Towards Joint Segmentation and Active Learning for Block-Structured Data Streams

Conrad Holtsclaw, Meet P. Vadera, and Benjamin M. Marlin
{choltscl,mvadera,marlin}@cs.umass.edu
University of Massachusetts Amherst

ABSTRACT

Stream-based active learning methods assume that data instances arrive in sequence and the decision must be made to query an instance or not as it arrives. In mobile health and human activity recognition, the data stream is often block-structured where instances in the same block have the same label, but the boundaries between blocks are unobserved. In this paper, we propose an approach to active learning in this setting where we simultaneously learn to segment the stream while learning an instance-level discriminative classifier. We show that by propagating collected labels into inferred segments, we can learn improved predictive models with significantly fewer queries.

KEYWORDS

datastreams, active learning, mobile health, activity recognition

ACM Reference Format:
Conrad Holtsclaw, Meet P. Vadera, and Benjamin M. Marlin . 2019. Towards Joint Segmentation and Active Learning for Block-Structured Data Streams. In *KDD '19: ACM SIGKDD Conference on Knowledge Discovery and Data Mining Workshop on Data Collection, Curation, and Labeling for Mining and Learning, August 5, 2019, Anchorage, Alaska.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

While unlabeled data are abundant in many traditional application areas of machine learning and data mining, the labeled data required to apply supervised learning methods remain scarce in many domains due to the time and cost required for data annotation. Active learning methods attempt to address this problem under a model where a learner creates a labeled data set by iteratively querying an oracle to obtain labels for instances [22]. The goal of active learning methods is to minimize the generalization loss that a given supervised learning model achieves given a sample of N labeled training data instances by selectively sampling (as opposed to randomly sampling) instances for labeling by the oracle.

Active learning methods fall into several categories based on how the learner interacts with the labeling oracle. In this work, we focus on the stream-based (or online) setting where the learner must make sequential decisions about whether to query for the label of each instance in a stream of instances [4]. We consider a specialization of active learning in the online setting where the data stream is block-structured, by which we mean that instances in the same block have the same label. However, the boundaries between blocks are not observed. This problem is motivated by online active learning for human activity recognition [10] and mobile health [9] where users must be queried in real-time for labels, but activity states naturally persist though time.

To leverage the resulting block structure of the label streams, we investigate methods for joint segmentation and active learning. In particular, we focus on the use of semi-supervised hidden Markov models to learn to segment streams, coupled with the learning of discriminative classification models. Our approach allows obtained label information to propagate along inferred segments, effectively enabling a single active learning query to label an inferred block of instances. We present results using both basic logistic regression models and neural network models showing that this approach can significantly out-perform standard stream based active learning methods that apply the results of queries to single data points.

2 BACKGROUND AND RELATED WORK

In this section, we begin by presenting technical background on active learning. We then discuss active learning in the human activity recognition and mobile health domains.

Notation: We denote data instances by $\mathbf{x} \in \mathcal{X}$ and corresponding labels by $\mathbf{y} \in \mathcal{Y}$. We let C denote the number of labels. Our main interest in this work is learning score-based, instance-level discriminative classification functions $f : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$. We will denote the space of possible classification functions by $\mathcal{F}$ (typically indexed by a continuous parameter vector $\mathbf{w}$). The canonical hard classification rule is given by $\arg\max_{y' \in \mathcal{Y}} f(y', \mathbf{x})$. We denote the oracle labeling function by $\Omega : \mathcal{X} \to \mathcal{Y}$. We define $L(\mathcal{D})$ to be a supervised classifier learning function, which takes a labeled training set $\mathcal{D}$ as input and outputs a classifier.

Stream-Based Active Learning Background: In classical stream-based or online active learning [4], we assume access to a stream $\mathcal{S}$ of unlabeled instances $\mathbf{x}_t$, $0 \leq t \leq T$ where T is the length of the stream.[1] T may be finite or infinite and may be known or unknown to the active learner. By assumption, the labeling oracle Ω can only be queried for the label of instance $\mathbf{X}_t$ at time t. We denote the learned classifier following the i^{th} query by f_i. The goal of the learner is to minimize the generalization loss of the classifier by optimally selecting instances for labeling from the stream. It is also common to start learning from a small initial data set.

