Learning From Multiple Imperfect Instructors in Sensor Networks

Nurali Virani, Shashi Phoha, and Asok Ray, *Fellow, IEEE*

***Abstract*— This paper presents a sequential learning framework for sensors in a network, where a few sensors assume the role of an instructor to train other sensors in the network. The instructors provide estimated labels for measurements of new sensors. These labels are possibly noisy, because a classifier of the instructor may not be perfect. A recursive density estimator is proposed to obtain the true measurement model (i.e., the observation density conditioned on the label) in spite of the training with noisy labels. Specifically, this paper answers the question "Can a sensor train other sensors?", provides necessary conditions for sensors to act as instructors, presents a sequential learning framework using recursive nonparametric kernel density estimation, and provides a convergence rate for the expected error in an observation density. The underlying concepts are illustrated and validated with simulation results.**

***Index Terms*— Information fusion, kernel density estimation, label noise, sequential learning.**

I. Introduction

SENSORS are used for monitoring of complex systems to assess the underlying state of the system, such as health monitoring in physical systems or target detection/classification in surveillance systems. Some of these sensors (e.g., cameras) are easy to train with supervised learning approaches, such as neural networks and support-vector machines, due to abundance of labeled training data. However, for other sensing modalities (e.g., geophones), which do not have adequate training data or which are more sensitive to the operating conditions of the environment, it is necessary to collect and label *in situ* training data to obtain the corresponding measurement models. In this paper, an alternative approach to obtain generative models for new sensors is presented by using the measurements from those sensors along with labels generated by existing sensors in the sensor network. The sensors which contribute (possibly incorrect) labels to *educate* the new sensor are called imperfect instructors. The objective of this paper is to enable the sensors to learn true measurement models even after being trained with noisy labels from a few sensors.

The relationship between observations/features and states can be represented as the conditional probability $p(Y|X)$, where Y is the measurement from the sensor when the system state is X. This conditional probability is called the measurement model of the sensor and it does not assume any specific structure of the relationship and noise characteristics. This paper presents a sequential online learning framework to estimate the true measurement model for a new sensor by using measurements and estimated labels of discrete states from multiple (possibly heterogeneous) sensors in a network. The concepts of probabilistic graphical models and conditional independence given the state have been used to develop a linear relationship between observation densities conditioned on the labels from instructors and the (true) measurement model. If the underlying linear operator is invertible, then it is implied that the existing sensor is qualified to be an instructor and that the (true) measurement model can be inferred eventually. This paper builds upon the theory of recursive density estimation [1], [2] to develop a sequential estimator of the conditional observation density using the observed feature and estimated labels from several sensors acting as instructors.

A. Motivation

Let us consider an example of a border surveillance network with cameras, geophones, and microphones. If more geophones are added to reduce the probability of false alarms and eventually reduce the computation and power used by cameras, then it might be necessary to train the target models for geophones. However, geophones are known to be sensitive to various soil conditions [3] and thus perform better with *in situ* training data [4]. Due to high sampling rate and persistent usage, abundant *in situ* data can be acquired; however, providing training labels to those samples manually may not be feasible. In the paradigm of dynamic data-driven application systems [5], where data are often used during operation to improve models, it is of interest to explore if the existing sensors can be used to train the target models online for the new sensors. Moreover, the high-fidelity sensors might not be used frequently after the new sensors are calibrated to the environment, which should reduce long-term operating cost of the network.

B. Prior Work

Frénay and Verleysen [6] have discussed several sources of label noise and have also reviewed several learning methods that are robust to label noise. The source of label noise in the current framework is due to imperfect classification performance of the sensors that assume the role of instructors for the new sensors. The problem of learning with label noise has been addressed in the literature for one-step classification problems using batch processing approaches [7]. In addition, in [8] and the references therein, a few different approaches to learn with noisy labels using deep learning is presented for convolutional neural networks. However, in most of the existing approaches (including [7]–[9]), the noise-corrupted training set is available at the outset. Hua *et al.* [10] have developed a collaborative active learning framework that uses labels for different images from multiple (possibly imperfect) human instructors to create a discriminative model using ensemble of classifiers. The proposed method relies on a sequential approach with streaming data to update the generative measurement models and to mitigate the effects of label noise. Scott *et al.* [9] have addressed a classification problem in which the class-dependent label noise is estimated from the training data and measurement models are recovered by maximal denoising. Unlike [9], the class-dependent label noise statistics are assumed to be known in this paper, because the measurement models and classifiers of sensors acting as instructors are assumed to be known a priori.

Manuscript received October 18, 2016; revised March 19, 2017 and November 3, 2017; accepted January 3, 2018. Date of publication February 1, 2018; date of current version September 17, 2018. This work was supported by the U.S. Air Force Office of Scientific Research under Grant FA9550-12-1-0270 and Grant FA9550-15-1-0400. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsoring agencies. *(Corresponding author: Asok Ray.)*

N. Virani was with the Department of Mechanical and Nuclear Engineering, The Pennsylvania State University, University Park, PA 16801 USA. He is now with GE Global Research, Niskayuna, NY 12309 USA (e-mail: nurali.virani88@gmail.com).

