A Maximum Entropy Framework for Semisupervised and Active Learning With Unknown and Label-Scarce Classes

Zhicong Qiu, David J. Miller, *Senior Member, IEEE*, and George Kesidis

***Abstract*— We investigate semisupervised learning (SL) and pool-based active learning (AL) of a classifier for domains with label-scarce (LS) and unknown categories, i.e., defined categories for which there are initially no labeled examples. This scenario manifests, e.g., when a category is rare, or expensive to label. There are several learning issues when there are unknown categories: 1) it is *a priori* unknown which subset of (possibly many) measured features are needed to discriminate unknown from common classes and 2) label scarcity suggests that overtraining is a concern. Our classifier exploits the inductive bias that an unknown class consists of the subset of the unlabeled pool's samples that are atypical (relative to the common classes) with respect to certain key (albeit *a priori* unknown) features and feature interactions. Accordingly, we treat negative log-p-values on raw features as nonnegatively weighted derived feature inputs to our class posterior, with zero weights identifying irrelevant features. Through a hierarchical class posterior, our model accommodates multiple common classes, multiple LS classes, and unknown classes. For learning, we propose a novel semisupervised objective customized for the LS/unknown category scenarios. While several works minimize class decision uncertainty on unlabeled samples, we instead preserve this uncertainty [maximum entropy (maxEnt)] to avoid overtraining. Our experiments on a variety of UCI Machine learning (ML) domains show: 1) the use of p-value features coupled with weight constraints leads to sparse solutions and gives significant improvement over the use of raw features and 2) for LS SL and AL, unlabeled samples are helpful, and should be used to preserve decision uncertainty (maxEnt), rather than to minimize it, especially during the early stages of AL. Our AL system, leveraging a novel sample-selection scheme, discovers unknown classes and discriminates LS classes from common ones, with sparing use of oracle labeling.**

***Index Terms*— Active learning (AL), cross entropy, inductive bias, logistic regression, p-value, rare classes, regularization, semisupervised learning (SL).**

I. Introduction

THE MAIN premise behind active learning (AL) of a statistical classifier is that, by judiciously choosing the samples to be labeled by an oracle (e.g., a human expert or user), one can potentially reduce the number of labeled examples that would otherwise be needed to learn an accurate classifier for the given domain. Such economization is of great practical importance, considering that ground-truth labeling is often an expensive and time-consuming exercise. AL also gives a flexible, user-interactive learning capability for domains where categorizations are highly subjective, e.g., consider learning to predict the music (or art) that an idiosyncratic user would find interesting. AL systems involve a closed-loop learning cycle, with the classifier's model parameters adapted/retrained in light of newly labeled examples, and with the resulting model in turn used to help select additional samples for labeling. The initial model, which starts this process, is generally learned based on a (relatively small) initially available labeled training set. In this paper, we focus on pool-based AL [1] [as well as on semisupervised learning (SL)], where the additional actively labeled samples are chosen from a captured pool of unlabeled samples. A pool-based AL system is fully determined by: 1) the features and associated feature representation used as an input to the classifier; 2) the chosen classifier model structure (e.g., decision tree or logistic regression); 3) the criterion or algorithm used for selecting samples to be labeled [e.g., choosing the sample with greatest class uncertainty (entropy)]; and 4) the training objective function and associated optimization method, used for adapting the classifier's parameters in light of each new labeled instance. We propose an AL scheme novel with respect to all four of these design elements.

Consider a classification domain with category set $\mathcal{C} = \{\Omega_i \in \mathbb{Z}^+, i = 1, \ldots, K\}$. In this paper, we develop novel SL and AL approaches addressing the compelling scenario where some of these categories have very few or even no examples amongst the initial set of labeled training samples.[1] More specifically, given a learning set $\mathcal{X} = \{\mathcal{X}_l, \mathcal{X}_u\}$, with $\mathcal{X}_l$ the labeled training examples and $\mathcal{X}_u$ the unlabeled examples (to be classified), an unknown category is one whose samples exclusively belong to $\mathcal{X}_u$. Likewise, a label-scarce (LS) category is one with very few labeled examples in $\mathcal{X}_l$. Accordingly, we will denote $\mathcal{C}_m = \{\Omega_{m_1}, \Omega_{m_2}, \ldots, \Omega_{m_M}\} \subset \mathcal{C}$ the set of common (majority) classes, with plentiful instances in $\mathcal{X}_l$, $\mathcal{C}_r = \{\Omega_{r_1}, \Omega_{r_2}, \ldots, \Omega_{r_R}\} \subset \mathcal{C}$ the set of LS (rare) categories,

Manuscript received October 26, 2014; revised October 25, 2015 and December 30, 2015; accepted December 31, 2015. Date of publication January 26, 2016; date of current version March 15, 2017. This work was supported in part by Pennsylvania State University, in part by the Project Cisco Systems URP gift, and in part by the Air Force Research Laboratory under Grant FA 8650-11-M1163.

The authors are with the School of Electrical Engineering and Computer Science, College of Engineering, The Pennsylvania State University, University Park, PA 16802 USA (e-mail: zzq101@psu.edu; djmiller@engr.psu.edu; kesidis@engr.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.2016.2514401

[1] In the AL setting, if there are initially no labeled examples from class Ω_k, this category may not even be known to exist for the given domain (until, through AL, one of its examples is chosen for labeling). Thus, when there are unknown categories, AL may also entail new class discovery.

and $\mathcal{C}_u$ the set of unknown categories, whose cardinality is unknown but is upper bounded by $|\mathcal{X}_u|$. Note that there are $M + R$ predefined categories, $\mathcal{C}_m \cap \mathcal{C}_r \cap \mathcal{C}_u = \emptyset$ and $\mathcal{C}_m \cup \mathcal{C}_r \cup \mathcal{C}_u = \mathcal{C}$.

One likely reason for label scarcity for some class is that one of the categories may be rare, i.e., with low probability of occurrence (ϵ). If we suppose that the initial labeled training set was chosen at random based on the class proportions, then, considering one rare category, for a given pair $(|\mathcal{X}_l|, \epsilon)$, it may be quite likely that few, if any, of the rare-category samples are chosen for labeling.[2] As some example rare-category domains, consider the following:

1) Known common and rare disease subtypes (with the latter, e.g., predisposed by rare mutations).
2) Astronomical object categorization, including some rarely observed object types.
3) Discriminating interesting (suspicious) detected anomalies forwarded by an anomaly detector (AD) from uninteresting (innocuous) anomalies [2]–[4]. A very serious practical issue with AD systems (e.g., for network intrusion detection [4]) is that, unless they are very conservatively designed to avoid false positives, they may tend to detect too many anomalies, flooding human operators with alerts. Most of these detections may be (uninterestingly) attributed to sensor failure or miscalibration, rather than a network intrusion attack. We will show that a classifier structure that is natural for discriminating interesting from uninteresting anomalies is in fact a good structure more generally for discriminating common from LS categories.

There are additional reasons for label scarcity for some categories, e.g., even if a category is relatively frequent, its examples may be much more expensive to definitively label, compared with other categories.

Label-Scarce and Unknown Categories:

As aforementioned, when there are no examples of a given category in the initial labeled training set, we refer to such a class as an unknown category. When there are few such labeled instances, we call the class LS. It is important to clearly distinguish the unknown and LS cases, for several reasons. First, some (supervised) classifier learning methods, e.g., support vector machines (SVMs), require at least one labeled example from every category, and thus are unsuitable when there are unknown categories (at least until the needed first labeled instances of the unknown category are identified by AL)—for a two-class problem with one unknown category, the one-class SVM (OC-SVM) [5], [6] could be invoked initially, until the first instance of the unknown category is actively labeled. Then, the AL system could switch to using the standard two-class SVM. Second, Dasgupta and Hsu [7] suggest that the most challenging phase of AL with unknown categories is identifying (and thus, labeling) the very first instance of an unknown category—a naive approach, such as random selection from $\mathcal{X}_u$, may require many labelings before identifying the first such instance. By contrast, once at least one labeled example has been identified for each category, subsequent active labeling requirements, in order to achieve good classification accuracy, may be more modest.

A. Inductive Bias: The Unknown Class as a Subset of What Is Anomalous

The framework that we propose for SL and AL with unknown and LS categories is inspired by statistical AD, and can be thought of as a generalization of (unsupervised) statistical AD. Specifically, for now considering a two-class problem,[3] we will treat the common category (Ω_1) as the null hypothesis (whose model can be learned based on the initial labeled training set of common category examples in $\mathcal{X}_l$), with the unknown (or LS) class denoted by Ω_2. Invoking a standard AD approach, given the learned null model, one would interrogate the unlabeled data batch $\mathcal{X}_u$, seeking to identify the set of anomalous samples—those which deviate most from the null, i.e., those with lowest p-values.[4] The inductive bias [8] that we posit on the common versus unknown (or LS) two-class problem is that the samples that come from the unknown (or LS) class are a (special) subset of the anomalous samples in $\mathcal{X}_u$—those whose atypicalities (low p-values) are with respect to certain key (albeit *a priori* unknown) features or feature combinations. While other inductive bias may be more suitable for learning a classifier when there are plentiful labeled examples of both classes, for our unknown or LS setting, we focus on anomalies with respect to the common class, in learning to discriminate between the common class and an unknown (or LS) class. The effectiveness of this choice will be borne out by our experiments in the sequel.

To exposit further, our inductive bias has three fundamental tenets.

1) A sample originates from the unknown class only if it is anomalous with respect to the common class, i.e., if it is typical of the common class with respect to all features, it should not belong to the unknown class.
2) The samples in the unlabeled batch that come from the unknown class exhibit anomalies with respect to a special (*a priori* unknown) common subset of the measured features, which we dub the informative feature subset.
3) The more atypical a sample is with respect to these informative features, the more likely it is to originate from the unknown class.

The first tenet simply assumes that the two classes are statistically distinct—if so, the unknown class samples should be atypical in some way, referenced to the common class distribution. The second tenet suggests that the anomalies of the unknown class samples are not expected to be random—since they come from the same class, these samples' anomalies are expected to exhibit a pattern, manifesting on a common subset of the measured features. Finally, the third tenet links

[2]The probability of no samples being selected from the rare category is $(1-\epsilon)^{|\mathcal{X}_l|}$. For example, if $\epsilon = 0.01$, this probability is 0.9 for $|\mathcal{X}_l| = 10$ and 0.61 for $|\mathcal{X}_l| = 50$.

[3]We will develop a general multicategory extension in the sequel.

[4]A p-value is the probability of seeing a datum more extreme than the given observation, under the null hypothesis.

the strength of the anomalous pattern to the discriminability of the two classes.

Example: Given a classification domain with a 100-D feature vector $\underline{X} \in \mathbb{R}^{100}$, the unknown class could consist of the samples in $\mathcal{X}_u$ that exhibit anomalies (low p-values) with respect to some subset of the following: 1) the common class's (marginal) distributions for individual features X_{17} and X_{32} and the common class's marginal distributions for the feature pairs (X_{79}, X_{93}), (X_{45}, X_{67}), and (X_{17}, X_{52}).

There are several important observations to make here. First, it should be clear from the above example, with a feature space of high-dimensionality $D = 100$ and a small (*a priori* unknown) subset of informative features, that, in general, an unknown class may not be easily identified without some supervising examples. At the same time, we note that there is prior work on purely unsupervised detection of unknown classes in a data batch [9], [10], e.g., the approach in [10] solely exploits the low likelihood under the null hypothesis that a (sizeable) collection of samples would manifest anomalies with respect to the same subset of features. In any event, in an AL setting, the supervision needed to make the unknown class identifiable/learnable will be oracle-supplied.