Common methods for stream-based active learning leverage a utility function $u : \mathcal{X} \times \mathcal{F} \to \mathbb{R}$ that estimates the usefulness of

[1] To simplify notation, we will refer to t as indexing time, although the interval between the arrival of instances in the stream need not be uniform in general.

KDD '19, August 5, 2019 , Anchorage, Alaska

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Algorithm 1 Utility maximizing stream-based active learning

Function SAL($\mathcal{S}$, $\mathcal{D}_0$, u, L, Ω, τ, B):
$t \leftarrow 0$; $f_0 \leftarrow L(\mathcal{D}_0)$; $i \leftarrow |\mathcal{D}_0|$
while $i < B$ **do**
  $\mathbf{x}_t \leftarrow \mathcal{S}$
  if $u(\mathbf{x}_t, f_{i-1}) \geq \tau$ **then**
    $y_t \leftarrow \Omega(\mathbf{x}_t)$
    $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \{(y_t, \mathbf{x}_t)\}$
    $f_i \leftarrow L(\mathcal{D}_i)$
    $i \leftarrow i + 1$
  $t \leftarrow t + 1$
Return f_B

obtaining a label for the current instance $\mathbf{X}_t$ based on the classifier f_i available at that time. To decide whether to query an instance, the utility is compared to a fixed threshold [22]. We show a general stream-based active learning method in Algorithm 1 where $\mathcal{S}$ is the stream, $\mathcal{D}_0$ is an initial training data set (possibly empty), u is the utility function, L is the underlying supervised learner, Ω is the labeling oracle, τ is the utility threshold and B is the active learning query budget.

Note that we do not require that the learner L used here operate in a fully online fashion where the current classifier must be updated on the basis of the current labeled instance only. Instead, as in the pool-based setting [12], we assume that all labeled instances collected to date can be used to update the classifier. This approach can easily be modified to update the classifier using a fixed-size buffer of instances of any size, including the current instance only, in which case the supervised learner may also take the previously learned classifier f_{i-1} as an input. The algorithm can also be modified to use an adaptive threshold $\tau(t)$, but in this work we restrict our investigation to the case of a fixed threshold.

The active learning literature includes many different utility functions. Within the uncertainty sampling family, the entropy and margin utilities are both commonly used. Both utilities require the classification score function $f(y, \mathbf{x})$ to provide a probability distribution over the values of y (e.g., $f(y, \mathbf{x}) = p(y|\mathbf{x})$). The entropy-based utility function $u_H(\mathbf{x}, f)$ corresponds to the standard Shannon entropy [23]. The margin utility corresponds to the difference in probability between the most likely class y_* given $\mathbf{x}$ and the second most likely class y'. Instances with smaller margins (higher uncertainty) are sampled, so the utility function $u_M(\mathbf{x}, f)$ must be negated for use in Algorithm 1.

$$u_H(\mathbf{x}, f) = -\sum_{y \in \mathcal{Y}} f(y, \mathbf{x}) \log(f(y, \mathbf{x})) \tag{1}$$

$$u_M(\mathbf{x}, f) = f(y_*, \mathbf{x}) - f(y', \mathbf{x}) \tag{2}$$

mHealth, HAR and Active Learning: This work is motivated by applications in the closely related areas of mobile health (mHealth) and human activity recognition (HAR). In these domains, a key challenge is developing accurate models for detecting a wide range of general and health-related activities and behaviors from multi-modal on-body sensor data including sleep [21], physical activity [25], eating and drinking [24] cigarette smoking [2, 20], drug use [7, 18] and more.

However, the data from commonly used sensors is generally not human interpretable and can generally not be labeled after collection without access to additional sources of data like video. The use of video data to support labeling is impractical outside of controlled lab settings, which themselves suffer from limited real-world validity resulting in other modeling challenges [16]. Alternative field data collection protocols rely on self-reported activity labels, which either places a large burden on study participants if temporally dense labels are requested, or results in substantial missingness if labels are sparsely requested. At the same time, many studies have indicated significantly better performance when personalizing models to individuals instead of using global models [20, 25].