S. Phoha is with the Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16801 USA (e-mail: sxp26@arl.psu.edu).

A. Ray is with the Department of Mechanical and Nuclear Engineering, The Pennsylvania State University, University Park, PA 16801 USA (e-mail: axr2@psu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2018.2791898

Since multiple heterogeneous sensors act as instructors, the label noise statistics are modeled to be sensor-dependent (i.e., instructor-dependent). In order to develop a sequential estimator, this paper has used the classical work in statistics on recursive density estimation from [1], [2], and [11], which extends the kernel density estimators developed by Rosenblatt *et al.* [12] and Parzen [13] for recursive estimation, and the results of consistency and convergence in these works have been leveraged in the proposed approach.

C. Contributions

This paper considers the case in which an expert may not be able to provide adequate labels to observations, and therefore one must rely on other sensors to aid in learning the measurement model and classifier for a new sensor. This perspective is fundamentally different from most of the work reported in the open literature and has not yet been explored to the authors' knowledge. Specifically, this paper derives a recursive density estimation approach to sequentially learn the true measurement model by using measurements, labels estimated from several sensors, and their corresponding state-dependent sensor-dependent label noise statistics. In this paper, we prove the consistency of the derived estimator and provide the necessary and sufficient conditions for the sensors to act as instructors, and the convergence rate is derived as a function of the number of updates and instructor-dependent label noise statistics. In addition, the performance of the proposed learning method for *intelligent student* is compared with those of two extreme cases—*blessed student*, where the correct labels are provided by an oracle and *naïve student*, where the new sensor may incorrectly assume that it always obtains the correct labels from its instructors. The performance evaluation is shown by means of simulations with synthetic data in order to compare the errors relative to the ground truth.

II. Learning Problem Formulation

This section formulates the sequential learning problem. Let $\mathcal{X} = \{1, 2, \ldots, L\}$ be the finite set of hypotheses on the random state X of the system, which one wishes to estimate through observations. The hypothesis set $\mathcal{X}$ is known *a priori*. We have a classification problem in this paper, where the state is to be assigned to one of the L possible classes. Let $\mathcal{S} = \{1, 2, \ldots, M\}$ be the finite set of all existing sensors. Let Y_s be the random observation of the state X from sensor $s \in \mathcal{S}$ and let $\widehat{X}_s$ be the estimated state from sensor $s \in \mathcal{S}$ using a deterministic mapping h_s from the measurement space $\mathcal{Y}_s$ to the state set $\mathcal{X}$. The measurement model $p(Y_s|X)$ and the classifier h_s are assumed to be known for all sensors $s \in \mathcal{S}$; thus, the conditional probability $p(\widehat{X}_s|X) = p(h_s(Y_s)|X)$ characterizing the performance of the sensor s is also known. It is noted that if the probability $p(\widehat{X}_s|X)$ is directly available, then the measurement model of the instructor is not needed.

Let Y_0 be the random observation or feature from the new sensor which takes the value $y_0^t \in \mathcal{Y}_0$ at the sampling instant $t \in \mathbb{N}$. At every sampling instant, the new sensor has access to $z^t = (y_0^t, \{\hat{x}_1^t, \hat{x}_2^t, \ldots, \hat{x}_M^t\})$, which includes the measurement y_0^t and the estimated labels from all other sensors. It is noted that $\hat{x}_s^t \in \mathcal{X} \cup \{\phi\}$, where ϕ is the null value implying that the label from sensor s was not available at the instant t; the provision of null value allows the existing sensors to say "I don't know" and also allows the sensors to have different sampling frequencies. The objective of this paper is to use the sequence $\{z^t\}$ to obtain a consistent estimate $\hat{p}(Y_0|X = x)$ of the measurement model of the new sensor for all states $x \in \mathcal{X}$. It is assumed that measurements are independent conditioned on the state, and the sequence of measurements from each sensor is also independent. Furthermore, the sensors are assumed to be collocated in the sense that the same state affects all these measurements at all time instants. The probabilistic graphical model in Fig. 1 exhibits dependences among the random variables of state, observations, and state estimates.

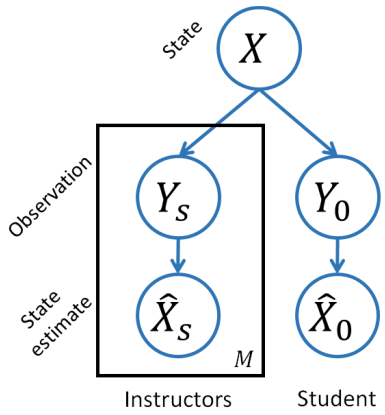

Fig. 1. Probabilistic graphical model showing dependences.

III. Technical Approach

This section first considers a single instructor (i.e., $M = 1$) to obtain the relationship between conditional densities for a given estimated label, and then derives the sequential estimator of the measurement model. This framework of a single instructor is then extended to allow multiple instructors.

A. Relation Between Conditional Densities

Although the probability density of observation of a new sensor conditioned on the label provided by an instructor can be computed, we are interested in the observation density conditioned on the true label. Hence, the relation between these observation densities conditioned on the estimated label and the true observation density is derived using the factorization shown in Fig. 1. Using marginalization, chain rule of probability, and conditional independence of Y_0 and $\widehat{X}_1$ given X (as shown in Fig. 1), the following relation is obtained:

$$\begin{aligned} p(Y_0|\widehat{X}_1 = i) &= \sum_{j=1}^{L} p(Y_0, X = j \mid \widehat{X}_1 = i) \\ &= \sum_{j=1}^{L} p(Y_0 \mid X = j, \widehat{X}_1 = i) p(X = j \mid \widehat{X}_1 = i) \\ &= \sum_{j=1}^{L} p(Y_0 \mid X = j) p(X = j \mid \widehat{X}_1 = i) \\ &= \sum_{j=1}^{L} \alpha_{ij}^1 \; p(Y_0 \mid X = j) \end{aligned} \tag{1}$$

where the constants α_{ij}^1 are obtained by using Bayes' rule as

$$\alpha_{ij}^1 = p(X = j|\widehat{X}_1 = i) = \frac{c_{ij}^1 w_j}{\sum_{k=1}^{L} c_{ik}^1 w_k}. \tag{2}$$

