On Collaboration in Machine Learning

Yu-Zhen Janice Chen
CICS
University of Massachusetts Amherst
Amherst, U.S.
yuzhenchen@cs.umass.edu

Don Towsley
CICS
University of Massachusetts Amherst
Amherst, U.S.
towsley@cs.umass.edu

Dinesh Verma
IBM T. J. Watson Research Center
Yorktown Heights, NY, U.S.
dverma@us.ibm.com

***Abstract*—During the surveillance of an area using ISR assets in coalition operations, the same object may be observed with different modalities by different coalition members. In those cases, the different members may reach different conclusions about the object. Understanding such systems is useful to design principles for coalition collaboration. We model these environments as requiring classification problems done on overlapping but different set of features. As a specific case, we consider a binary classification problem with two features and ask the following question: suppose different observers observe different features, can they reduce the error on classifying random observations by collaboratively learning a model or performing inference collaboratively? If so, can we quantify accuracy improvements? We study this problem by considering two classes C_1, C_2 as bivariate Gaussian distributions with variables x and y; one observer observes variable x for both C_1 and C_2; the other observes variable y for both classes. We consider four strategies: 1) independent learning, independent inference; 2) collaborative learning, independent inference; 3) independent learning, collaborative inference; 4) collaborative learning, collaborative inference. Assuming all models are perfectly learned, we show the relation between these four strategies and the advantage of collaboration. We also analyze the problem assuming the amount of training data affects model accuracy. Although we formulate the problem using simple bivariate Gaussian distribution models, the ideas can lead to deeper insights into the advantage gained by collaborations in machine learning.**

I. Motivation

In coalition settings, the same object may be observed with different modalities by different coalition members. Take finding and tracking person of interest as an example. One coalition member may have GPS signal data; another coalition member may have visual data collected by local infrastructures. In this case, is it worthwhile for them to collaborate in the machine learning processes for classifying whether the target is of interest?

Specifically, Verma et al. [1] introduce two modes of machine learning collaboration in coalition settings: 1) *collaboration during learning phase*, in which coalition partners exchange training data in order to create a joint model that all partners could use; 2) *collaboration during inference phase*, in which coalition partners exchange models learnt locally and share the inferences they reach. This work builds upon [1], which examines and addresses a number of challenges in data sharing and model sharing, to answer more general questions: can classification error be reduced by learning or inferring collaboratively? If so, can we quantify the improvements?

We study this problem by considering a binary classification problem with classes C_1 and C_2. In finding a person of interest example, C_1 denotes the class of high value targets, while C_2 stands for the class of targets not of interest. We assume there are two coalition members, Alice and Bob, involved in this task. In the following section, we introduce our problem formulation.

II. Problem Formulation

Let C_1 and C_2 be two bivariate Gaussian distributions:

$$C_1 : \mathcal{N}(x, y \,|\, \mu_{1x},\, \mu_{1y},\, \sigma_{1x},\, \sigma_{1y},\, \rho),$$
$$C_2 : \mathcal{N}(x, y \,|\, \mu_{2x},\, \mu_{2y},\, \sigma_{2x},\, \sigma_{2y},\, \rho).$$

The probability density function (PDF) of the bivariate Gaussian distribution is:

$$\mathcal{N}(x, y \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[\frac{-1}{2(1-\rho^2)} Q(x, y)\right],$$

$$Q(x, y) = \left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right),$$

where μ_x and μ_y are the means of variables X and Y, respectively; σ_x^2, σ_y^2 are the variances of X, Y, respectively; and ρ is the Pearson correlation coefficient, $\rho \in [-1, 1]$.

And the marginal distributions of C_1, C_2 are:

$$C_{1x} : \mathcal{N}(x \,|\, \mu_{1x}, \sigma_{1x}), \qquad C_{1y} : \mathcal{N}(y \,|\, \mu_{1y}, \sigma_{1y}),$$
$$C_{2x} : \mathcal{N}(x \,|\, \mu_{2x}, \sigma_{2x}), \qquad C_{2y} : \mathcal{N}(y \,|\, \mu_{2y}, \sigma_{2y}).$$

Note that, given any set of two Gaussian distributions, we can always shift and re-scale the coordinate to let C_1 have zero means and unit standard deviations; hence, we assume $\mu_{1x} = 0, \mu_{1y} = 0, \sigma_{1x} = 1, \sigma_{1y} = 1$ without loss of generality. Further, we assume both distributions to have the same correlation ρ and variances $\sigma_{2x} = 1, \sigma_{2y} = 1$ (i.e. C_2 is a shift of C_1) for simplicity of analysis.