Second, note that our inductive bias instructs that a special (nonlinear) feature transformation should be applied to the original, raw features: p-values (on individual features and/or low-order feature combinations), treated as derived features (classifier inputs), are hypothesized to be effective for discriminating between the unknown and common classes.

Third, as illustrated by the above example, it is possible that the unknown class may only conspicuously manifest anomalies (low p-values) with respect to a small subset of the measured features (the informative features).[5] The presence of many uninformative or nuisance features is plausible because, without any initial labeled examples of the unknown class, one does not know *a priori* which features will be needed to discriminate this class from the common class, i.e., one may need to overprovision with features; consequently, many uninformative features may be measured. Likewise, it will also be *a priori* unknown which low-order feature combinations (and associated p-values) will be class-discriminating. Thus, in general, one must agnostically measure all of them, e.g., restricting just to feature pairs to limit complexity, all $(D(D-1)/2)$. Thus, for large D, there may be many marginal and, especially, many pairwise, feature p-values that are uninformative. It is thus incumbent upon AL to figure out which (potentially sparse subset of) individual and low-order feature combination p-values are strongly discriminating. This amounts to a (high dimensional) feature selection problem. Note also that distance-based methods, such as [12], which discriminate on the basis of all (or most) features, without performing feature selection, should be suboptimal when many of the features are uninformative.

Fourth, suppose that we have identified the informative features. The three tenets together imply a constraint should be imposed on the classifier's parameter space. Specifically, as seen in the sequel, negative log p-values should be nonnegatively weighted, with zero weights for irrelevant features (tenet 2), and with larger positive weights for more informative features (tenet 3).

Finally, note that our above assumptions do not necessitate that an unknown class should consist of a compact cluster, in either the full-feature space (as assumed in [12] and [13]) or even in the informative feature subspace. That is, two samples atypical with respect to the same pair of features need not be proximal in the corresponding 2-D feature space. As an example, consider a bivariate Gaussian null, whose equidensity contour is an ellipse. A pair of samples lying on the same ellipse will have the same p-value.[6] However, the two samples could be at the opposite ends of the ellipse along the principal axis, and thus be quite distant from each other. What this suggests is that clustering-based approaches that work in the original feature space using a Euclidean distance measure [12], [13] may not be effective at identifying unknown classes (at least those consistent with our inductive bias). Our use of p-values as derived features can accordingly be considered an (intuitively motivated) alternative to the agnostic use of a (e.g., Gaussian) kernel for achieving a nonlinear mapping of the original features. For some domains, the use of such kernel-based mappings may give poor class discrimination [14].[7]

B. Novel Contributions of This Paper

There are several primary contributions of this paper.

1) To represent the null model for a (in general) high-dimensional feature vector, we use comprehensive low-order distributions defined on individual features and feature pairs, modeling these using Gaussian mixture models (GMMs), which facilitates measuring mixture-based p-values.
2) Our proposal of a novel (p-value-based) feature representation and a companion class posterior model that together encode our inductive bias both for SL and for AL in the presence of unknown and LS categories. This model handles multiple common, multiple LS, and unknown classes. When there are rare categories, this model will be shown to outperform conventional classifiers that have similar representation power and which are trained in the same way as our model, but which work on raw features rather than p-values.
3) A novel SL framework with built-in capability for avoiding overtraining, as especially needed for SL and AL

[5]The common and unknown classes may have very similar distributions for all other features. Alternatively, both common and unknown class samples in the data batch may manifest anomalies with respect to these other features. Either way, these other features will not have much power to discriminate the two classes and are thus uninformative. Moreover, considering finite sample sizes available for classifier training, it is well known that the use of features with weak discrimination power can compromise classification accuracy [11], i.e., these uninformative features may effectively be nuisance features.

[6]The p-value is obtained by integrating the bivariate null density over the region defined by the exterior of the ellipse.

[7]The Gaussian kernel, for example, implicitly defines generalized coordinates that include product features (and, thus, will combine informative and irrelevant features through their products). In this way, the contribution to the classifier from informative features may be distorted or masked by the use of the kernel.

with unknown (and LS) categories. Unlike [15], our approach preserves, rather than minimizes decision uncertainty, and will be shown to give superior results for SL and AL under the unknown and LS class scenarios, especially during the early stages of AL.

4) An AL framework that exploits the first two contributions and employs a novel sample selection mechanism for active labeling focused on new class discovery. Our AL scheme adapts the classifier model to accommodate each new class discovered through the labeling process of AL. Our AL system discovers novel classes and achieves good classification accuracy with sparing use of oracle labeling. This approach will be shown to outperform several conventional benchmark methods, including SVM-based AL.

The rest of this paper is organized as follows. In Section II, we introduce our proposed null distribution modeling, which comprehensively evaluates low-order (marginal and pairwise feature) p-values. Section III introduces our proposed (hierarchical) class posterior model, based on negative log p-value derived features. Section IV develops our SL objective function, which is tailored for SL and AL with unknown (or LS) classes. Section V introduces the sample selection criteria we used for AL. Section VI presents comprehensive experiments on a variety of UCI domains. Section VII discusses related work. Finally, conclusions are drawn in Section VIII.

II. Comprehensive Low-Order Null Modeling

Suppose that we are given a training set of known common (i.e., null hypothesis) class (Ω_1) examples $\tilde{\mathcal{X}} = \{\underline{\tilde{x}}^{(i)}, i = 1, \ldots, T\}$, $\underline{\tilde{x}}^{(i)} \in \mathbb{R}^D$, from which we learn our probability model under the null.[8] One possibility is to learn the joint density function for $\underline{X} \in \mathbb{R}^D$ under Ω_1; then, for a given observed vector $\underline{x} \in \mathcal{X}_u$, one would measure a single p-value with respect to this joint density. However, there are several issues here. First, if D is large, the training database $\tilde{\mathcal{X}}$ may need to be huge in order to accurately estimate the joint density, i.e., there is a curse of dimensionality [16].[9] This statement may hold even if one invokes strategies for limiting how the model size grows as a function of D [17]–[19], which efficiently model the covariance matrices for GMM-based joint densities. Second, the anomalous signature may be too weak (with the p-value too large) for large D. For example, suppose that the features are statistically independent under the null, and that a sample $\underline{x}$ exhibits an anomaly only for one (or very few) of the D features. In this case, the joint log-likelihood for $\underline{x}$ under the null is the sum of the marginal (single feature) log-likelihoods, with the effect of a single (anomalous) feature on the joint log-likelihood (and, thus, on the joint p-value) averaged out for increasing D.

To avoid both these problems, we instead propose to represent the null (and assess p-values) using low-order distributions defined on subvectors of $\underline{X}$. Since it is *a priori* unknown which individual features and feature combinations may be informative about the unknown class, we must be unbiasedly comprehensive, considering all of them. Note that there are $\binom{D}{k} = D!/(k!(D-k)!)$ such distributions at order k; even assuming that the estimation of these distributions is performed offline, the highest order models that may be computationally practicable to learn (while at least capturing statistical dependence between pairs of features) is $k = 2$. Accordingly, assuming large D, we propose to estimate all D marginal densities $\{f_{X_i}(\cdot), i = 1, \ldots, D\}$ and all $D(D-1)/2$ pairwise densities $\{f_{X_i,X_j}(\cdot), i, j = 1, \ldots, D, i < j\}$. Moreover, since distributions under the null will, in general, be complicated and multimodal, we will apply a widely invoked approach, modeling each marginal and pairwise feature density as a finite GMM. We learn each such model separately via the expectation-maximization (EM) algorithm [20], with the number of components chosen to minimize the Bayesian information criterion (BIC) [21] model order selection objective.[10] While this learning is computationally heavy, in practice it can often be performed off-line prior to AL. Moreover, this GMM learning is also highly parallelizable.

Mixture-Based p-Values as Derived Features

Consider any pair of features $\underline{Y} = (X_i, X_j)$ modeled by a bivariate GMM under the null. Let $\{\alpha_l, l = 1, \ldots, L_{ij}\}$, $0 \leq \alpha_l \leq 1$, $\sum_{l=1}^{L_{ij}} \alpha_l = 1$, be the prior probabilities for the (L_{ij}) mixture components, with associated component densities $f_{\underline{Y}|l}(\underline{y}|\theta_l)$, where $l = 1, \ldots, L_{ij}$ and $\theta_l = (\underline{\mu}_l, \Sigma_l)$ the (mean vector, covariance matrix) parameter set for the lth density. The mixture density is thus $f_{\underline{Y}}(\underline{y}) = \sum_{l=1}^{L_{ij}} \alpha_l f_{\underline{Y}|l}(\underline{y}|\theta_l)$. Given such a mixture null, we would like to calculate the p-value—the probability that a 2-D feature vector will be more extreme than the given observed vector $\underline{y} = (x_i, x_j)$. This has been previously considered in [10], with the result dubbed mixture-based p-values. First, consider a single bivariate Gaussian density $\mathcal{N}(\underline{\mu}, \Sigma)$. We call a vector $\underline{y}'$ more extreme

[8] For the unknown class case, $\tilde{\mathcal{X}} = \mathcal{X}_l$. For the LS case, $\tilde{\mathcal{X}}$ is the subset of $\mathcal{X}_l$ that originates from the common classes.

[9] This problem is exacerbated by the potential of finding poor local optima of the (generally nonconvex) likelihood objective function that is maximized in learning the joint density function.

[10] Separately learning each marginal and pairwise feature GMM using the common training set $\tilde{\mathcal{X}}$ will not ensure consistency with respect to feature marginalizations. Specifically, a marginal-consistent collection of univariate and bivariate density functions should satisfy the following: if we consider any feature pairs (i, j) and (j, k), marginalizing out feature i from the (i, j) bivariate density and marginalizing out feature k from the (j, k) bivariate density should lead to the same marginal density for feature j. However, when the univariate and bivariate distributions are Gaussian mixtures, with a nonconvex log-likelihood function (and with BIC-based model order selection separately applied to choose the number of components for each GMM), the separate application of EM-plus-BIC to learn each GMM density function does not ensure a set of marginal-consistent distributions. This property is not centrally important here, however, since our main concern is only to learn marginal and pairwise density functions that allow accurate assessment of p-values. Accordingly, in this paper, we will apply EM-plus-BIC separately to learn each low-order GMM.