These constants depend on prior probabilities $\{w_j = p(X = j)\}$ and classification performance $\{c_{ij}^1 = p(\widehat{X}_1 = i|X = j)\}$ of sensor 1. It is noted that, in (1), the observation density conditioned on an estimated label is a mixture model of true measurement densities with known

mixture weights but unknown components. The objective here is to identify the components of the mixture model.

Let us denote $p(y|X=i)$ by $p_i(y)$ and $p(y|\widehat{X}_1=i)$ by $\tilde{p}_i(y;1)$ for brevity. Then, (1) is rewritten for all $i \in \mathcal{X}$ and for each $y \in \mathcal{Y}_0$ in a matrix form as follows:

$$\begin{bmatrix} \tilde{p}_1(y;1) \\ \vdots \\ \tilde{p}_L(y;1) \end{bmatrix} = \begin{bmatrix} \alpha_{11}^1 & \cdots & \alpha_{1L}^1 \\ \vdots & \ddots & \vdots \\ \alpha_{L1}^1 & \cdots & \alpha_{LL}^1 \end{bmatrix} \begin{bmatrix} p_1(y) \\ \vdots \\ p_L(y) \end{bmatrix}$$
$$\widetilde{P} = \mathbf{A}_1 \, P.$$

Here, for each observation y, $\widetilde{P}$ is a column vector of label-conditioned observation densities that can be estimated, and P is a column vector of true observation densities that are the desired outputs. The matrix $\mathbf{A}_1$ consists of all values $[\alpha_{ij}^1]$ from (2). If $\mathbf{A}_1$ is invertible, then

$$P = \mathbf{A}_1^{-1} \, \widetilde{P} = \mathbf{B}_1 \, \widetilde{P}. \tag{3}$$

Matrix $\mathbf{A}_1$ depends on prior state probabilities and classification performance of Sensor 1; thus if $\mathbf{A}_1$ is invertible, then Sensor 1 can be an instructor for the new sensor. The implication of the invertibility requirement is discussed further in Section III-C. Using (3), the measurement model is obtained as a linear combination of the probability densities conditioned on the estimated labels as

$$p_i(y) = \sum_{j=1}^{L} \beta_{ij}^1 \; \tilde{p}_j(y;1) \tag{4}$$

where the constant β_{ij}^1 is an element of the matrix $\mathbf{B}_1$ and the equality holds for each $y \in \mathcal{Y}_0$. Thus, if $\tilde{p}_j(y;1)$ is estimated for all $j \in \mathcal{X}$ using the sequence $z^t = (y^t, \hat{x}_1^t)$, then (4) can be used to estimate the true measurement model for each state $i \in \mathcal{X}$. In Section III-B, a sequential estimator for the measurement models is derived.

B. Recursive Density Estimation

In this part, a recursive density estimator is used to sequentially estimate $\tilde{p}_j(y;1)$ for all $j \in \mathcal{X}$ and an estimator is derived for $p_i(y)$ using (4). All the observation-label pairs z^t, where the label from instructor $\hat{x}_1^t$ is j are used sequentially to estimate the conditional density $p(Y|\widehat{X}_1=j)$. Using the recursive density estimator from [1], the observation density conditioned on a label is updated as

$$\tilde{p}_j^t(y;1) = \frac{n_j^t - 1}{n_j^t} \tilde{p}_j^{t-1}(y;1) + \frac{1}{n_j^t h\left(n_j^t\right)^d} K\left(\frac{y - y^t}{h\left(n_j^t\right)}\right) \tag{5}$$

where n_j^t is the number of observations that have been assigned label $j \in \mathcal{X}$ in t time steps, $K(\cdot)$ is a kernel function, $h(n)$ is the kernel width after n updates, and d is the dimensionality of the observation. The number of observations is updated by incrementing $n_j^t = n_j^{t-1} + 1$. For all other labels $i \in \mathcal{X} \setminus j$, the densities remain the same $\tilde{p}_i^t(y;1) = \tilde{p}_i^{t-1}(y;1)$ and observation counts remain the same $n_i^t = n_i^{t-1}$. It is noted that if the label from an instructor is unavailable, i.e., $\hat{x}_1^t = \phi$, then all densities and observation counts stay the same for that instructor.

Considering certain assumptions on kernel function $K(\cdot)$, kernel width sequence $\{h(n)\}$, and assuming that measurements are independent and identically distributed (IID) conditioned on state, we leverage the results from [1] and [2] to show that $\tilde{p}_i^t(y;1)$ is an unbiased and consistent estimator of $\tilde{p}_i(y;1)$. The assumptions and relevant results are presented in Section III-C. Using all observation densities conditioned on the estimated label, an estimate of the measurement model is obtained by sequentially using (4) as

$$p_i^t(y) = \sum_{j=1}^{L} \beta_{ij}^1 \; \tilde{p}_j^t(y;1). \tag{6}$$

C. Analysis

This section first presents the assumptions on the kernel (i.e., Assumptions 1 and 2) and the kernel bandwidth sequence (i.e., Assumption 3). These assumptions are used to leverage the results from [1] to prove that the proposed density estimator is consistent and to provide an upper bound of the error.