As a result, the online, real-time nature of the activity labeling problem is an excellent match for stream-based active learning methods where methods generate queries in real time and each individual serves as the labeling oracle for their own sensor data streams. Interestingly, this problem has much richer structure than the classical streaming setting. For example, activity states typically persist through time. It is also often feasible to query individuals for the labels of events in the recent past, although with possibly degraded accuracy in the provided labels. Finally, while not all individuals are the same, they are also not all completely different, resulting in the possibility of reducing labeling burden by sharing data across similar individuals.

In terms of prior work, Miu et al. investigated methods for online active learning in the HAR setting, but used the ground truth activity segments on the more challenging case of non-periodic activities that we investigate here, while for the case of periodic activities they leverage a basic change-point method in the feature space [14]. Martindale et al. use a hierarchical hidden Markov model (HMM) as part of a smart data annotation tool in an active learning-like setting for labeling gait data sets [13]. However, they condition on high-level gait labels and use the tool to assist with labeling individual gait cycles. Similarly, Alemdar et al. use an HMM-based model to perform active learning for activity recognition based on ambient (as opposed to on-body) sensors [1]. However, their approach relies on issuing multiple queries once per day and is thus not a fully online streaming active learning method. There is also related work on clustering-based methods for active learning in data streams; however, this work focuses on a batch incremental setting where multiple instances are selected for labeling from a stream at once instead of in a fully online fashion [8]. Finally, we note that past work in the mHealth setting has leveraged between-subject similarity structure when applying active learning to a cohort of individuals simultaneously, but this work was performed in the pool-based active learning setting [17].

3 PROPOSED APPROACH

In this section, we present our proposed approach to the problem of active learning for block-structured data streams.

Overview: As mentioned in the introduction, our ultimate goal is to more efficiently learn an instance-level discriminative classifier by exploiting the block structure of the data stream. To do so, we proposed a hybrid approach that combines online semi-supervised learning of a segmentation model with online active learning of a discriminative instance classifier. The central idea is

Algorithm 2 Joint Segmentation and Active Learning

Function JSAL($\mathcal{S}$, $\mathcal{D}_0$, u, L, Ω, τ, σ):
$t \leftarrow 0$; $f_0 \leftarrow L(\mathcal{D}_0)$; $i \leftarrow |\mathcal{D}_0|$; $b_s \leftarrow 0$
while $i < B$ **do**
  $\mathbf{x}_t \leftarrow \mathcal{S}$
  if $\sigma(\mathbf{x}_t)$ **then**
    $b_e \leftarrow t - 1$
    if $\max_{b_s \le k \le b_e} (u(\mathbf{x}_k, f_{i-1})) \ge \tau$ **then**
      $y_{t-1} \leftarrow \Omega(\mathbf{x}_{t-1})$
      $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \{(\mathbf{x}_k, y_{t-1}) | b_s \le k \le b_e\}$
      $f_i \leftarrow L(\mathcal{D}_i)$; $b_s \leftarrow t$; $i \leftarrow i + 1$
  $t \leftarrow t + 1$
Return f_B

that by introducing a segmenter, which makes predictions about the block-structure of the input data stream, we can infer block boundaries and use single queries to label blocks. We only trigger an evaluation of the active learning utility function at inferred block boundaries and the utility that is assessed is the maximum utility among all instances in the inferred block. If a query is issued, the whole inferred block of labeled instances can then be added to the training set of the discriminative classifier.

Algorithm 2 illustrates the approach using a fixed segmenter σ that returns *True* when it predicts that a new block has started. We operate the segmenter in a conservative, over-segmentation mode, which mitigates the risk of introducing label error due to inaccurate inference of block boundaries at the cost of reduced label propagation. While active learning of the classifier is used to drive the active learning in the proposed framework, the segmenter can also be trained online in a semi-supervised fashion using both abundant unlabeled instances and labeled instances returned by active learning. The segmenter can also condition on observed labels to improve segmentation inference. We use both extensions in this work.

We also note that this algorithm makes assumptions about the availability of unlabeled instances that differ from the more stringent classical streaming setting where it is assumed that instance $\mathbf{x}_t$ must be processed at time t and then discarded. As we can see, Algorithm 2 assumes access to all unlabeled instances from the previous block and the query that is issued is actually for the most recent instance in the previous block. In practice, the additional assumptions pose little practical concern as the individual instances in the mHealth and HAR setting are typically short in duration (for example, one to several minutes long). Similarly, maintaining a buffer of instances for the current block is not a problem as blocks may last from minutes to hours depending on the application. If needed, a maximum allowable block length can also asserted. When learning the segmenter online, it will also typically be necessary to retain additional unlabeled data. Again, a buffer of the desired size can be used for this purpose that satisfies any storage constraints. In our experiments, we do not apply storage restrictions as the data volumes do not require it.