One approach to obtain marginal-consistent low-order distributions is to simply learn the single GMM for the joint distribution on the full feature vector, $\underline{X}$. This determines (via marginalization) all lower order distributions (which are also GMMs, and which are guaranteed to be marginal-consistent). However, if available data are limited, this strategy suffers from the curse of dimensionality [16]. Alternatively, it is possible to directly, jointly learn a marginal-consistent set of low-order GMMs.

than $\underline{y}$ if it lies on a lower equidensity (elliptical) contour than $\underline{y}$. Accordingly, the p-value for $\underline{y}$ denoted by $p_{ij}(\underline{y})$ is the integral of the bivariate density over the exterior of the ellipse defined as the squared Mahalanobis distance from $\underline{y}$ to $\underline{\mu}$, i.e., $r^2(\underline{y}) = (\underline{y} - \underline{\mu})'\Sigma^{-1}(\underline{y} - \underline{\mu})$. This integral can be calculated exactly by applying a whitening transformation [22], leading to the result that the p-value is $p_{ij}(\underline{y}) = e^{-r^2(\underline{y})/2}$, i.e., it is one minus the Rayleigh cumulative distribution function evaluated at $r(\underline{y})$. Extending to a GMM by applying the law of total probability and conditioning on the mixture component of origin, we find that the p-value for $\underline{y} = (x_i, x_j)$ is $p_{ij}(\underline{y}) = \sum_{l=1}^{L_{ij}} P[M = l|\underline{y}]e^{-r_l^2(\underline{y})/2}$. Here, the mixture posterior is $P[M = l|\underline{y}] = \alpha_l f_{\underline{Y}|l}(\underline{y}|\theta_l)/(\sum_{k=1}^{L_{ij}} \alpha_k f_{\underline{Y}|k}(\underline{y}|\theta_k))$, and $r_l^2(\underline{y})$ is the squared Mahalanobis distance between $\underline{y}$ and $\underline{\mu}_l$ for the lth Gaussian Mixture component. Note that $p_{ij}(\underline{y})$ is the expected p-value, with the expectation taken with respect to the mixture posterior probability mass function (pmf). In a similar fashion, one can also calculate the mixture-based p-values for single (univariate) features, denoted by $p_i(x_i), i = 1, \ldots, D$. In this case, complementary error functions are used to measure the p-value conditioned on each mixture component, with the mixture-based p-value again the expected p-value.

III. Classifier Model

A. Two Class Case

For a given feature vector $\underline{x} \in \mathbb{R}^D$, define the set of low-order (1-D and 2-D) mixture p-values $\mathcal{P}(\underline{x}) = \{p_{ij}(x_i, x_j), i, j = 1, \ldots, D, i < j\} \cup \{p_i(x_i), i = 1, \ldots, D\}$.[11] Our AL approach can be applied for various class posterior models. Here, it is developed assuming a nearly standard logistic regression class posterior form, albeit with negative logs of the p-values $\mathcal{P}(\underline{x})$ as derived feature inputs and with nonnegative weight constraints. Specifically, for a two-class problem, with one common class (Ω_1) and one unknown or LS class (Ω_2), let

$$f(\underline{x};\Lambda) = \exp\left(w_0 - \sum_{i=1}^{D} w_i \log p_i(x_i) - \sum_{j>i} \beta_{ij} \log p_{ij}(x_i, x_j)\right)$$

where the model parameters are $\Lambda = \{w_0, \{w_i\}, \{\beta_{ij}\}\}$. We then have

$$P(\Omega_2|\mathcal{P}(\underline{x});\Lambda) = \frac{f(\underline{x};\Lambda)}{1 + f(\underline{x};\Lambda)}, \quad w_i \geq 0, \quad \beta_{ij} \geq 0 \quad \forall i, j$$
$$P(\Omega_1|\mathcal{P}(\underline{x});\Lambda) = 1 - P(\Omega_2|\mathcal{P}(\underline{x});\Lambda). \tag{1}$$

We can make the following observations about this model.

1) If we initialize at $w_1 = w_2 = \cdots = w_D = w > 0$ and set $\beta_{ij} = \beta > 0$, $\forall i, j$, then ordering samples by their *a posteriori* probability of belonging to the unknown class is equivalent to ordering them by their aggregate atypicality (the sum of the arithmetic averages of their first-and second-order log p-values), i.e., initially, before there are any supervising examples, the model is unbiased about which features/combinations are informative, treating the most anomalous sample as the one most likely to come from the unknown class. Thus, before there are supervising examples, our model is akin to a conventional AD.
2) We enforce nonnegative weights on negative log p-values (which are positive), consistent with our inductive bias that unknown is a subset of what is anomalous with respect to Ω_1—no increase in a feature's atypicality should reduce the probability of a sample belonging to the unknown class. However, if a single feature's (i) or a feature pair's ((i, j)) atypicality is irrelevant to common–unknown discrimination, this can be properly reflected by setting $w_i = 0$ or $\beta_{ij} = 0$, respectively. In fact, as will be seen in the sequel, our approach learns highly sparse models, with a very few nonzero weights. Thus, once learned, our model is very different from a conventional AD.
3) *Monotonic Sample Ordering (Generalization) Property:* Suppose that the classifier parameters have been chosen so that $P(\Omega_2|\mathcal{P}(\underline{x}^{(n)})) < 0.5$, such that sample n is maximum *a posteriori* (MAP)-classified to Ω_1. Now, consider any other sample $\underline{x}^{(m)}$, such that $p_i(x_i^{(m)}) \geq p_i(x_i^{(n)}), \forall i = 1, \ldots, D$ and $p_{ij}(x_i^{(m)}, x_j^{(m)}) \geq p_{ij}(x_i^{(n)}, x_j^{(n)}), \forall i, j$. Clearly, $P(\Omega_2|\mathcal{P}(\underline{x}^{(m)})) \leq P(\Omega_2|\mathcal{P}(\underline{x}^{(n)})) < 0.5$, i.e., if we have learned to MAP-assign sample n to Ω_1, we have in fact also learned to MAP-assign all samples more typical than n under the null to Ω_1. The same also holds in the alternative case, supposing that $\underline{x}^{(n)}$ is MAP-assigned to Ω_2 and considering the set of samples more atypical than $\underline{x}^{(n)}$. This is a type of generalization that comes from constraining the weights (on derived features that are positive) to be nonnegative.

B. Extension for Multiple Common and Label-Scarce Classes

Suppose, now, that there are multiple common classes $\Omega_{m_i} \in \mathcal{C}_m, i \in \{1, \ldots, M\}$ and multiple LS classes $\Omega_{r_j} \in \mathcal{C}_r$, $j \in \{1, \ldots, R\}$. In order to generalize our two-class approach to handle multiple common and LS classes, we propose a hierarchical class posterior, defined on all common and LS classes, that is

$$P[C = \Omega_{r_j}|\underline{x}] = P[C \in \mathcal{C}_r|\underline{x}]P[C = \Omega_{r_j}|\mathcal{C}_r, \underline{x}] \tag{2}$$

and

$$P[C = \Omega_{m_i}|\underline{x}] = P[C \in \mathcal{C}_m|\underline{x}]P[C = \Omega_{m_i}|\mathcal{C}_m, \underline{x}]. \tag{3}$$

Here, the top-level posterior is the two class subset ($\{\mathcal{C}_r, \mathcal{C}_m\}$) posterior, that is

$$P[C \in \mathcal{C}_r|\underline{x};\Lambda] = \frac{f(\underline{x};\Lambda)}{1 + f(\underline{x};\Lambda)}, \quad w_i \geq 0, \quad \beta_{ij} \geq 0 \quad \forall i, j$$

and

$$P[C \in \mathcal{C}_m|\underline{x};\Lambda] = 1 - P[C \in \mathcal{C}_r|\underline{x};\Lambda]. \tag{4}$$

The second-level posteriors $P[C = \Omega_{r_j}|\mathcal{C}_r, \underline{x}]$ and $P[C = \Omega_{m_i}|\mathcal{C}_m, \underline{x}]$ divide up the top-level probabilities $P[C \in \mathcal{C}_r|\underline{x}]$

[11] We restrict to first- and second-order p-values in order to ensure the scalability of the method with increasing D.

and $P[C \in \mathcal{C}_m|\underline{x}]$ amongst $\Omega_{r_j} \in \mathcal{C}_r$ and $\Omega_{m_i} \in \mathcal{C}_m$, respectively.

The second-level posterior for the LS classes, $P[C = \Omega_{r_j}|\mathcal{C}_r, \underline{x}]$, is of the multilogit form and can be defined either in raw feature space or in p-value space, e.g., for raw feature space, let

$$f(\underline{x}; A^j) = \exp\left(a_0^{(j)} + \sum_{k=1}^{D} a_k^{(j)} x_k + \sum_{l>k} a_{kl}^{(j)} x_k x_l\right)$$

where the model parameters are $A^j = \{a_0^{(j)}, \{a_k^{(j)}\}, \{a_{kl}^{(j)}\}\}$. We then have, $\forall j \in \{1, \ldots, R\}$

$$P[C = \Omega_{r_j}|\mathcal{C}_r, \underline{x}; A^j] = \frac{f(\underline{x}; A^j)}{\sum_{\Omega_{r_{j'}} \in \mathcal{C}_r} f(\underline{x}; A^{j'})}. \quad (5)$$

Experimentally, we have found that while using p-values in the top level (to discriminate between the common and LS class subsets) significantly outperforms the use of raw features, there are only small differences in performance resulting from using p-values versus raw features for the second-level posterior to discriminate between the LS classes. This is not so surprising, since our inductive bias is really associated with the top-layer posterior (discriminating the common class subset from the rare class subset).

The second-level posterior for the known classes $P[C = \Omega_{m_i}|\mathcal{C}_m, \underline{x}]$ is also of the multilogit form and works in the raw feature space, i.e., for the ith common class, let

$$f(\underline{x}; B^i) = \exp\left(b_0^{(i)} + \sum_{k=1}^{D} b_k^{(i)} x_k + \sum_{l>k} b_{kl}^{(i)} x_k x_l\right)$$

where the model parameters are $B^i = \{b_0^{(i)}, \{b_k^{(i)}\}, \{b_{kl}^{(i)}\}\}$. We then have, $\forall i \in \{1, \ldots, M\}$

$$P[C = \Omega_{m_i}|\mathcal{C}_m, \underline{x}; B^i] = \frac{f(\underline{x}; B^i)}{\sum_{\Omega_{i'} \in \mathcal{C}_m} f(\underline{x}; B^{i'})}. \quad (6)$$

Note that both second-level posteriors are valid pmfs, summing to one over their respective class subsets, with the hierarchical posterior summing to one over all classes. Note also that, unlike the top-level posterior, there are no constraints on values for parameters of the bottom-level posterior pmfs.

C. Extension to Include Unknown Classes

In the AL setting, in addition to $\mathcal{C}_m$ and $\mathcal{C}_r$, we have the unknown class subset, $\mathcal{C}_u$, consisting of an (in general) unknown number of unknown classes, which need to be discovered through AL. Once a new class is discovered via AL sample selection and oracle labeling, this category (now with one labeled instance) is moved from $\mathcal{C}_u$ to the LS class subset, $\mathcal{C}_r$. Thus, the subsets $\mathcal{C}_u$ and $\mathcal{C}_r$ change as a function of oracle labelings. To reflect this, we use the superscript (t), denoting $\mathcal{C}_u^{(t)}$ and $\mathcal{C}_r^{(t)}$ after the tth oracle labeling. For concise notation in representing the event that a sample belongs to the unknown class subset, we assign (the unused) class $C = 0$, i.e., we let $C \in \mathcal{C}_u \Leftrightarrow C = 0$. Letting $\mathcal{C}_{\text{ru}}^{(t)}$ denote $\mathcal{C}_r^{(t)} \cup \mathcal{C}_u^{(t)}$, the top-level posterior for $\mathcal{C}_{\text{ru}}^{(t)}$ and $\mathcal{C}_m$ is of the same form as (4), with $P[C \in \mathcal{C}_{\text{ru}}^{(t)}|\underline{x}]$ and $P[C \in \mathcal{C}_m|\underline{x}]$, respectively. However, we modify the bottom-level posterior for $\mathcal{C}_r$ as follows, to allow inference on the unknown class subset. We have

$$P\left[C = \Omega_{r_j}|C_{\text{ru}}^{(t)}, \underline{x}; A^j\right] = \frac{f(\underline{x}; A^j)}{1 + \sum_{\Omega_{r_{j'}} \in \mathcal{C}_r^{(t)}} f(\underline{x}; A^{j'})} \quad \forall j \in \{1, \ldots, R\} \quad (7)$$

and for $C \in \mathcal{C}_u^{(t)}$,

$$P\left[C = 0|C_{\text{ru}}^{(t)}, \underline{x}; A^j\right] = \frac{1}{1 + \sum_{\Omega_{r_{j'}} \in \mathcal{C}_r^{(t)}} f(\underline{x}; A^{j'})}. \quad (8)$$

Note that (7) and (8) define a valid pmf, i.e., $P[C = 0|C_{\text{ru}}^{(t)}, \underline{x}; A^j] + \sum_{j=1}^{R} P[C = \Omega_{r_j}|C_{\text{ru}}^{(t)}, \underline{x}; A^j] = 1$. Note also that, at the outset of AL, if there are no LS classes, all probability is assigned to the unknown class ($C = 0$). As the cardinality of $\mathcal{C}_r^{(t)}$ grows during AL, the probability of the unknown category ($C = 0$) tends to diminish. To control the amount of contribution from the unknown class, one can further replace the value 1 in both (7) and (8) with a hyperparameter, and tune its value based on cross validation (CV) or a related measure. We leave consideration of this for future work.