Assumption 1: Kernel K is a Borel-measurable function, which is bounded, i.e., $\sup_{y\in\mathcal{Y}} |K(y)| < \infty$, absolutely integrable, i.e., $\int_{y\in\mathcal{Y}} |K(y)| < \infty$, and $\lim_{\|y\|\to\infty} |yK(y)| = 0$.

Assumption 2: Let the function K have Fourier transform K^*, i.e., $K^*(u) = \int_{-\infty}^{\infty} e^{-iuy} K(y) \; dy$. Then, for some $r \in \mathbb{N}$, $\lim_{u\to 0}\{[1 - K^*(u)]/|u|^r\} = k_r$ is finite and the r^{th} derivative of the observation density, i.e., $\tilde{p}_i^{(r)}(y;s)$, exists for all $s \in \mathcal{S}$.

Assumption 3: The kernel bandwidth sequence satisfies $h(n) \to 0$ and $nh(n) \to \infty$. More specifically, the sequence is considered to be $h(n) = bn^{-\gamma}$ with $\gamma \in (0,1)$ and $b > 0$.

Assumption 4: The constant $r \in \mathbb{N}$ in Assumption 2 and the constant $\gamma \in (0,1)$ in Assumption 3 satisfy the strict inequality: $r\gamma < 1$.

Using these assumptions on the kernel and the kernel bandwidth sequence, the error in conditional density given a label from instructor becomes uniformly bounded based on the following lemma.

Lemma 5: If Assumptions 1, 2, 3, and 4 hold, then the density estimation error satisfies the following relation:

$$\sup_{y\in\mathcal{Y}_0} \left|\tilde{p}_j(y;1) - E\,\tilde{p}_j^t(y;1)\right| = O\left(\left(n_j^t\right)^{-\gamma r}\right)$$

where E denotes the expectation and O is the big-O notation denoting order.

Proof: The proof follows directly from [1, Th. 1(*b*)] for $h(n) = bn^{-\gamma}$; further details are given in the Appendix. ■

The label-conditioned density can be computed if all error probabilities are known, that is, $p(\widehat{X}_s = i|X = j)$ is known for all $i, j \in \mathcal{X}$. A sensor may assume the role of an instructor if it is capable of providing unique estimation of the true measurement density model for the new sensor. The necessary and sufficient condition for a sensor $s \in \mathcal{S}$ to be an instructor is that matrix $\mathbf{A}_s = [\alpha_{ij}^s]$, where α_{ij}^s is given in (2), is invertible. The implications of the invertibility condition and the approach to address the noninvertible case are given in the remark as follows.

Remark 6 (Invertibility Condition and Noninvertible Cases): The invertibility condition depends on the prior probability and the instructor's performance as described below.

1) Prior Probability: The matrix $\mathbf{A}_s$ is not invertible if the prior probability of any class j is 0, i.e., $w_j = 0$. In this case, the measurement model $p(Y|X=j)$ for all j with $w_j = 0$ cannot be obtained from this technique as the prior probability suggests that the class j cannot occur. However, an estimate for models of other classes can be obtained by removing the columns corresponding to the zero probability classes from $\mathbf{A}_s$ to obtain matrix $\widetilde{\mathbf{A}}_s$ with full column rank. The pseudoinverse of this nonsquare matrix $\widetilde{\mathbf{A}}_s$ is then computed as $\mathbf{B}_s = (\widetilde{\mathbf{A}}_s^{\mathrm{T}}\widetilde{\mathbf{A}}_s)^{-1}\widetilde{\mathbf{A}}_s^{\mathrm{T}}$.

2) Instructor's Classification Performance: If all prior probabilities are positive, then the matrix $\mathbf{A}_s$ is not invertible if the columns are linearly dependent. Here, linear dependence implies that the instructor is confused between certain states or performs poorly to detect some specific states. For example, $\mathbf{A}_s = [0.5, 0.5; 0.5, 0.5]$

implies that the instructor is confused between states 1 and 2, and $\mathbf{A}_s = [0.8, 0.8; 0.2, 0.2]$ shows that the instructor performs poorly to detect state 2. The measurement model corresponding to these classes cannot be recovered from this technique as the instructor more often wrongly labels the observations. If the linearly dependent columns are removed from $\mathbf{A}_s$ to obtain matrix $\widetilde{\mathbf{A}}_s$, then the pseudoinverse of $\widetilde{\mathbf{A}}_s$ can be used to obtain the density estimates for other classes.

It is noted that if $M = 1$, then the invertibility condition would also be the condition for identifiability of the unknown measurement models, and that condition does not depend on the underlying true observation densities.

Theorem 7 (Main Result): Suppose that Assumptions 1–4 hold. If $\mathbf{A}_1 = [\alpha_{ij}^1]$, from (2), is invertible and the inverse is given by $\mathbf{B}_1 = [\beta_{ij}^1]$, and if the proposed estimator given in (6) is used to estimate the conditional density $p(Y_0 \mid X = i)$ given as $p_i(y)$, then the error in estimation is given by

$$\sup_{y \in \mathcal{Y}_0} \left| p_i(y) - E p_i^t(y) \right| = O\left(\sum_{j=1}^{L} \left| \beta_{ij}^1 \right| \left(n_j^t \right)^{-\gamma r} \right)$$

where n_j^t is the number of observations labeled as $\widehat{x}_1 = j$ after t total updates.

