathbf{S}_{3-i}^n \wedge \mathbf{S}_i^{n-1}) > P(\overline{\mathbf{S_{3-i}^n}} \wedge \overline{\mathbf{S}_i^{n-1}})$ and let $\gamma = 1 - P(\mathbf{S}_{3-i}^n \wedge \mathbf{S}_i^{n-1})$, if $P(\mathbf{S}_i^{n+1} | \mathbf{S}_{3-i}^n \vee \mathbf{S}_i^{n-1}) > 1 - \frac{\gamma\epsilon}{8}$ and $P(\mathbf{S}_{3-i}^n | \mathbf{S}_{3-i}^n \vee \mathbf{S}_i^{n-1}) > 1 - \frac{\gamma\epsilon}{8}$, then $P(\mathbf{S}_i^{n+1} \wedge \mathbf{S}_{3-i}^n) \geq (1 + \frac{\epsilon}{2}) P(\mathbf{S}_{3-i}^n \wedge \mathbf{S}_i^{n-1})$*

Proof

$$\begin{aligned}\gamma &= P(\mathbf{S}_{3-i}^{n} \oplus \mathbf{S}_{i}^{n-1}) + P(\overline{\mathbf{S}_{3-i}^{n}} \wedge \overline{\mathbf{S}_{i}^{n-1}}) \\ &\geq (1+\epsilon)P(\overline{\mathbf{S}_{3-i}^{n}} \wedge \overline{\mathbf{S}_{i}^{n-1}}) \\ &\geq (1+\epsilon)(1 - P(\mathbf{S}_{3-i}^{n} \vee \mathbf{S}_{i}^{n-1}))\end{aligned} \tag{4}$$

From inequality 4 we can get $P(\mathbf{S}_{3-i}^{n} \vee \mathbf{S}_{i}^{n-1}) \geq 1 - \frac{\gamma}{1+\epsilon}$. Thus

$$\begin{aligned}P(\mathbf{S}_{i}^{n+1} \wedge \mathbf{S}_{3-i}^{n}) &\geq (1 - \frac{\gamma\epsilon}{4})(1 - \frac{\gamma}{1+\epsilon}) \\ &\geq (1-\gamma)(1 + \frac{\gamma\epsilon}{8}) \\ &\geq (1 + \frac{\gamma\epsilon}{8})P(\mathbf{S}_{3-i}^{n} \wedge \mathbf{S}_{i}^{n-1})\end{aligned} \tag{5}$$

From Lemma 1 and Lemma 2, we present that with fine tuned confidence condition, classifiers trained in a serial way possess same character compared with classifiers built paralleled after each iteration. Therefore, we conclude that with the modified ϵ-expanding condition fulfilled, after same number of iterations, classifiers trained serially can achieve same error rate with same confidence as shown in the original ϵ-expanding theorem (Balcan et al., 2004).

References

Balcan, Maria-Florina, Blum, Avrim, and Yang, Ke. Co-training and expansion: Towards bridging theory and practice. In *Advances in neural information processing systems*, pp. 89–96, 2004.