IV. Learning Objective Function for Semisupervised and Active Learning

In this section, we focus on classifier learning, developing a novel SL objective well suited both for SL and AL with unknown (and LS) classes. Our principal focus, in devising a learning framework for these scenarios, is to make appropriate, effective use of the (assumed to be many) unlabeled samples in the data batch ($\mathcal{X}_u$), in addition to the current cadre of labeled samples ($\mathcal{X}_l$). In this section, let $\mathcal{I}_l$ denote the sample index set for the labeled samples, $\mathcal{X}_l$, with $\mathcal{I}_u$ the index set for the unlabeled samples.

If we were to only make use of the labeled samples, a standard (supervised) approach for learning a class posterior model is maximization of the posterior log-likelihood over the labeled training samples [16]. For our purposes, it is convenient to exploit the equivalence between posterior log-likelihood maximization and minimization of a sum of cross entropies objective function, specified as follows. For a given labeled feature vector instance $\underline{x}_n$ from class c_n, we define the target distribution vector

$$Q[C|n] = (Q[\Omega_1|n], \ldots, Q[\Omega_{|\mathcal{C}|}|n]) = (0, 0, \ldots, 1, 0, \ldots, 0)$$

with the 1 in the position of $c_n \in \{\Omega_1, \ldots, \Omega_K\}$. Then, after the tth oracle labeling of AL, maximizing the log-likelihood of the posterior model $P[C|\underline{x}]$ given in (2), (3) is equivalent to minimizing

$$D^{(t)} = \sum_{n \in \mathcal{I}_l^{(t)}} d(Q[C|n] \| P[C|\underline{x}_n]) = -\sum_{\Omega \in \mathcal{C}_m \cup \mathcal{C}_r^{(t)}} \sum_{n \in \mathcal{I}_l^{(t)}: c(n) = \Omega} \log P[\Omega|\underline{x}_n] \quad (9)$$

where $d(Q\|P) = \sum_k q_k \log(q_k/p_k)$ is the Kullback–Leibler distance [23] (cross entropy) between pmfs $Q = \{q_k\}$ and $P = \{p_k\}$ defined on the same support, with Q the target distribution and P its (model-constrained) approximation. Note that in the AL case, the labeled index set $\mathcal{I}_l$, the LS class set $\mathcal{C}_r$, and the unknown class set $\mathcal{C}_u$ are functions of time. $\mathcal{I}_l$ grows with each oracle labeling, while $\mathcal{C}_r$ grows when a new unknown class is discovered through active labeling.[12] Note further that we do not consider here the case where a rare class Ω_{r_j} may transition (after AL has produced a sufficient number of labeled instances from this class) from $\mathcal{C}_r$ to $\mathcal{C}_m$.[13]

We seek to modify (9) in two respects: 1) primarily, to sensibly exploit for learning the (typically large) current subset of unlabeled samples, $\mathcal{X}_u^{(t)}$ and 2) to also account for large imbalance between the number of labeled samples from common classes versus LS classes. Let us consider the first goal. In [15], [24], and [25], SL objective functions were proposed by adding, to the purely supervised training objective function, a cost term which penalizes posterior solutions with high class uncertainty (i.e., high class entropy) on the unlabeled sample subset. Specifically, following the approach taken in [15], one would modify the above training objective function to add an entropy regularization term:

$$D_{\min-H}^{(t)} = D^{(t)} + \mu \sum_{n \in \mathcal{I}_u^{(t)}} H(C|n) \tag{10}$$

where

$$H(C|n) = -\sum_{\Omega \in \mathcal{C}_m \cup \mathcal{C}_{\text{ru}}^{(t)}} P[C = \Omega|\underline{x}_n] \log P[C = \Omega|\underline{x}_n]$$

and with μ a nonnegative hyperparameter, giving the weight on the entropy term.[14] This general approach, together with a weight decay term to avoid overfitting, has been dubbed minimum entropy (minEnt) regularization, and has been motivated from the principle of making decisions with confidence [15]. A related approach given in [26] seeks to locally minimize class overlap. For the scenario considered here, however, with a deficient amount of labeled training data for some of the classes, it appears that the pivotal learning issue is how much training of the classifier's model parameters one should do, commensurate with the available labeled data, to glean as much information as possible from the labeled examples but while avoiding overtraining. The entropy penalty term in (10) will, in general, give a tendency for more learning, rather than less. This can be understood as follows. Consider the solution to the purely supervised posterior log-likelihood maximization problem, i.e., minimizing (9). Suppose that this solution achieves a level of unlabeled subset class entropy $H_0 = \sum_{n \in \mathcal{I}_u^{(t)}} H(C|n)$. Clearly, since the semisupervised modification of (9) positively penalizes this class entropy, minimizing (10) should yield solutions with lower class entropy than those which minimize (9). However, such reduction in class entropy is generally achieved by increasing the magnitudes of the classifier's weights on the features as given in the posterior, i.e., by performing more learning. But this gives propensity for overtraining. In fact, in [15], [24], and [25], there is some hedging on the minEnt solution—Grandvalet and Bengio [15] imposed a smoothness constraint on the weight vector, while Fujino *et al.* [24] introduced an explicit, additional weight vector norm ($L2$) regularization term. Regardless, in both of these approaches, the impetus in the cost function coming from the unlabeled samples is to seek solutions with minimum class entropy.

By contrast, rather than penalizing solutions with high class decision uncertainty on the unlabeled samples, we penalize solutions with low decision uncertainty, as they may be consistent with possible overtraining, especially during the early stages of AL and/or in the SL case, given LS classes, i.e., we propose maximum entropy (maxEnt), rather than minEnt regularization.[15]

Tikhonov-style regularization, such as $L2$ or lasso, by imposing a prior belief on the parameter space, also limits overtraining. However, such a universal prior has no dependence on (is unaffected by) unlabeled samples. By contrast, our maxEnt regularizer is unlabeled-sample-dependent, seeking to keep the classifier as noncommittal as possible in the regions of the feature space solely occupied by unlabeled samples. An example highlighting the difference between maxEnt and $L2$ regularization will be given in the sequel.

There are different ways that maxEnt regularization can be mathematically formulated in practice. We effect such regularization by choosing the same (cross entropy) cost function for the unlabeled samples as for the labeled samples, albeit with a different choice for the target distribution. For every unlabeled sample, $n \in \mathcal{I}_u^{(t)}$, we introduce maxEnt target distributions for both the top-level and bottom-level posteriors.

Denote $S(\Omega)$ as the class subset ($\mathcal{C}_{\text{ru}}$ or $\mathcal{C}_m$) to which Ω belongs. For the top level, we choose the uniform target distribution

$$Q[S^{(t)}(C)|n] = \left(\frac{1}{2}, \frac{1}{2}\right).$$

Likewise, for the bottom level, we choose uniform target distributions. In the AL case, where we may have both $\mathcal{C}_r^{(t)}$ and $\mathcal{C}_u^{(t)}$, we choose

$$Q\left[C|C \in C_{\text{ru}}^{(t)}, n\right] = \left(\frac{1}{|\mathcal{C}_r^{(t)}| + 1}, \Omega \in \left\{0, \mathcal{C}_r^{(t)}\right\}\right)$$

$$Q[C|C \in \mathcal{C}_m, n] = \left(\frac{1}{|\mathcal{C}_m|}, \Omega \in \mathcal{C}_m\right).$$

In the SL case, if we only have $\mathcal{C}_r$, we choose

$$Q[C|C \in \mathcal{C}_r, n] = \left(\frac{1}{|\mathcal{C}_r|}, \Omega \in \mathcal{C}_r\right)$$

$$Q[C|C \in \mathcal{C}_m, n] = \left(\frac{1}{|\mathcal{C}_m|}, \Omega \in \mathcal{C}_m\right).$$

[12] In the SL case, the superscript (t) should be omitted from these sets because, in this case, $\mathcal{X}_l$ and $\mathcal{X}_u$ remain fixed.

[13] For our experiments in the sequel, with the maximum number of oracle labelings set to 100, rare categories, in general, remain rare during our AL.

[14] The value of the hyperparameter can be chosen, e.g., so as to minimize a cross-validated classification error rate measure.

[15] In the latter stages of AL, when there are sufficient labels from all classes, minimizing class uncertainty on the unlabeled samples may be a good choice. However, in the early AL stages, a main concern should be avoiding overtraining.

That is, at the top level, target probability is equally apportioned between $C_{\text{ru}}^{(t)}$ and $\mathcal{C}_m$ and, at the bottom level, given a class subset, equally apportioned amongst the classes belonging to that subset.[16] Consistent with these target distributions, we have the following maxEnt regularization costs (exposited for the AL case, where there are unknown classes). At the top level

$$R_{\text{top}}^{(t)} = \sum_{n \in \mathcal{I}_u^{(t)}} d(Q[S^{(t)}(C)|n] \| P[S^{(t)}(C)|\underline{x}_n])$$

where $P[S^{(t)}(C)|\underline{x}] = (P[C_{\text{ru}}^{(t)}|\underline{x}], P[\mathcal{C}_m|\underline{x}])$, and at the bottom level

$$R_{\text{bot}}^{(t)} = \sum_{n \in \mathcal{I}_u^{(t)}} \left(P\left[C_{\text{ru}}^{(t)}|\underline{x}_n\right] \cdot d_{\text{QP}}^r + P[\mathcal{C}_m|\underline{x}_n] \cdot d_{\text{QP}}^m \right)$$

where

$$d_{\text{QP}}^r = d\left(Q\left[C|C \in C_{\text{ru}}^{(t)}, n\right] \| P\left[C|C \in C_{\text{ru}}^{(t)}, \underline{x}_n\right]\right)$$
$$d_{\text{QP}}^m = d(Q[C|C \in \mathcal{C}_m, n] \| P[C|C \in \mathcal{C}_m, \underline{x}_n]).$$

Note that the bottom-level regularizer weighs each of the cross entropy terms by the probability that the sample originates from that class subset. Clearly, minimizing the costs $R_{\text{top}}^{(t)}$ and $R_{\text{bot}}^{(t)}$ will encourage entropy maximization on the unlabeled samples. Accordingly, our proposed semisupervised objective function is a weighted sum of cross entropies over both the labeled and unlabeled data subsets, specified as follows[17]:

$$D_{\text{max}-H}^{(t)} = -\sum_{\Omega \in \mathcal{C}_m \cup \mathcal{C}_r^{(t)}} \sum_{n \in \mathcal{I}_l^{(t)}: c(n) = \Omega} \alpha_\Omega^{(t)} \log P[\Omega|\underline{x}_n] + \alpha_{\text{top}}^{(t)} R_{\text{top}}^{(t)} + \alpha_{\text{bot}}^{(t)} R_{\text{bot}}^{(t)}. \quad (11)$$

Note that, in general, this cost function is nonconvex in its parameters—this is due to the regularization term $R_{\text{bot}}^{(t)}$. However, in the special case where there is a single known class and a single unknown (or LS) class (and, thus, no bottom-layer regularization term $R_{\text{bot}}^{(t)}$), the cost function is in fact convex. The objective (11) can be minimized by various optimization techniques, e.g., projected gradient or interior point methods, where the descent direction can either be the gradient or Newton's direction. In our experiments, (11) is minimized, after the tth oracle labeling, via a gradient descent procedure, which we have found to be effective. We applied the projected gradient descent method, i.e., at each step, we update each parameter in its steepest descent direction. For parameters constrained to be nonnegative, we project their values to 0 if the gradient-updated value goes negative. Inexact line search is used to select the learning rate (common to all parameters) at each iteration to ensure that each gradient update satisfies the Wolfe condition [27]. For each gradient update, with K known classes, N samples, and D raw features, the computational complexity of the gradient is $O(\text{KND}^2)$.