Proof: Using the result from Lemma 5 and definition of big-O notation, it follows that there exists a positive constant $\theta_j \in \mathbb{R}^+$ and a positive integer $\tilde{n}_j$, such that:

$$\sup_{y \in \mathcal{Y}_0} \left| \tilde{p}_j(y; 1) - E \tilde{p}_j^t(y; 1) \right| \leq \theta_j \ \left(n_j^t \right)^{-\gamma r} \tag{7}$$

holds for all $n_j^t \geq \tilde{n}_j$ for each $j \in \mathcal{X}$. On the other hand, taking expectation on (6) and using (4), we obtain

$$p_i(y) - E p_i^t(y) = \sum_{j=1}^{L} \beta_{ij}^1 \ \left(\tilde{p}_j(y; 1) - E \tilde{p}_j^t(y; 1) \right).$$

Using Jensen's inequality [14] for absolute value function, which is known to be a convex function, we obtain

$$\left| p_i(y) - E p_i^t(y) \right| \leq \sum_{j=1}^{L} \left| \beta_{ij}^1 \right| \left| \tilde{p}_j(y; 1) - E \tilde{p}_j^t(y; 1) \right|.$$

Combining the above result with (7) and making use of $\theta_{max} = \max_{j \in \mathcal{X}} \theta_j$, we obtain

$$\left| p_i(y) - E p_i^t(y) \right| \leq \sum_{j=1}^{L} \left| \beta_{ij}^1 \right| \theta_j \ \left(n_j^t \right)^{-\gamma r}$$

$$\left| p_i(y) - E p_i^t(y) \right| \leq \theta_{\max} \sum_{j=1}^{L} \left| \beta_{ij}^1 \right| \left(n_j^t \right)^{-\gamma r}$$

for some $n_j^t \geq \tilde{n}_j$ for all $j \in \mathcal{X}$ for any $y \in \mathcal{Y}_0$. Then using multivariate big O notation, we obtain

$$\sup_{y \in \mathcal{Y}_0} \left| p_i(y) - E p_i^t(y) \right| = O\left(\sum_{j=1}^{L} \left| \beta_{ij}^1 \right| \left(n_j^t \right)^{-\gamma r} \right).$$

■

This theorem gives the error as a function of number of different training labels from a particular instructor, the performance of classifier of the instructor, and prior class probabilities. In this formulation, a perfect instructor is modeled with the identity confusion matrix, i.e., $c_{ij}^1 = 1$ for $i = j$ and $c_{ij}^1 = 0$ otherwise. If the instructor is considered to be perfect, then the following remark explains the error relationship.

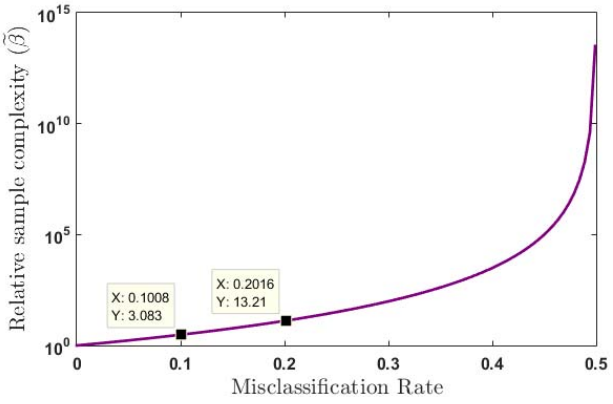

Fig. 2. Relative sample complexity for different misclassification rates.

Remark 8 (Oracle Instructor): If prior class probabilities $w_j \in (0, 1)$, then the perfect instructor is modeled by using the matrix $\mathbf{A}_1$ to be an identity matrix of size L. This implies that if $i = j$, then $\beta_{ij}^1 = 1$, else $\beta_{ij}^1 = 0$. Using Theorem 7, it follows that $\sup_{y \in \mathcal{Y}_0} |p_i(y) - E p_i^t(y)| = O((n_i^t)^{-\gamma r})$. This result agrees with the known result for recursive density estimation and is an extreme case of the proposed model for imperfect instructors.

The next remark addresses the number of examples needed from an imperfect instructor to obtain an error of the same order as that from an oracle.

Remark 9 (Relative Sample Complexity): Let us assume that we have equal number of observations for all labels after N updates from an imperfect instructor and after N_o updates from an oracle. In order to have an error of the same order, the following relation must hold for all $i \in \mathcal{X}$ from Remark 8 and Theorem 7:

$$\left(\frac{N_o}{K} \right)^{-\gamma r} = \sum_{j=1}^{K} \left| \beta_{ij}^1 \right| \left(\frac{N}{K} \right)^{-\gamma r}.$$

A rearrangement of this relation yields

$$\widetilde{\beta} = \frac{N}{N_o} = \left(\sum_{j=1}^{L} \left| \beta_{ij}^1 \right| \right)^{\frac{1}{\gamma r}}.$$

Thus, the error introduced by imperfect instructions can be compensated by the proposed estimator in (5) and (6) by using $\widetilde{\beta}$ times more observations.

For example, let us consider a binary state scenario, where one sensor serves as the instructor with equal misdetection rate and false alarm rate, and the prior probability of each class is 0.5, then relative sample complexity from Remark 9 shows that $\widetilde{\beta} \approx 3$, if misclassification rate is 0.1; whereas, $\widetilde{\beta} \approx 1.7$, if misclassification rate is 0.05. Thus, learning from a noisy sensor is indeed feasible according to this analysis. The value of $\widetilde{\beta}$ for different values of misclassification rates from $[0.0, 0.5)$ is given in Fig. 2. It is noted that as misclassification rate approaches 0.5, which is a noninvertible case (Remark 6), we see that $\widetilde{\beta} \to \infty$.