In this paper, if not otherwise specified, learning is referred to as estimating bivariate or univariate Gaussian distribution

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

TABLE I
TABLE OF COLLABORATION APPROACHES

Abbrev.	ILII	CLII	ILCI	CLCI
Learning Train Data	Independent $\{x_i\}$	Collaborative $\{(x_i, y_i)\}$	Independent $\{x_i\}$	Collaborative $\{(x_i, y_i)\}$
Inference Use Model	Independent $\hat{C}_{1x}$, $\hat{C}_{2x}$	Independent $\hat{C}_1$, $\hat{C}_2$	Collaborative $\hat{C}_{1x}$, $\hat{C}_{2x}$ $\hat{C}_{1y}$, $\hat{C}_{2y}$	Collaborative $\hat{C}_1$, $\hat{C}_2$
Test Data	$\{x_j\}$	$\{x_j\}$	$\{(x_j, y_j)\}$	$\{(x_j, y_j)\}$

parameters using data samples; a model is referred to as a learned bivariate or univariate Gaussian distribution; and inference is referred to as classifying random observations using likelihood-ratio test. By perfect learning, we mean the estimated value of parameter matches the exact value (e.g. when $\hat{\mu}_{1x} = \mu_{1x}$, $\hat{\sigma}_{1x} = \sigma_{1x}$, we have a perfectly learned marginal model $\hat{C}_{1x}$). As for imperfect learning, we refer to practical settings that we do not have infinite amount of data to learn from but we could design our estimators to get estimates converged asymptotically to the exact value.

Suppose coalition partner Alice observes only variable X of both C_1 and C_2 and has labeled training datasets $\mathcal{X}_1 : x \sim C_{1x}$ and $\mathcal{X}_2 : x \sim C_{2x}$. Suppose Bob observes only variable Y of both C_1 and C_2 and has labeled training data sets $\mathcal{Y}_1 : y \sim C_{1y}$ and $\mathcal{Y}_2 : y \sim C_{2y}$. Alice and Bob may take one of the following four collaboration approaches: 1) independent learning, independent inference (**ILII**); 2) collaborative learning, independent inference (**CLII**); 3) independent learning, collaborative inference (**ILCI**); 4) collaborative learning, collaborative inference (**CLCI**). Table I describes what training and testing data and model(s) are used in each collaboration approach from Alice's perspective.

III. ANALYSIS

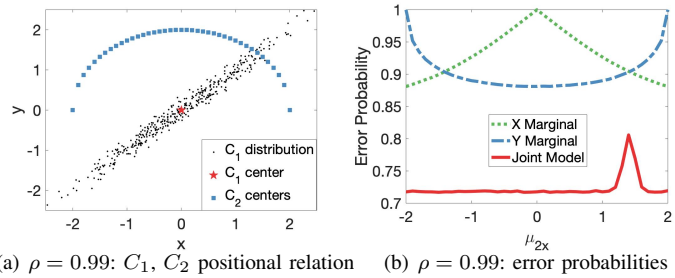
A. Perfect Learning

In this subsection, we assume models are perfectly learned (from an infinite amount of data). Therefore, by independent learning, Alice and Bob each obtain accurate marginal models $\hat{C}_{1x}$, $\hat{C}_{2x}$ and $\hat{C}_{1y}$, $\hat{C}_{2y}$ respectively; by collaborative learning, they both obtain accurate joint models $\hat{C}_1$, $\hat{C}_2$. Next, we consider the error probabilities during the inference phase. We denote the probability that a likelihood-ratio test (LRT) classifies an observation as C_2 when it is from C_1 as α and the probability that the LRT classifies an observation as C_1 when it is from C_2 as β. To evaluate the error probabilities $\alpha + \beta$, we apply total variation distance [2], defined as

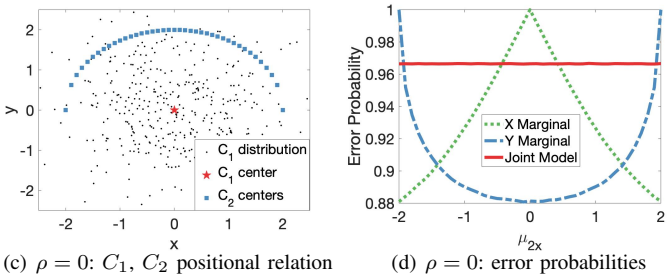
$$\mathcal{V}_T(\mathbb{P}_1, \mathbb{P}_2) = \frac{1}{2}||p_1(x) - p_2(x)||_1,$$

where $p_1(x)$ and $p_2(x)$ are densities of $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively, and $||\cdot||_1$ is the $\mathcal{L}_1$ norm. For likelihood-ratio test, error probabilities

$$\alpha + \beta = 1 - \mathcal{V}_T(\mathbb{P}_1, \mathbb{P}_2).$$



(a) $\rho = 0.99$: C_1, C_2 positional relation (b) $\rho = 0.99$: error probabilities