[16]Note also that in the special case where there is a single common class and a single unknown class, there is only the top level, with target distribution $(1/2, 1/2)$.

[17]Again, in the SL case, the superscript (t) should be omitted, since in this case, the labeled and unlabeled data subsets are fixed.

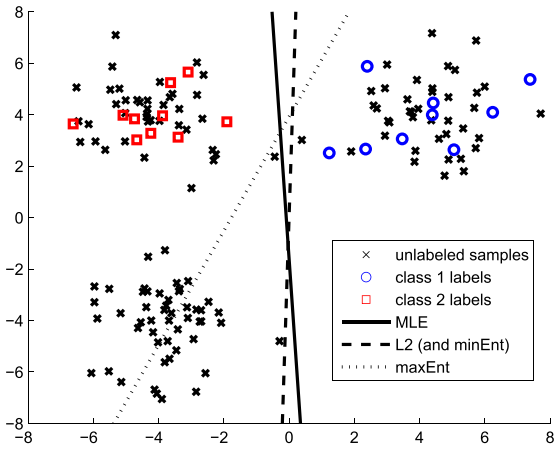


Fig. 1. Three-cluster example with decision boundaries trained by different objectives. MLE: maximum likelihood estimate. $L2$: MLE with $L2$-norm. minEnt: minimum entropy with weight decay ($L2$) [15]. maxEnt: maximum entropy proposed.

To start the learning, the weights $\Lambda, \{A^j\}$, and $\{B^i\}$ were (unbiasedly) initialized to the same small value $\sim 10^{-6}$, except for $w_0, \{a_0^{(j)}\}$, and $\{b_0^{(i)}\}$, which were initialized to zero.

A second important issue is how to choose the weights $\alpha_{\text{top}}^{(t)}$, $\alpha_{\text{bot}}^{(t)}$, and $\alpha_\Omega^{(t)}, \forall \Omega \in \mathcal{C}_r^{(t)} \cup \mathcal{C}_m$. We propose to choose the weight $\alpha_\Omega^{(t)}$ to balance the contribution to the objective function coming from the common and LS class subsets. That is, we let $\alpha_\Omega^{(t)} = 1, \forall \Omega \in \mathcal{C}_m$ and $\alpha_\Omega^{(t)} = (|n \in \mathcal{I}_l : c(n) \in \mathcal{C}_m| / |n \in \mathcal{I}_l : c(n) \in \mathcal{C}_r^{(t)}|), \forall \Omega \in \mathcal{C}_r^{(t)}$. It appears more difficult to find a principled, universal prescription for choosing $\alpha_{\text{top}}^{(t)}$ and $\alpha_{\text{bot}}^{(t)}$—the right level of class decision uncertainty at the tth labeling step may depend, in a complicated way, on the sizes of the labeled and unlabeled data subsets, on K, and could also be domain-dependent. We suggest in practice to set these hyperparameters to minimize (over a finite grid of candidate values) a cross-validated error rate measure, based on the labeled samples seen until now.[18] The choice of these hyperparameters is further discussed in Section VI.

While the design of our SL objective (11) and its maxEnt regularization are largely motivated to avoid overlearning, it is fundamentally different from Tikhonov-style regularization, such as $L2$. To illustrate this and also compare with other objectives, we synthetically generated a three-cluster example shown in Fig. 1. We labeled ten samples from each of classes 1 and 2, treating the rest of the samples as unlabeled. Here, we want to discriminate between classes 1 and 2, and make inference on all the unlabeled samples. Because the classification task involves two common categories and no rare categories, our hierarchical posterior reduces to a two-class logistic regression, using raw features as input, i.e., no top-level posterior and with bottom-level posteriors solely on the two common classes. We trained different objective functions and plotted the resulting optimized decision boundaries. Note that MLE, $L2$, and maxEnt are all convex functions, while

[18]Again, in the SL case, the superscript (t) on α_Ω, α_{top}, and α_{bot} should be omitted, because in this case, the unlabeled and labeled data subsets are fixed.

minEnt is not. As shown in Fig. 1, all the objectives are able to confidently discriminate between classes 1 and 2. However, all objectives except maxEnt classify the unknown cluster (lower left) samples to class 2. On the other hand, maxEnt places the decision boundary in the middle of this high density region with no labeled samples, reflecting minimum classifier confidence (a maxEnt posterior) in this region—this solution makes sense, since this is a region with purely unlabeled samples. Note that $L2$, MLE, and minEnt give very similar solutions—the main difference is that with a smoothness constraint on the weight magnitudes, $L2$ achieves smaller margin between classes 1 and 2 than MLE. The cross-validated minEnt solution is identical to $L2$'s, because in this case, the cross-validated error rate is 0 when the hyperparameter on the unlabeled sample's entropy term is set to 0.

V. Active Learning: Sample Selection Criteria for Oracle Labeling

Tailored for our semisupervised AL learning objective function, we propose a new sampling criterion suitable for both classification and new class discovery. Specifically, at the tth oracle labeling, we select $\underline{x}_{i^*}$ to be the most likely unknown class sample from $\mathcal{X}_u$

$$i^* = \underset{i \in \mathcal{I}_u^{(t)}}{\operatorname{argmax}} P\left[C \in \mathcal{C}_{\text{ru}}^{(t)} | \underline{x}_i\right] P\left[C = 0 | \mathcal{C}_{\text{ru}}^{(t)}, \underline{x}_i\right]. \quad (12)$$

Intuitively, the unlabeled sample most likely to belong to the unknown category should also deviate the most from the current common and LS categories, and once labeled, it is expected to either be from a new class or to significantly improve classification accuracy on the existing classes. At the initial AL phase, before any rare category is known, (12) reduces to $\arg\max_{i \in \mathcal{I}_u^{(t)}} P[C \in \mathcal{C}_{\text{ru}}^{(t)} | \underline{x}_i]$. Thus, our sample selection rule specializes, in this case, to choosing the sample least likely to belong to the common class subset. We call our new sample selection criterion (12) "most likely unknown".

Other than our proposed criterion, we also investigate the following AL sample selection criteria.

1) *Most Uncertain:* The unlabeled sample $\underline{x}_{i^*}$ with greatest class uncertainty (as measured by Shannon's entropy function $H(C|i^*)$ over all the known classes and the unknown class) is forwarded to the oracle. Intuitively, if an unlabeled sample is most uncertain for the current classifier, the removal of this uncertainty (by oracle labeling) should provide valuable information to the classifier.
2) *Least Common:* The unlabeled sample $\underline{x}_{i^*}$ with largest *a posteriori* probability of not belonging to the common class subset (as measured by $P[\mathcal{C}_{\text{ru}}^{(t)} | \underline{x}_{i^*}]$) is forwarded to the oracle. This choice prioritizes accuracy in discriminating between the common and LS/unknown classes.
3) *Random:* The unlabeled sample $\underline{x}_{i^*}$ is randomly chosen for oracle labeling. This choice gives a lower bound baseline for AL performance.

TABLE I
Data Set Properties. N: Number of Samples. D: Number of Attributes. K: Number of Classes. K_r: Number of Rare Classes. SCP: Smallest Class Proportion in Percentage

Data Set	N	D	K	K_r	SCP
Page Block	5473	10	5	4	0.50%
Statlog Shuttle	58000	9	7	4	14e-3%
Yeast	1484	8	10	5	0.34%
(White) Wine Quality	4898	11	7	4	0.10%
Abalone (age 3-23)	4177	8	21	13	0.15%
Spam Base	4601	57	2	1	5.0%
Seismic Pump	2584	18	2	1	5.0%
KDD'99 (10%)	494021	42	15	14	2.4e-3%

VI. Experimental Results

In this section, we will demonstrate: 1) the advantage of p-values as features over other feature mapping techniques (raw, radial basis function (RBF) kernel) when one class is rare. Thus, we validate our choice of p-values as derived features; 2) that p-values in conjunction with weight constraints yield sparse solutions; 3) the advantage of the proposed entropy-preserving regularizer (maxEnt) as opposed to minEnt regularization and other model variations in SL settings with unknown and LS classes; and 4) that, when applied to AL, our maxEnt SL together with our new AL sample selection criterion achieves a performance advantage over alternative methods, especially during the early phase of AL.

A. Performance Metrics

In all of our experiments, the available data set (including in aggregate both the training and test data files) was partitioned into two mutually exclusive subsets—one used for learning and the other for (heldout) testing of generalization accuracy. This was repeated ten times, with the test performance results averaged. For common/rare-category discrimination, there are two fundamental performance measures of interest on the test batch: 1) true detection rate—the proportion of truly rare class samples that have been correctly classified to $\mathcal{C}_r$ and 2) false alarm rate—the proportion of truly common class samples classified to $\mathcal{C}_r$. We evaluate these measures in all our experiments and, in an AL setting, for the current classifier (after every iteration of active labeling). As a comprehensive performance measure encompassing true detections and false positives, we measure average test set area under the receiver operating characteristic (ROC) curve (ROC area under the curve (AUC)). For discriminating among different common and rare classes, we also measure the per-class classification accuracy, averaged over all classes. In the AL setting, before an unknown class is discovered, all of its samples are assumed misclassified.

B. Data Set Description

Data sets were selected from the UCI ML repository, with both real and categorical features and with naturally rare classes. Most of the chosen data sets have been used in prior related studies [12], [28]. For each of these data sets, we define rare categories as those comprising less than 5% of the whole data set. The main properties of these data sets are shown

in Table I. The majority of these data sets have multiple normal and/or multiple rare classes. For the two-class domains (Spam Base and Seismic Pump), normal emails and normal recordings, respectively, are treated as the common category. Since Spam Base has balanced normal/spam email subsets, we randomly subsampled the spam subset to consist of only 5% of the whole batch.