This section has developed a sequential estimator for the measurement density of a single instructor (i.e., $M = 1$). The framework is now extended for multiple instructors (i.e., $M \geq 2$) in the following part. The linear combination of unknown measurement densities in (1) has been shown to be equal to the label-conditioned observation densities, obtained from labels provided by an instructor. Since the unknown measurement densities are the same irrespective of the

instructor, the relation in (1) is generalized for $M \geq 2$ as follows:

$$\begin{bmatrix} \tilde{p}_1^t\ (y;1) \\ \vdots \\ \tilde{p}_L^t\ (y;1) \\ \hline \tilde{p}_1^t\ (y;2) \\ \vdots \\ \tilde{p}_L^t\ (y;2) \\ \hline \vdots \\ \hline \tilde{p}_1^t\ (y;M) \\ \vdots \\ \tilde{p}_L^t\ (y;M) \end{bmatrix} = \begin{bmatrix} \alpha_{11}^1 & \dots & \alpha_{1L}^1 \\ \vdots & \ddots & \vdots \\ \alpha_{L1}^1 & \dots & \alpha_{LL}^1 \\ \hline \alpha_{11}^2 & \dots & \alpha_{1L}^2 \\ \vdots & \ddots & \vdots \\ \alpha_{L1}^2 & \dots & \alpha_{LL}^2 \\ \hline \vdots & \vdots & \vdots \\ \hline \alpha_{11}^M & \dots & \alpha_{1L}^M \\ \vdots & \ddots & \vdots \\ \alpha_{L1}^M & \dots & \alpha_{LL}^M \end{bmatrix} \begin{bmatrix} p_1\ (y) \\ p_2\ (y) \\ \vdots \\ p_L\ (y) \end{bmatrix}$$

$$\widetilde{P} = \mathbf{A}\ P$$

$$P = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\ \widetilde{P} = \mathbf{B}\ \widetilde{P}.$$

If $\mathbf{A}$ has the full column rank, then $\mathbf{B}$ is well-defined and a linear estimator of the measurement density is obtained as

$$p(Y_0 \mid x = i) = \sum_{s=1}^{M} \sum_{j=1}^{L} \beta_{ij}^s\, p(Y_0 \mid \hat{x}_s = j) \tag{8}$$

where, for brevity, the elements of $\mathbf{B}$ are denoted as β_{ij}^s instead of $\beta_{i,L(s-1)+j}$.

It is noted that the coefficients β_{ij}^s in the estimator may have negative values. Consequently, after finitely many updates, the estimate of density at few values of $y \in \mathcal{Y}_0$ can be negative. Thus, for practical usage, only the positive part is considered and the estimator could be modified as

$$p(Y_0 \mid x = i) = \eta \left| \sum_{s=1}^{M} \sum_{j=1}^{L} \beta_{ij}^s\, p(Y_0 \mid \hat{x}_s = j) \right|^{+} \tag{9}$$

where η is a normalization constant and the notation $|\cdot|^{+} = \max(0, \cdot)$.

IV. Simulation Results and Discussion

In this section, the proposed sequential density estimator is validated by using a simulation example. In the example, we consider a scenario in which two existing sensors with known models use a maximum likelihood classifier to generate state estimates. These estimates serve as noisy training labels for a new sensor with a different and unknown sensor model. Referring to Section I, the density estimator for *intelligent* student is compared with that for *naïve* student (which assumes that the estimated label is the true label), and also with that for *blessed* student (which obtains the true label from an oracle) (see Remark 8). The details of the simulation and the results are presented ahead in this section.

A binary classification problem is considered, i.e., $\mathcal{X} = \{0, 1\}$ and $L = 2$, with two instructors, i.e., $\mathcal{S} = \{1, 2\}$ and $M = 2$. The sensor models in the simulation are chosen to have independent Gaussian distributions. Specifically, $p(Y_i|X) \sim \mathcal{N}(\mu_X^i, \Sigma_X^i)$, where $\Sigma_X^i = 1$ for all $i \in \{0, 1, 2\}$, $\mu_0^1 = \mu_0^2 = -1$, $\mu_0^0 = -3$, $\mu_1^1 = \mu_1^2 = 1$, and $\mu_1^0 = 3$. It is noted that the unknown model $p(Y_0|X)$ is less overlapping compared to those of the instructors $p(Y_1|X)$ and $p(Y_2|X)$. The state estimates are generated by the instructor $s \in \mathcal{S}$ by using the observation $y_s \in \mathcal{Y}_s$; and the maximum likelihood classifier is obtained as

$$\widehat{x}_s(y_s) = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\ p(y_s|x).$$