In the next sections, we discuss the choices we explore for the segmenter and classifier in this work.

Segmentation: As noted in the related work section, hidden Markov models (HMMs) [3] have been used as the classifiers in some previous work on active learning for activity recognition. We adopt HMMs as well to exploit their ability to accomodate semi-supervised training. However, we decouple their application as segmenters from their application as classifiers to enable the use of discriminative classifiers.

In this work, we use a Gaussian emission model $Pr(X_t = x|Y_t = y) = N(x|\mu_y, \Lambda_y)$ where μ_y is the class conditional mean and Λ_y is the class conditional covariance matrix. We consider diagonal and full covariance matrices. For the label transition matrix $Pr(Y_t = y'|Y_{t-1} = y) = \pi_{yy'}$, we consider a full parameterization of $\pi_{yy'}$ as well as a simplified parameterization where the probability of no transition is ρ and if a transition occurs, all values of y are equally likely (including the current class).

To train the model, we use the Expectation-Maximization (EM) algorithm [5] with Laplace smoothing on the transition probability parameters. We train the model in a semi-supervised fashion using acquired instance labels together with unlabeled instances seen to date. We choose to update the model only when a new labeled instance is obtained or K time steps have passed since the last query was issued. The update schedule parameter K is introduced for increased control of computation required.

Given the current trained model, the decision rule for block boundaries is straightforward. Given a hyper-parameter ϵ, the segmenter will predict that $\mathbf{x}_t$ belongs to a new segment if:

$$\sum_{y \in \mathcal{Y}} Pr(Y_t = Y_{t-1} | X_1, \ldots, X_t) \le \epsilon$$

In other words, the segmenter predicts that a point is the start of a new block if the probability that it shares the same label as its predecessor is not sufficiently high. This probability is easily obtainable as a pairwise marginal from the messages computed by the forward-backward algorithm used within the EM algorithm [15]. We note that the structure of this inference shows its advantage over using the instance classifier itself to decide when block boundaries occur since it marginalizes over the possible *joint* label assignments of the two instances give the transition structure and history of the HMM. We further note that this inference can also condition on whatever labels have previously been obtained. Finally, we note that the parameter ϵ directly controls how conservative the segmenter is. The larger the value of ϵ, the more conservative the segmenter and the smaller the blocks will be on average.

Discriminative Classification: We consider two probabilistic classification models in this work: a basic multi-class logistic regression model and a Bayesian feed-forward neural network model. The logistic regression model [6] parameterizes the conditional probability $P(Y = y|\mathbf{x})$ of the labels given the features using a log linear function of the features. Due to the fact that the number of parameters equals the number of features, the model has minimal variance and can be efficiently trained using standard maximum likelihood methods at small data volumes. This makes it a good choice for active learning with low label budgets, which is the scenario we focus on here, although the model will have high bias in cases where the data are not close to being linearly separable.

We also consider Bayesian feed-forward neural network models as an approach to achieving a more flexible bias-variance trade-off

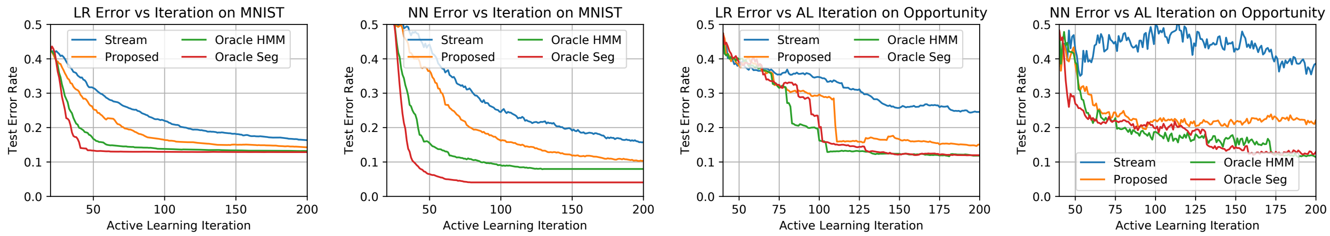

Figure 1: Results of active learning experiments on the MNIST and Opportunity data sets.