C. Mixture p-Values as Features for Supervised Learning

Here, we compare several feature mapping techniques on UCI ML repository data sets, using the SVM as the supervised learning model. For the p-value feature mapping, we first used all the available common class samples in $\mathcal{X}_l$ to build GMM null models; then we measured first- and second-order mixture p-values as derived features for all the samples. We then applied different kernel mappings (linear, RBF) to train an SVM, with the hyperparameters (on margin slackness and on the RBF's width) chosen via tenfold CV on the learning batch. We compared against linear and RBF kernel SVMs that use the original (raw) features, as well as against SVMs that use both raw and all pairwise feature products. For discriminating common versus rare-category subsets, we measured ROC AUC performance as a function of the rare-category proportion (from 2% up to 20%). For data sets with multiple classes, we lumped the common classes together to form a single (majority) class and all rare classes together to form a single (minority) class. For each data set, we split it into a learning subset and a test subset of equal sizes. We then randomly subsampled without replacement from the learning subset to achieve specified majority/minority class proportions. This was also done on the test set. Again, the learning/testing split was performed ten times, with the test set classification performance averaged over these ten trials. As shown in Fig. 2, on every data set, p-value feature mappings (linear and RBF) gave the overall best results, compared with other feature mappings that were tested. Moreover, the performance advantage is most obvious when the class proportions are highly skewed (2% and 5%). In addition, as demonstrated for Abalone in Fig. 2(c), p-value with the linear kernel gives an almost perfect ROC when a few unknown class samples are present, but with the AUC decreasing as the number of unknown class samples is increased; on the other hand, p-values used with the RBF kernel mapping show a reverse trend. This result is not surprising as: the size of the unknown class grows, so does the Vapnik Chervonenkis dimension, and a more complicated (nonlinear) decision boundary may be needed to well separate the two classes.

D. Evaluating Semisupervised Learning Method Variations

Here, we compare our proposed scheme (P), with maxEnt regularization, p-values as features, our hierarchical posterior, with nonnegative constraints on weights in the top level, and employing our sample weighting choice for $\alpha_\Omega^{(t)}$ against alternatives that each vary some of our design choices: 1) using raw and pairwise product features and unconstrained weights, instead of p-value-based features in the top level (raw); 2) allowing unconstrained weights (positive and negative

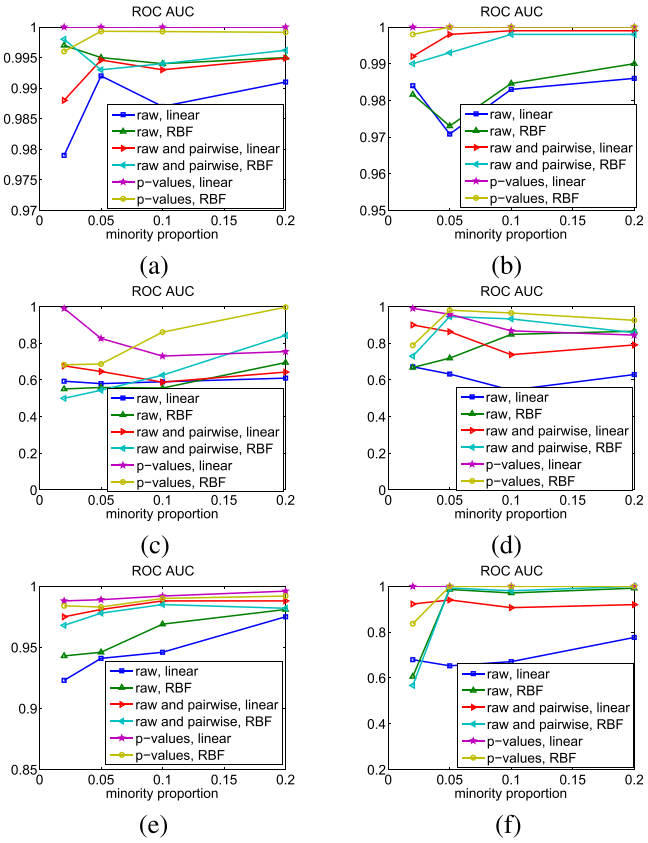


Fig. 2. Supervised classification experiment. Average ROC AUC performance comparison of different feature mappings input to linear and RBF kernel SVMs on various UCI data sets. (a) Page Block. (b) Spam Base. (c) Abalone. (d) Yeast. (e) KDD'99. (f) Seismic Pump.

valued) on p-value features (Uc); 3) applying equal weighting to all samples (Uw); and 4) maximizing the posterior log-likelihood without performing maxEnt regularization (MLE). Ten times we repeated a 50–50 split into learning and test subsets, with generalization performance averaged over these ten trials. In each trial, in the learning subset, we randomly labeled 20% of samples from each of the common classes, which are used to train the null GMM models. The p-values for each sample are calculated from the null models. Besides the 20% labeled common samples, we also labeled one sample from each of the rare categories. Thus, all categories are known. Then, we trained our proposed maxEnt classifier and these four method variations on several data sets. Hyperparameters [$\alpha_{\text{top}}^{(t)}$ and $\alpha_{\text{bot}}^{(t)}$] are chosen via tenfold CV. Although it seems that the best CV method is to minimize the hierarchical posterior's error, averaged over all classes, in the LS case, we do not have enough samples from each class to perform such validation. As an alternative, we minimized (over a finite grid of candidate values) a tenfold cross-validated average error rate measure of the top-level posterior. In the case of CV ties, a larger value of $\alpha_{\text{top}}^{(t)}$ and a larger value of $\alpha_{\text{bot}}^{(t)}$ were chosen.

As shown in Table II, imposing a nonnegative constraint on the top-level weights gives superior performance in all of these experiments, compared with a scheme with unconstrained weights. There is large improvement that comes from

TABLE II

COMPARE MODEL VARIATIONS. AVERAGE CLASSIFICATION ACCURACY PER-CLASS AND FRACTION OF INACTIVE MODEL PARAMETERS IN Λ. P: PROPOSED. Uc: UNCONSTRAINED PARAMETERS. Uw: UNIFORM WEIGHTING. MLE: MAXIMUM LIKELIHOOD

Data Set	Average Accuracy Per Class					Fraction of Inactive Model Parameters				
	P	Uc	Uw	Raw	MLE	P	Uc	Uw	Raw	MLE
Page	**0.41**	0.39	0.25	0.37	0.37	0.78	0	**0.80**	0	**0.80**
Statlog Shuttle	**0.78**	0.76	0.61	0.51	0.66	0.49	0	**0.54**	0	0.51
Yeast	0.47	0.47	0.27	0.48	**0.50**	0.78	0	**0.89**	0	**0.89**
Wine	**0.29**	0.22	0.19	0.23	0.23	**0.74**	0	0.42	0	**0.74**
Abalone	**0.21**	0.17	0.11	0.12	0.16	**0.45**	0	0.31	0	**0.45**
Spam	**0.88**	0.74	0.82	0.61	0.82	0.95	0	0.76	0	**0.97**
Seismic	**0.54**	0.50	0.51	0.49	0.53	0.84	0	**0.86**	0	**0.86**
KDD'99	**0.50**	0.45	0.48	0.46	0.48	0.90	0	0.90	0	**0.91**

zero-weighting the features that are deemed uninformative. Moreover, the use of weight constrained p-value features improves generalization performance compared with the use of raw and pairwise product features. But when the top-level weights are unconstrained, the use of p-value features performs similar to the use of raw and pairwise product features on the majority of the data sets. Hence, our inductive bias (both nonnegative weight constraints and p-value features) acts jointly to achieve superior performance. Third, maxEnt regularization is seen to be vital to achieving good performance, substantially outperforming classifiers trained without this regularization. That is, unlike MLE, the proposed regularizer avoids overtraining based on the few labeled rare samples. Finally, the effect of sample weighting is data set- dependent, but when nonuniform weighting is effective, it gives a large improvement over equal weighting. Thus, the experimental comparison in Table II supports our method's use of: 1) maxEnt regularization; 2) p-values; 3) nonnegative top-level weights; and 4) nonuniform sample weighting.

Table II also shows that the p-values as features and nonnegativity constraints on top-level weights act jointly to achieve a substantial level of parameter sparsity (zero weights) in the solution—P, Uw, and MLE use p-values and nonnegative top-level weights and achieve high top-level feature sparsity. Note also that this is achieved with only one labeled sample from each rare category. Thus, the degree of sparsity does not appear to strongly depend on the number of labeled rare (unknown) category samples. On Spambase, only $\sim$5% of top-level weight parameters are nonzero for our proposed method. The maximum degree of nonsparsity for our method is on Statlog Shuttle and Abalone, with more than 50% of weight parameters nonzero. By contrast, the method variations that allow unconstrained weights and which use raw features rather than p-values achieve no degree of sparsity (all nonzero weights). Finally, note that sparsity by itself is not an indicator of good performance—the MLE method (without maxEnt regularization) achieves the highest degree of sparsity on all data sets, but it is outperformed, accuracywise, by the proposed method on all the data sets.

E. Semisupervised Learning Evaluation

Here, we compare the performance of our maxEnt scheme against minEnt regularization [15] and $L2$ regularization, with all of these methods learning our p-value-based hierarchical discriminative model, with nonnegative weight constraints in the top level and unconstrained weights in the second level. The experimental protocol (training/test splits and performance averaging) was the same as in the Section VI. D. We are interested in two scenarios of SL: 1) with labeled samples from only the common classes (completely unknown rare categories) and 2) with LS rare categories. For the first (unknown class) scenario, from the learning subset, we randomly labeled 20% of samples from each of the common classes, treating the remaining samples as unlabeled ($\mathcal{X}_u$). Thus, all rare categories are unknown. For the latter scenario, besides the 20% labeled common class samples, we also labeled one sample from each of the rare categories. The labeled common class samples in the learning batch were used to learn the first- and second-order null GMM models; we then evaluated both first- and second-order p-values as derived features for all samples. We measured ROC AUC performance for the unknown class scenario, and measured both ROC AUC and average classification accuracy for the LS case. For the minEnt approach presented in [15], we used a pair of hyperparameters, one on the unlabeled sample's entropy and the other on the $L2$ norm. The hyperparameter values were chosen via tenfold CV applied to the learning batch, so as to minimize average classification error of the top-level posterior. In the case of CV error ties, a smaller weight on the unlabeled sample's entropy term and a larger weight decay term were selected. For the unknown class scenario (with no ability to measure CV error), all hyperparameters were fixed at $|\mathcal{X}_l|/|\mathcal{X}_u|$ for all the regularizers, so as to balance the effective sample size between the labeled and unlabeled data subsets.