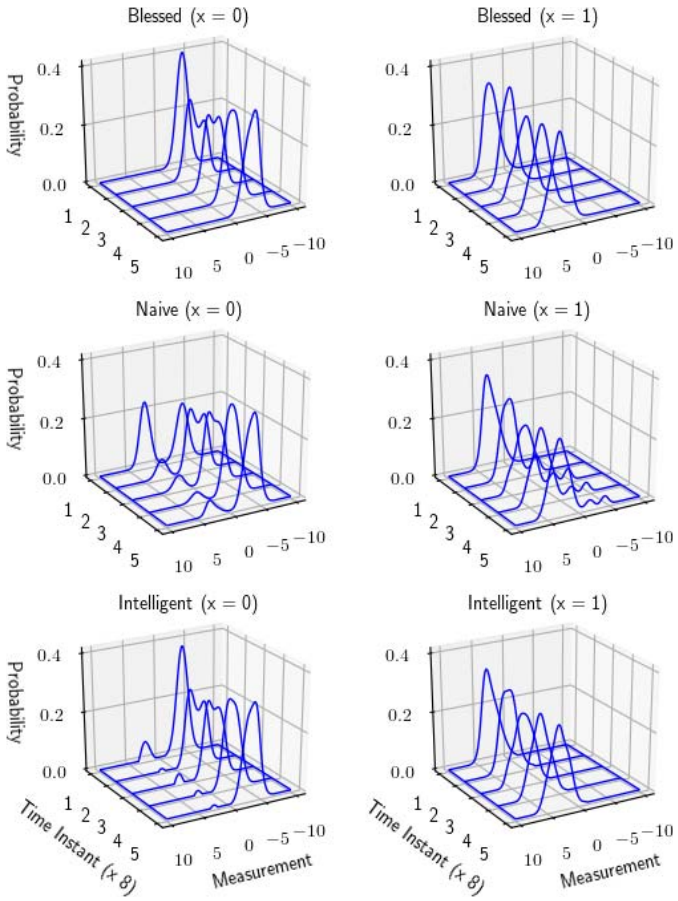

Fig. 3. Sequential update of density estimates for each state using different approaches (40 updates).

The normalized confusion matrix for the instructor's classifier, i.e., $p(\widehat{X}_s|X)$, is computed either by sampling or by using the known cumulative density function. Results of the sequential estimator for the first 40 updates are shown in Fig. 3. The estimated model after 40 and 1000 updates from different approaches is compared in Fig. 4(a) and (b), respectively. It is noted that, for the *naïve* student, the label noise leads to an incorrect measurement model with spurious components as compared to that of the *blessed* student. The *intelligent* student (in Fig. 3) removes the spurious components after using more observations to be closer to the estimate of the *blessed* student. Since the true model is known for the new sensor, it is observed that the estimate after 1000 updates is closer to the true model as compared to that after 40 updates. The estimated models are compared with the true model by computing the maximum absolute error over both states. The resulting plot is shown in Fig. 4(c). The *naïve* student converges to a relatively poor model, whereas the proposed approach keeps improving the estimate as more samples become available; however, as expected, the improvement is slightly less than learning from an oracle. The estimate from two instructors is only slightly better than that from a single instructor, because the proposed method computes a weighted average of the estimates from each instructor and the weights depend on classification performance of the individual instructor. In this simulation example, a noisy sensor with overlapping classes has been used to successfully train a sensor with relatively less overlapping classes. Since the generative model for data in each class is obtained sequentially, the estimate of the maximum likelihood classifier is obtained after each update. In this way, the proposed method is validated to sequentially update

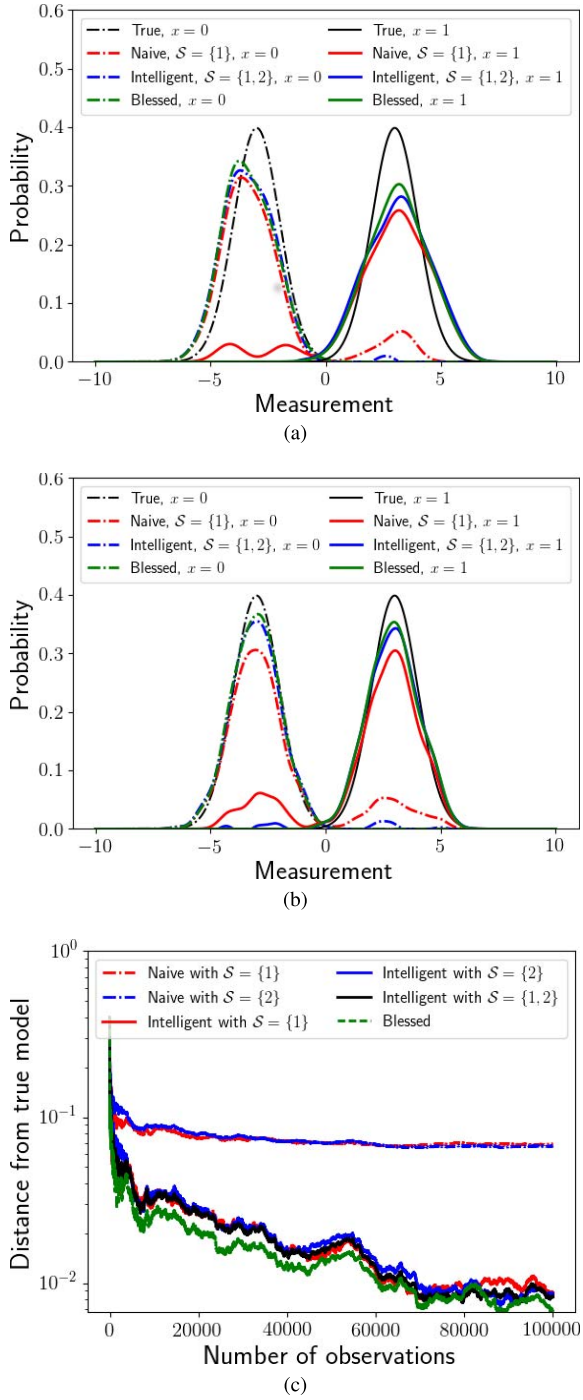

Fig. 4. Comparison of errors in density estimation. (a) Density estimate after 40 updates. (b) Density estimate after 1000 updates. (c) Average error in density estimation.

the density estimate using a streaming sequence of labels from the instructors along with the sensor's own measurements/features.