during active learning. Specifically, we adopt Stochastic Gradient Langevin Dynamics (SGLD) as an efficient Markov chain Monte Carlo (MCMC) method for implementing Bayesian inference in neural network models [26]. Given a neural network architecture and a prior over the network's parameters $\mathbf{w}$, SGLD allows for sampling model parameters $\mathbf{w}_s$ from the model's posterior distribution over parameters when conditioning on a training data set. The SGLD algorithm constructs a Markov chain using gradient information that can be computed from mini-batches, which allows for the per-iteration cost to be controlled as the data set size grows.

In this work, we use an incremental, parallel-chains variant of SGLD. We run M completely separate SGLD chains. Each chain provides a sample of the weights that is independent of the other chains. The model's posterior predictive distribution $P(Y = y|\mathbf{x}, \mathcal{D})$ is then obtained as a Monte Carlo average of the predictions obtained from the current sample provided by each chain. When a new labeled instance is obtained, the posterior is updated by advancing all chains for a specified number of steps. The prediction cost of this approach is M times higher than when using a single neural network model. However, the use of MCMC methods instead of point-estimation has the potential to enable the effective use of neural network models in active learning despite their high variance, while avoiding the approximation bias that occurs when using variational Bayesian inference.

4 EXPERIMENTS AND RESULTS

In this section, we present initial experimental results comparing the proposed approach (Proposed) to the traditional stream-based active learning setting where a single query is used to label a single instance (Stream). We also consider two oracle baselines including an oracle HMM segmenter trained on all available data for a given data set (Oracle HMM), and an oracle segmenter that uses the true segment boundaries (Oracle Seg).

Data Sets: We consider two data sets: MNIST [11] and the Opportunity Human Activity Recognition data set [19]. MNSIT contains 60,000 training images and 10,000 test images from digit classes 0 to 9. We convert the data set into a stream by sampling from a Markov chain model over the labels and then randomly assigning instances to positions in the stream with a matching label. We use a model matching the simplified transition structure with $\rho = .98$ as the no transition probability. The assignment process is performed without replacement thus producing a block-structured permutation of the original labeled training data set. We generate 10 training data streams at random using this procedure and average the results on the test set. We use a 50-dimensional PCA basis as a pre-processing step for MNIST to reduce the dimension of the data. This is a fully unsupervised operation. The Opportunity data set consists of multiple activity sessions performed by four different individuals in a lab setting. We use the "locomotion" label set ("Stand", "Walk", "Sit", "Lie"). We note that that data set includes missing/other labels, which we discard for these experiments. We learn a separate model for each person and average the results over test sessions for each person. Further details on the session structure an data set sizes for Opportunity can be found in [19]. We initialize the models with 2 labels per class for MNIST and 10 labels per class for opportunity.

Methods: For the MNIST data set, we use the HMM segmenter with simplified transition structure. We use logistic regression and a one-hidden layer SGLD neural network model with 50 hidden units. On the opportunity data set we use logistic regression and a one-hidden layer SGLD neural network model with 10 hidden units. In all cases we use 0.99 as the segment boundary threshold. We use 0.003 as the utility threshold on MNIST and 0.1 on Opportunity. These thresholds were selected so that the methods spread their label budgets over most of the length of the data streams. We use the margin utility in all experiments. We use a total budget of $B = 200$.

Results: The results of these experiments are shown in Figure 1. We can see that the proposed joint segmentation and active learning method out-performs the standard stream-based method up to the limit of 200 labeled data points for both data sets and both classifiers. We can also see that the proposed method is able to obtain a sizeable fraction of the error reduction achieved by the oracle methods (which condition on substantially more information) in all cases. Finally, we note that while the SGLD neural network models obtain improved performance on the MNIST data set relative to LR, this is not the case for Opportunity, which requires further investigation.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a new framework for joint segmentation and active learning that is designed to exploit block structure in data streams. We have explored two instances of this framework based on semi-supervised HMM segmenters combined with logistic regression and Bayesian neural network classifiers. We have presented promising initial results showing that the proposed approach can significantly outperform standard stream-based methods. There are many possible future directions for this work including using the HMM to assign confidence weights to instances within a block to enable soft label propagation, investigating the use of additional utility functions, integrating methods for adapting utility and segmentation thresholds, and investigating the effect of storage budgets for longer or unbounded streams.