TABLE III

SEMISUPERVISED EXPERIMENT FOR THE UNKNOWN CLASS SCENARIO: ROC AUC PERFORMANCE

Data Set	OC-SVM	GMM	Proposed	minEnt	L2
Page Block	0.500	0.969	**0.971**	0.5	0.5
Statlog Shuttle	**0.946**	0.927	0.943	0.5	0.5
Yeast	0.802	0.706	**0.824**	0.5	0.5
Wine	0.747	**0.751**	0.738	0.5	0.5
Abalone	0.741	0.725	**0.780**	0.5	0.5
Spam Base	0.500	0.702	**0.874**	0.5	0.5
Seismic Pump	0.554	0.587	**0.704**	0.5	0.5
KDD'99	0.500	0.862	**0.985**	0.5	0.5

The ROC AUC performance for the unknown class scenario is shown in Table III. Note that when there are no labeled examples from the rare categories, both minEnt and $L2$ yield

TABLE IV
SEMISUPERVISED EXPERIMENT FOR THE LS CLASS SCENARIO: ROC AUC AND AVERAGE ACCURACY PER-CLASS. P: PROPOSED. M: minEnt. $L2$: $L2$ REGULARIZATION

Data Set	ROC AUC			Average Accuracy		
	P	M	L2	P	M	L2
Page Block	**0.97**	0.93	0.97	**0.41**	0.39	0.35
Statlog Shuttle	**0.93**	**0.93**	0.92	**0.78**	0.73	0.68
Yeast	**0.78**	0.74	0.71	**0.47**	0.42	0.39
Wine	**0.72**	0.70	0.71	**0.29**	0.23	0.17
Abalone	**0.66**	0.62	0.62	**0.21**	0.17	0.19
Spam Base	**0.91**	0.87	0.88	**0.88**	0.73	0.76
Seismic Pump	0.61	0.50	**0.62**	**0.54**	0.50	0.53
KDD'99	**0.98**	0.86	0.86	**0.50**	0.49	0.41

degenerate solutions, classifying all samples as coming from the common classes. Thus, these methods wholly fail under the first scenario. The performance of two standard ADs is also shown: 1) using OC-SVM [5], [6] with the RBF kernel and 2) a GMM, trained on the labeled common class samples, modeling the full-feature space. For these two methods, we ranked the test samples based on their respective anomaly scores (distance from origin in the case of OC-SVM and data likelihood for the GMM) and used the ordered list to sweep out an ROC curve. We also lower bound the ROC AUC by 0.5. Both of these ADs lag performance-wise behind our proposed model on most of the investigated data sets. Compared with these full-feature space models, our proposed method's inductive bias (p-values and nonnegative constraints) yields a better classifier on these data sets. Table IV shows both the ROC AUC and average classification accuracy for the second (LS) scenario. Again, we see superior performance of our maxEnt regularizer compared with minEnt and $L2$-norm regularizers, presumably because the maxEnt regularization avoids overtraining when there are few labeled rare samples, while minEnt encourages weight magnitudes to increase. However, during AL, as the number of labeled rare samples increases, minEnt tends to perform well (see the Section VI. F for further discussion). An uninformative prior ($L2$-norm) also underperforms compared with our proposed data-dependent regularizer.

F. Active Learning

In this section, we investigate the use of our semisupervised maxEnt learning approach within a pool-based AL setting, with samples chosen for oracle labeling one-by-one. In principle, the null model GMMs could be adapted when a new oracle labeling is from a common class. However, we fixed the null models after their initial training because there were an adequate number of common class samples in the initial $\mathcal{X}_l$. We also evaluate several different schemes for selecting which sample to label within AL. To assess performance, we measured test set ROC AUC and average classification accuracy as a function of the number of oracle labelings. We assumed the unknown class scenario at the outset, with no initially known (labeled) rare class samples in the learning batch, and with 20% from each of the common classes samples labeled. The experimental protocol was the same as in Sections VI. D and VI. E, with results averaged over ten learning/test splits.

1) Comparison of Sample Selection Schemes: AL performance may be affected by the choice of classifier structure, learning method, as well as by the scheme for choosing which samples to label. First, we compared different active sample selection strategies, using our maxEnt model as the classifier. Initially, since there are no labeled rare class samples, it may appear tempting to select for labeling the sample least likely (as measured by the model) to belong to the common class subset. Indeed, when using this strategy as the AL sample selection mechanism, we are able to identify many more rare class samples than either most likely unknown or most uncertain sampling. However, as next shown, the generalization performance of a classifier using this least common strategy can be poor. Besides the strategies mentioned in Section V, we also show the performance achieved when all the samples in the learning batch are labeled (this should upper bound the performance of AL).

In Fig. 3, we show both ROC AUC and average classification accuracy as a function of oracle labelings, while in Table V, we show the average number of true detections and the fraction of rare classes discovered after 100 oracle labelings. We emphasize that the average number of true detections is not a classification performance measure—it merely indicates how frequently the sample selection strategy forwards unknown/rare class samples to the oracle for labeling. As shown in Table V, the least common strategy forwards many more unknown/rare class samples to the oracle on average than most likely unknown and most uncertain, with uncertainty sampling forwarding the fewest. This can be explained by the fact that the number of common class samples close to the decision boundary dominates the rare class ones. However, as shown in Fig. 3, both least common and most uncertain have poorer generalization performance than most likely unknown. Note the initial drop in the performance of most uncertain sampling on the Wine data set. Here, the labelings of most uncertain samples are clearly not the most helpful for classification. Moreover, while the most likely unknown and least common curves in Fig. 3 start out together because the two criteria are equivalent until the first unknown category sample is labeled, the two curves diverge after this first labeling. The least common strategy has a tendency to identify many rare class samples that are very similar to each other and far from the decision boundary—it does not tend to select informative samples that either are close to the decision boundary or are instances of novel classes. Thus, as shown in Table V, it tends to have higher true detections but a lower fraction of rare classes discovered compared with most likely unknown (see the results on Yeast and KDD'99). It also has poorer generalization performance than most likely unknown shown in Fig. 3. Note in particular results on KDD'99, Wine, Page Block, and Yeast.

In some applications, where the rare class samples are suspicious and, hence, urgently actionable, there may thus be a need for two (concurrent) operational labeling streams [3]. One is the most actionable samples: these are the unlabeled samples most likely to be rare (suspicious), whose identification may lead to subsequent action by a human operator. The other is the most informative samples: these are the samples whose

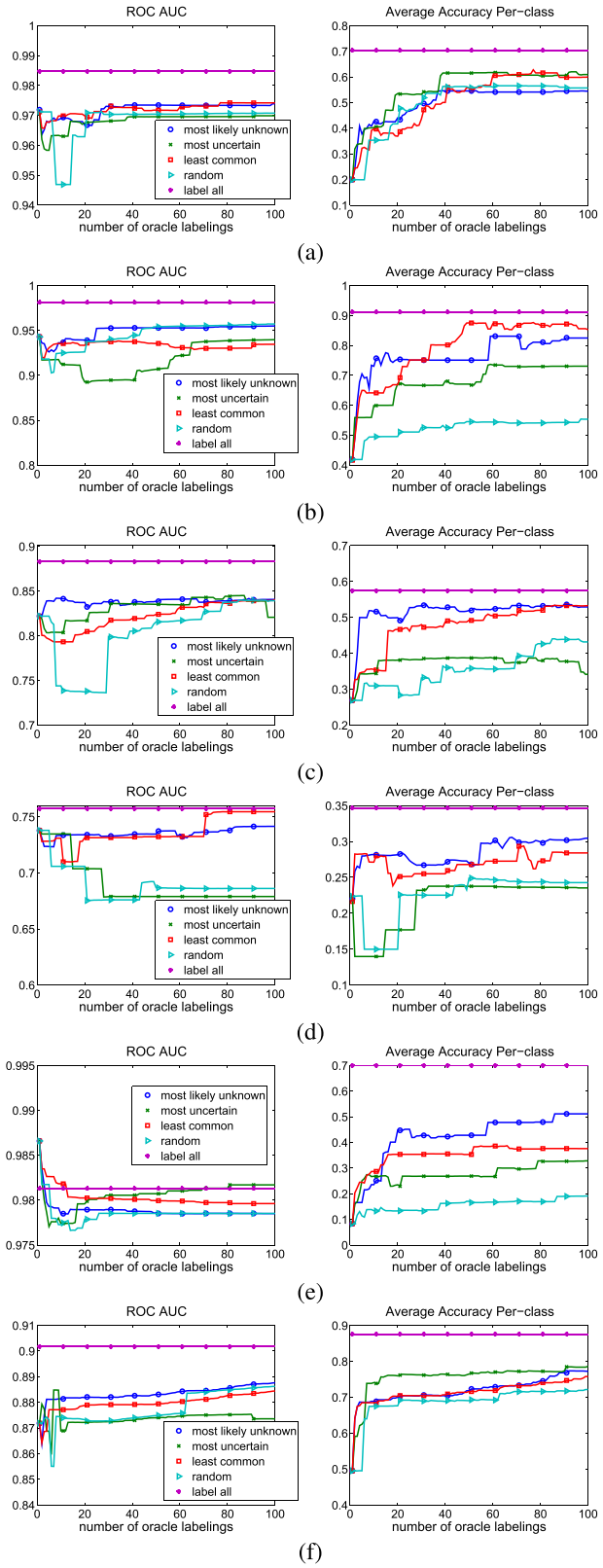


Fig. 3. AL experiment for different sample selection strategies. (a) Page Block. (b) Statlog Shuttle. (c) Yeast. (d) Wine Quality. (e) KDD'99. (f) Spam Base.

labeling will help to improve the generalization accuracy of the classifier the most. The new most likely unknown method achieves the best results with the highest fraction of rare classes discovered, and sometimes comes close to the performance upper bound line.

Note that, for all these strategies, there is an initial drop in ROC AUC, while the average accuracy increases on all the data sets. The reason is that when the first few rare class samples are labeled, the nonuniform weighting takes effect, and there is a tradeoff between false positives and classification accuracy between the common and rare class subsets. As more rare samples are labeled and new classes are discovered in latter AL stages, ROC AUC tends to increase and surpass the initial level.

2) Comparison of Active Learning Models: Next, we compared our proposed maxEnt model (using the most unlikely unknown strategy for sample selection) with other alternatives: SVM-based, minEnt-based, and a generative model-based, elaborated as follows. For an SVM-based approach, we first used OC-SVM with the RBF kernel and queried the most anomalous samples, until the first rare class sample was selected and labeled; then, we switched to using the standard multiclass SVM with RBF kernel, labeling the most uncertain sample. A related work can be found in [29]. However, this approach requires the knowledge of the number of classes and, to initiate the learning, requires labeled instances from every class. As for a minEnt-based approach (as opposed to our maxEnt model), we chose the approach presented in [15]. We also compared with the recent model presented in [28], which we dub "unified", a generative model that uses a Dirichlet process to model the true class distribution, with expected classifier improvement as the sample selection mechanism. This approach was shown to outperform a number of previous works on rare-category detection and characterization, i.e., [12], [30], and [31].

In our initial AL experiments, we found that the minEnt approach achieves no generalization capability, i.e., the method classifies all samples as common class (see the Section VI. E and Table III) and never selects any samples from the unknown class. This represents a fundamental breakdown of the minEnt approach within an initially unknown class AL setting. Subsequently, in order to at least allow minEnt to achieve some generalization performance, we fed the minEnt classifier with the labeled samples selected by our maxEnt approach. In this way, informative samples determined by our proposed model were also used to train the minEnt classifier. As shown in Fig. 4, it is apparent that minEnt, even when fed the labeled samples selected by the maxEnt approach, achieves no generalization power in the early AL stage, on all the data sets. However, as observed for KDD'99, as more and more rare class samples are included into the labeled subset, minEnt improves and performs quite well in the latter iterations, sometimes even better than maxEnt at a certain point. Compared with [28], our maxEnt approach achieves overall superior performance on all the data sets except KDD'99, especially during the early AL phase. Note that our model's ROC AUC curve outperforms all others on all data sets except Statlog Shuttle and Wine.