(c) $\rho = 0$: C_1, C_2 positional relation (d) $\rho = 0$: error probabilities

Fig. 1. Comparing errors of joint models and marginal models

Next, we write down the error probabilities of all four collaboration approaches from Alice's perspective:

$$E_{\text{ILII}} = 1 - \frac{1}{2}||\hat{C}_{1x} - \hat{C}_{2x}||_1. \tag{1}$$

$$E_{\text{CLII}} = 1 - \frac{1}{2}||\hat{C}_{1x} - \hat{C}_{2x}||_1. \tag{2}$$

$$E_{\text{ILCI}} = \min\{1 - \frac{1}{2}||\hat{C}_{1x} - \hat{C}_{2x}||_1, 1 - \frac{1}{2}||\hat{C}_{1y} - \hat{C}_{2y}||_1\}. \tag{3}$$

$$E_{\text{CLCI}} = 1 - \frac{1}{2}||\hat{C}_1 - \hat{C}_2||_1. \tag{4}$$

Note that, in (2), marginal PDFs $\hat{C}_{1x}$, $\hat{C}_{2x}$ are used instead of joint PDFs $\hat{C}_1$, $\hat{C}_2$. Because, in independent inference, Alice only has marginal testing data; hence, we assume Alice will marginalize the joint model for inference. In (3), the smaller error probability between the two is applied as we assume Alice picks the model with highest accuracy.

In the following, we evaluate (1)-(4). In Fig. 1, C_1's center, (μ_{1x}, μ_{1y}), denoted by the red star marker, is fixed at position $(0, 0)$; we let C_2's center (denoted by blue square markers) revolve around C_1's center with radius 2. In Fig. 1(a) and 1(b), correlation $\rho = 0.99$. Fig. 1(a) shows what the shape of distribution and the positional relationship of C_1 and C_2 are. Fig. 1(b) shows, when C_2's center is at $(\mu_{2x} = 0, \mu_{2y} = 2)$, as $\mu_{1x} = 0$, the marginal models of X (denoted by the green dotted line) are useless to distinguish between the two classes and the error probability is 1; similarly, when C_2's center (μ_{2x}, μ_{2y}) is at $(-2, 0)$ or $(2, 0)$, because $\mu_{1y} = 0$, we cannot distinguish between the two distributions using the marginal models of Y (denoted by the blue dashed line). Fig. 1(b) also shows that the error probability of joint models (denoted by the red line) is stable w.r.t. the positional change of C_2 and only grows when C_2's center is around $(\mu_{2x} = \sqrt{2}, \mu_{2y} = \sqrt{2})$ where C_1 and C_2 distributions have the most overlap. In this case, the relation between error probabilities of the four

approaches is

$$E_{\text{CLCI}} < E_{\text{ILCI}} < E_{\text{CLII}} = E_{\text{ILII}}.$$

In Fig. 1(c) and 1(d), correlation $\rho = 0$. Performances of the X and Y marginal models in Fig. 1(d) are similar to the performances in $\rho = 0.99$ case in Fig. 1(b). As for the performances of joint models, Fig. 1(d) shows that they are stable w.r.t. C_2's positional change as the amount of overlap between the two classes are the same. In this case, the relation between error probabilities of the four approaches is

$$E_{\text{ILCI}} < E_{\text{CLCI}} \lesseqgtr E_{\text{CLII}} \lesseqgtr E_{\text{ILII}}.$$

Notice that, in both $\rho = 0$ and $\rho = 0.99$ cases, collaboration in the inference phase improves accuracy.

B. Imperfect Learning

In this subsection, we consider the practical scenario that the number of data samples is not infinite and the efficiency of estimators affect the accuracy of models learnt. In this case, can we learn better model with both coalition members' data than just using data from one member? We examine three learning tasks: 1) one mean unknown, 2) one mean and correlation unknown, and 3) two means unknown.

1) μ_y unknown, μ_x, σ_x, σ_y, ρ known: We study the scenario where Alice has established an accurate model of variable X with a precise estimation of μ_x and σ_x while Bob aims to learn μ_y, given σ_y. Assume correlation ρ between X and Y is known. We revisit results from Bishwal and Pena [3], which show it is worthwhile for coalition members to learn collaboratively.