V. Summary and Conclusion

This paper presents a new perspective of learning from other sensors in a network and it is proven that existing sensors can teach new sensors. The proposed framework has been developed by using recursive density estimation and probabilistic graphical models to obtain a sequential estimator for the measurement models. The developed estimator uses estimated labels from existing sensors and their state-dependent statistics of label noise to obtain a generative model of true observation density. The estimation error bound is obtained by using results from recursive kernel density estimation [1].

There are several areas of theoretical work to develop the proposed method for real-life implementation. The authors suggest the following topics for future research.

1) *Incorporating the Effects of Number of Labels:* The proposed framework fuses the densities from each instructor based only on their classification accuracy and does not consider the number of labels contributed by each instructor. The number of labels is expected to affect the uncertainty in the density estimate. The framework can be extended to include the number of labels contributed from each instructor.
2) *Relaxation of IID and Conditional Independence Assumption:* This paper assumes that the sequence of measurements conditioned on state is IID; however, consistency and asymptotic normality properties of recursive kernel estimators have been shown under certain dependency conditions [15]. The future research may address relaxation of the assumption of conditional independence given state for measurements from different sensors.
3) *Estimation of Instructor's Performance:* This paper assumes that the performance of the instructor is accurately known at the outset, which for example may not be true for labels obtained from annotations of web images [8]. The framework can be updated to estimate the instructor's performance for incorporation in the sequential estimation.
4) *Field Experimentation:* The future research may demonstrate the proposed learning method with heterogeneous sensors (e.g., training of geophones using cameras).

Appendix

The theorem used in the proof of Lemma 5 from [1] is given in this appendix for completeness of this paper. The unknown density is represented by $f(y) = \tilde{p}_i(y; s)$ and the recursive estimator from (5) is $f_t^*(y)$.

Theorem 10 [1]: Let the kernel K satisfy the Assumptions 1 and 2, and let the kernel bandwidth sequence $h(n)$ satisfy the conditions: $h(n) \to 0$ and $nh(n) \to \infty$. Let $nh(n)^r \to \infty$ and $1/\left(nh(n)^r\right)\sum_{j=1}^{n} h(j)^r$ converge to some $\gamma_r \in \mathbb{R}$, where r is given in Assumption 2. Then

$$\frac{Ef_t^*(y) - f(y)}{h(n)^r} \to \gamma_r k_r f^r(y).$$

It is noted that if the kernel sequence is chosen to be $h(n) = bn^{-\gamma}$ (as in this paper), then $\gamma_r = 1/(1 - \gamma r)$, so $Ef_t^*(y) - f(y) = O(n^{-\gamma r})$, because $\|f^r\|_\infty$ is bounded (Assumption 2).

References

[1] E. J. Wegman and H. Davies, "Remarks on some recursive estimators of a probability density," *Ann. Statist.*, vol. 7, no. 2, pp. 316–327, 1979.
[2] E. Masry and L. Györfi, "Strong consistency and rates for recursive probability density estimators of stationary processes," *J. Multivariate Anal.*, vol. 22, no. 1, pp. 79–93, 1987.
[3] J. R. McKenna and M. H. McKenna, "Effects of local meteorological variability on surface and subsurface seismic-acoustic signals," DTIC, Fort Belvoir, VA, USA, DTIC Rep. ADA481553, 2006.
[4] N. Virani, S. Sarkar, J.-W. Lee, S. Phoha, and A. Ray, "Algorithms for context learning and information representation for multi-sensor teams," in *Context-Enhanced Information Fusion* (Advances in Computer Vision and Pattern Recognition), L. Snidaro, J. García, J. Llinas, and E. Blasch, Eds. Cham, Switzerland: Springer, 2016, pp. 403–427.

[5] F. Darema, "Dynamic data driven applications systems: A new paradigm for application simulations and measurements," in *Proc. Int. Conf. Comput. Sci.*, 2004, pp. 662–669.
[6] B. Frénay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.
[7] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 1196–1204.
[8] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2691–2699.
[9] C. Scott, G. Blanchard, and G. Handy, "Classification with asymmetric label noise: Consistency and maximal denoising," in *Proc. Conf. Learn. Theory*, 2013, pp. 489–511.
[10] G. Hua, C. Long, M. Yang, and Y. Gao, "Collaborative active learning of a kernel machine ensemble for recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1209–1216.
[11] E. Masry, "Nonparametric estimation of conditional probability densities and expectations of stationary processes: Strong consistency and rates," *Stochastic Process. Appl.*, vol. 32, no. 1, pp. 109–127, 1989.
[12] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, no. 3, pp. 832–837, 1956.
[13] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, Sep. 1962.
[14] T. Needham, "A visual explanation of Jensen's inequality," *Amer. Math. Monthly*, vol. 100, no. 8, pp. 768–771, 1993.
[15] G. G. Roussas and L. T. Tran, "Asymptotic normality of the recursive kernel regression estimate under dependence conditions," *Ann. Statist.*, vol. 20, no. 1, pp. 98–120, 1992.