6 ACKNOWLEDGEMENTS

This work was supported by grants from the National Science Foundation (IIS-1722792).

REFERENCES

[1] Hande Alemdar, Tim LM van Kasteren, and Cem Ersoy. Using active learning to allow activity recognition on a large scale. In *International Joint Conference on Ambient Intelligence*, pages 105–114. Springer, 2011.

[2] Amin Ahsan Ali, Syed Monowar Hossain, Karen Hovsepian, Md Mahbubur Rahman, Kurt Plarre, and Santosh Kumar. mpuff: automated detection of cigarette smoking puffs from respiration measurements. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, pages 269–280. ACM, 2012.

[3] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[4] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[5] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[6] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[7] Syed Monowar Hossain, Amin Ahsan Ali, Md Mahbubur Rahman, Emre Ertin, David Epstein, Ashley Kennedy, Kenzie Preston, Annie Umbricht, Yixin Chen, and Santosh Kumar. Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In *Proceedings of the 13th international symposium on Information processing in sensor networks*, pages 71–82. IEEE Press, 2014.

[8] Dino Ienco, Albert Bifet, Indrė Žliobaitė, and Bernhard Pfahringer. Clustering based active learning for evolving data streams. In *International Conference on Discovery Science*, pages 79–93. Springer, 2013.

[9] Santosh Kumar, Wendy Nilsen, Misha Pavel, and Mani Srivastava. Mobile health: Revolutionizing healthcare through transdisciplinary research. *Computer*, 46(1):28–35, 2012.

[10] Oscar D Lara and Miguel A Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[13] Christine F Martindale, Sebastijan Sprager, and Bjoern M Eskofier. Hidden markov model-based smart annotation for benchmark cyclic activity recognition database using wearables. *Sensors*, 19(8):1820, 2019.

[14] Tudor Miu, Paolo Missier, and Thomas Plötz. Bootstrapping personalised human activity recognition models using online active learning. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 1138–1147. IEEE, 2015.

[15] Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

[16] Annamalai Natarajan, Gustavo Angarita, Edward Gaiser, Robert Malison, Deepak Ganesan, and Benjamin M Marlin. Domain adaptation methods for improving lab-to-field generalization of cocaine detection using wearable ecg. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 875–885. ACM, 2016.

[17] Annamalai Natarajan, Deepak Ganesan, and Benjamin M Marlin. Hierarchical active learning for model personalization in the presence of label scarcity. In *Proceedings of the IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks*, 2019.

[18] Annamalai Natarajan, Abhinav Parate, Edward Gaiser, Gustavo Angarita, Robert Malison, Benjamin Marlin, and Deepak Ganesan. Detecting cocaine use with wearable electrocardiogram sensors. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 123–132. ACM, 2013.

[19] Hesam Sagha, Sundara Tejaswi Digumarti, José del R Millán, Ricardo Chavarriaga, Alberto Calatroni, Daniel Roggen, and Gerhard Tröster. Benchmarking classification techniques using the opportunity human activity dataset. In *2011 IEEE International Conference on Systems, Man, and Cybernetics*, pages 36–40. IEEE, 2011.

[20] Nazir Saleheen, Amin Ahsan Ali, Syed Monowar Hossain, Hillol Sarker, Soujanya Chatterjee, Benjamin Marlin, Emre Ertin, Mustafa Al'Absi, and Santosh Kumar. puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 999–1010. ACM, 2015.

[21] Akane Sano, Weixuan Chen, Daniel Lopez-Martinez, Sara Taylor, and Rosalind W Picard. Multimodal ambulatory sleep detection using lstm recurrent neural networks. *IEEE journal of biomedical and health informatics*, 2018.

[22] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[23] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27, 1948.

[24] Edison Thomaz, Irfan Essa, and Gregory D Abowd. A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 1029–1040. ACM, 2015.

[25] Yonatan Vaizman, Katherine Ellis, and Gert Lanckriet. Recognizing detailed human context in the wild from smartphones and smartwatches. *IEEE Pervasive Computing*, 16(4):62–74, 2017.

[26] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.