From information viewpoint, a random vector (x, y) contains more Fisher information [4] about parameter μ_y than a marginal random variable y when x and y are dependent:

$$\mathcal{I}_{(X,Y)}(\mu_y) - \mathcal{I}_Y(\mu_y) = \frac{1}{(1-\rho^2)\sigma_y^2} - \frac{1}{\sigma_y^2} = \frac{\rho^2}{(1-\rho^2)\sigma_y^2} \geq 0.$$

From the estimation perspective, The uniformly minimum-variance unbiased estimator is given by

$$\delta_2 = \bar{y} - \beta(\bar{x} - \mu_x), \;\; \beta = \rho\sigma_y/\sigma_x,$$

which leverages joint observations. Estimator δ_2 is more efficient, i.e., has smaller variance, than the sample mean, $\delta_1 = \overline{y} = \sum_{i=1}^n y_i/n$ The takeaway message is that observations from Alice, whose mean μ_x is known, can be leveraged to reduce the variance of the estimate of μ_y.

2) μ_y, ρ unknown, μ_x, σ_x, σ_y known: When ρ is not known, Bishwal and Pena [3] propose the following estimator for the mean,

$$\delta_3 = \overline{y} - \hat{\beta}(\overline{x} - \mu_x), \text{ where } \hat{\beta} = \frac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}.$$

δ_3 is less efficient than δ_2 but according to Theorem 1 in [3],

$$\frac{\text{Var}(\delta_3)}{\text{Var}(\delta_1)} = \left(\frac{1}{1-\rho^2}\right)\left(\frac{n-3}{n-2}\right),$$

δ_3 is more efficient than the sample mean if and only if $|\rho| > 1/\sqrt{n-2}$. Which shows collaborative learning is beneficial most of the time for this learning task. We refer the reader to [3] for details about δ_3.

3) μ_y, μ_x unknown, σ_x, σ_y, ρ known: Here we study the scenario where Alice and Bob both aim to learn μ_x and μ_y. Having the correlation information, we would expect x samples to be useful for the estimation of μ_y and vice versa. Surprisingly, the maximum likelihood estimator for μ_y in this scenario is the sample mean $\bar{y}$ (similarly, for μ_x):

$$\begin{cases} \frac{\partial}{\partial \mu_x} \log \mathcal{N}(x, y \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(1-\rho^2)} \left(\frac{x-\mu_x}{\sigma_x^2} - \frac{\rho(y-\mu_y)}{\sigma_x \sigma_y} \right) = 0 \\ \frac{\partial}{\partial \mu_y} \log \mathcal{N}(x, y \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(1-\rho^2)} \left(\frac{y-\mu_y}{\sigma_y^2} - \frac{\rho(x-\mu_x)}{\sigma_x \sigma_y} \right) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} (x - \hat{\mu_x})\sigma_y - \rho\sigma_x(y - \hat{\mu_y}) = 0 \\ (y - \hat{\mu_y})\sigma_x - \rho\sigma_y(x - \hat{\mu_x}) = 0 \end{cases}$$

$$\Rightarrow \hat{\mu_y} = \bar{y}$$

This shows there is no advantage to collaborate across coalition members. This result generalizes to the multivariate case. That is, when the means are unknown and the covariance matrix is known, the maximum likelihood estimators are also sample means $\boldsymbol{\mu}$. The log-likelihood function of k-variate normal distribution is

$$\begin{aligned} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{x}_1, \ldots \mathbf{x}_n) = &- \frac{nk}{2} \ln(2\pi) - \frac{n}{2} \ln(\det(\boldsymbol{\Sigma})) \\ &- \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}). \end{aligned}$$

$$\begin{aligned} &\nabla_{\boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{x}_1, \ldots \mathbf{x}_n) \\ &= \nabla_{\boldsymbol{\mu}} \left[-\frac{nk}{2} \ln(2\pi) - \frac{n}{2} \ln(\det(\boldsymbol{\Sigma})) - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \\ &= - \sum_{i=1}^n \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = -\boldsymbol{\Sigma}^{-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}) = 0 \end{aligned}$$

Therefore, the maximum likelihood estimator for $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$.

IV. Conclusion

In this work, we study four collaboration approaches in coalition settings, compare their classification error probabilities assuming the models learnt are perfect, and show when collaboration is beneficial for learning a model. Future work will examine the classification errors or error probabilities with imperfect models.

References

[1] D. Verma, S. Calo, S. Witherspoon, E. Bertino, A. A. Jabal, A. Swami, G. Cirincione, S. Julier, G. White, G. de Mel *et al.*, "Federated learning for coalition operations," *arXiv preprint arXiv:1910.06799*, 2019.

[2] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[3] J. Bishwal and E. A. Peña, "A note on inference in a bivariate normal distribution model," *PO Box*, vol. 14006, pp. 27 709–4006, 2008.

[4] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.