For comparing our method with the SVM, note that for the leftmost part of the curve, before any rare class samples have been labeled, we are really comparing our approach with OC-SVM. The SVM approach, which does not make

TABLE V
AL TRUE DETECTIONS AND FRACTION OF RARE CLASSES DISCOVERED OUT OF 100 ITERATIONS FOR DIFFERENT SAMPLE SELECTION STRATEGIES. MLU: MOST LIKELY UNKNOWN. MU: MOST UNCERTAIN. LC: LEAST COMMON. R: RANDOM

Data Set	AL True Detections				Fraction of Rare Classes Discovered			
	MLU	MU	LC	R	MLU	MU	LC	R
Page Block	35	12	**85**	15	**1.0**	**1.0**	**1.0**	0.75
Statlog Shuttle	60	4	**68**	2	**0.93**	0.75	**0.93**	0.12
Yeast	24	6	**49**	12	**0.93**	0.62	0.74	0.64
Wine	14	5	**34**	9	**1.0**	0.83	0.99	0.70
KDD'99	**33**	18	31	4	**0.71**	0.57	0.32	0.16
Spam Base	48	11	**49**	6	**1.0**	**1.0**	**1.0**	**1.0**

effective use of unlabeled samples to regulate the learning, performs poorly on all the data sets.

VII. RELATED WORK

Training set class imbalance for supervised learning of classifiers has been addressed in a number of prior works. Widely used approaches include class rebalancing (by discarding samples from the common classes) [32] and class reweighting [33], [34], where greater weights in the samplewise learning objective function are given to samples from rare classes. Similar to these works, our approach assigns a larger weight to rare class samples. Other approaches, such as boosted ensembles [35], implicitly address this problem, by focusing the training of successive classifier stages on the most difficult examples. That is, the rare class examples and the examples from common classes nearest to them in feature space are likely the most difficult examples to correctly classify.

AL with rare/unknown categories was first addressed by Pelleg and Moore in [2], motivated by trying to find interesting objects in satellite images. Assuming that the data are generated by a Gaussian mixture, starting from an entirely unlabeled batch, they first clustered samples using EM. Subsequently, after some samples are labeled, they learn a semisupervised GMM [36]. Then, an interleave algorithm is introduced to first rank order the unlabeled samples in each of the mixture components by increasing likelihood, then round-robin querying samples from the top of the ordered lists. After oracle-supplied labels are determined, the Gaussian mixture is retrained in light of the newly labeled samples and the process is repeated. However, the number of rare categories has to be known and set in advance for this approach to work. A related approach was later introduced by Vatturi and Wong [37], wherein they used hierarchical mean shift clustering with increasing bandwidth settings to capture anomalies at different scales. Dasgupta and Hsu [7] proposed a sampling scheme, which exploits cluster structure (and class purity of the same) in the data to potentially improve AL efficiency. They demonstrated their method on the Statlog Shuttle data domain, for which the rarest category occurs with frequency just 0.014%. Their scheme reduced the number of labelings required to identify a labeled example from every class by a factor of almost ten, compared with random selection.

The most focused, notable research effort, of which we are aware, on AL with rare categories is the work by He [12]. This thesis separately addresses rare-category detection (identifying the first labeled instances of all classes), rare-category characterization (to subsequently identify all instances of a rare category), and a related unsupervised problem, where rare groups should be identified in the absence of class labels for any samples in the pool. There are some fundamental differences between the approaches taken in [12] and in this paper. First, He [12] focuses on identifying examples of the rare category—not on overall classification accuracy. For a two-class problem with one unknown category, the most valuable samples to label may be the samples that are nearest to the true (optimal) decision boundary. It need not be the case that such samples predominantly come from the unknown class. In fact, if the unknown class is a rare class, one would expect, within the unlabeled pool, that more common category samples are close to the decision boundary than rare-category samples. Accordingly, we expect that the most effective AL sample selection strategy (to optimize overall classification performance) should sample accordingly. In fact, this was borne out by our experimental results—prioritizing the labeling of most/all rare-category samples may not help a classifier to learn how to accurately discriminate common category samples that are near these rare-category samples. Many of these rare-category samples may be largely redundant. Second, some of the proposed methods in [12] assume the class priors are known. While this may be realistic for some domains (e.g., rare disease categories with known prevalence), such information will not be available in general. Third, in detecting rare-category examples, He [12] uses a nearest neighbor framework, with distances, in general, measured using all the features. By contrast, our approach seeks to be robust to the presence of many noisy features that are irrelevant to (and which may thus confound) discrimination of the rare and common categories. Accordingly, it accommodates implicit (soft) feature selection.

The most recent work that unifies class discovery and learning is the method presented in [28]. Hospedales *et al.* [28] proposed a Dirichlet Process to model the true class distribution, considering both known and unknown categories. They also developed an expected classifier improvement sample selection strategy that unifies both class discovery and learning into one objective. This method was shown to outperform several existing methods, i.e., [12], [30], and [31]. However, the sample selection strategy requires trial-learning before every query, with extremely high complexity as the number of categories grows. In addition, unlike [28], the dimension sparsity is discriminately trained in our method, allowing it to focus on the relevant dimensions and ignore irrelevant ones.

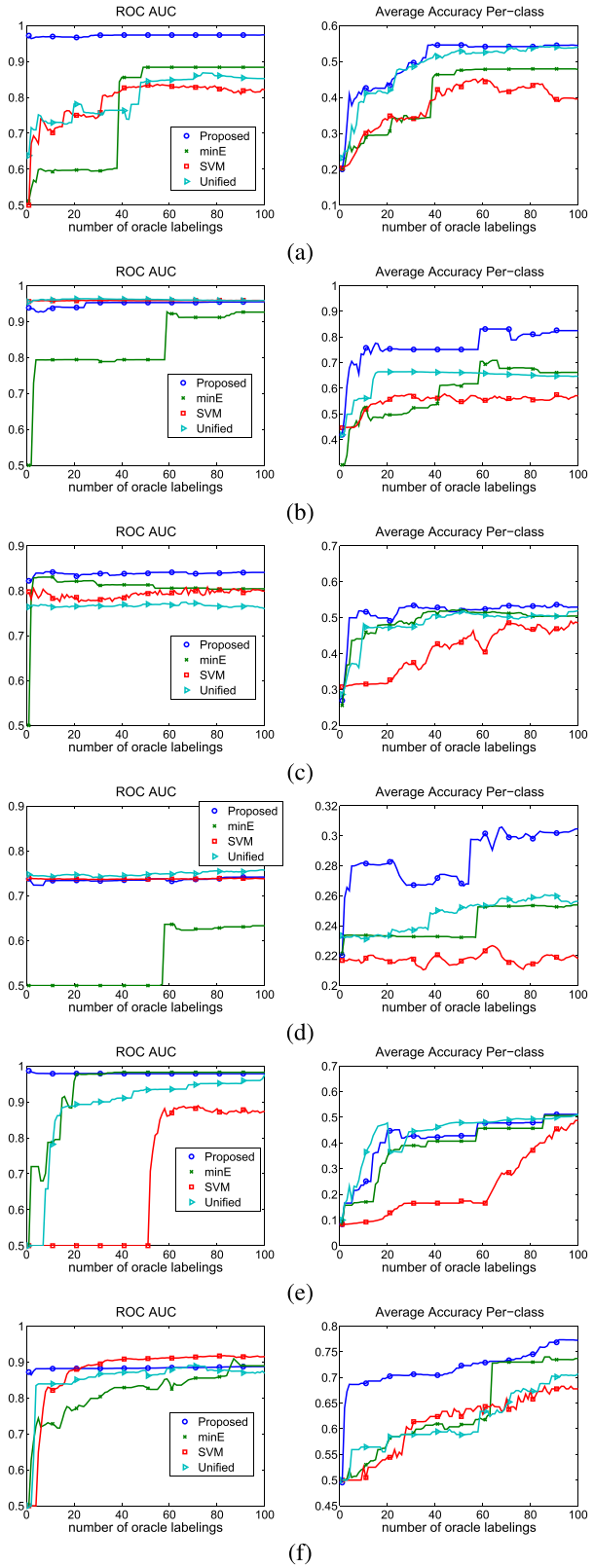


Fig. 4. AL experiment. Comparison of methods' ROC AUC performance and average classification accuracy versus number of oracle labelings, on various data sets. (a) Page Block. (b) Statlog Shuttle. (c) Yeast. (d) Wine Quality. (e) KDD'99. (f) Spam Base.

SL, wherein unlabeled samples are used to achieve a type of regularization, has been proposed in a number of prior works [15], [24]–[26], [38]. All of these approaches introduce a cost term on the unlabeled samples, which encourages solutions with minEnt on the unlabeled samples, i.e., consistent with most confident decisions [15] and class label smoothness [26]. Grandvalet and Bengio [15] is also related to transductive SVM as a type of maximal margin separator [39]. Gomes *et al.* [25] goes one step further than [15] by introducing an empirical estimate of the class prior. The proposed objective function in [25] is shown to work well in both unsupervised clustering and SL tasks. Mann and McCallum [38] extends [15] by adding an expected class estimate regularization term, whose prior belief is provided by a domain expert. However, a concern with these approaches is that they encourage overtraining in the case where there are too few or no labeled rare class examples. Accordingly, we modified these approaches, essentially by changing the sign on the unlabeled term—rather than minimizing decision uncertainty, our approach encourages maximizing decision uncertainty so as to limit the amount of classifier learning performed based on a limited number of supervising rare class examples. Moreover, for a two-class (common versus unknown/LS class subset) setting, our optimization problem is convex, whereas the minEnt problem is not.

Finally, there is our earlier work [3], which specifically addressed AL to discriminate suspicious from innocuous anomalies, applied in the domain of vehicle tracking. Here, we have generalized this approach to address multiple common and multiple unknown classes, we introduced a novel AL sample selection criterion, and we have shown that the model and basic learning strategy are more generally applicable to SL, to common versus unknown/LS class discrimination, and to a variety of application domains.

VIII. Conclusion

In this paper, we proposed novel SL and AL frameworks, along with a class posterior model, for learning to discriminate unknown from common classes. Our use of a hierarchical class posterior accommodates multiple common, multiple LS, and unknown classes. Our top-level class posterior imposes a constraint on the parameter space (nonnegative weights on p-values) consistent with the inductive bias that what is unknown is a subset of what is anomalous relative to the common classes. Experimentally, we found that imposing this inductive bias leads to sparsity in the solution, with only a small subset of feature p-values contributing to the class decision. Our SL approach exploits unlabeled samples in an unconventional way—to control the amount of learning given few labeled samples. That is, we proposed a novel (maxEnt) sample-dependent semisupervised regularizer, suitable for AL with unknown categories and for SL with high-class imbalance in general. We also proposed a sample selection criterion for AL which focuses on the discovery of novel classes. Our framework is particularly suitable for learning a mapping between a human operator's concept of an unknown class (which could be based on complementary, human-digestible information sources, e.g., images or text) and an information-rich, albeit possibly very low-level high-dimensional feature space. The classifier achieves this mapping, identifying which

(low level) features prominently capture the operator's concept. Future work could investigate encoding higher order interactions into our posterior model. It could also explore a hybrid approach, with maxEnt regularization used during the early AL stage, transitioning to minEnt regularization in the later learning phase. We could consider online learning of the null. We may also explore a hyperparameter to control the mass of the unknown category. Finally, we could investigate the effect of nonconvexity of the learning objective on the quality of classifier solutions we obtain.

ACKNOWLEDGMENT

The authors would like to thank O. Mendoza-Schrock and B. Stieber for their helpful discussions.

REFERENCES

[1] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin–Madison, Madison, WI, USA, Tech. Rep. 1648, 2010.
[2] D. Pelleg and A. Moore, "Active learning for anomaly and rare-category detection," in *Proc. Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA, Dec. 2004, pp. 1073–1080.
[3] Z. Qiu, D. J. Miller, B. Stieber, and T. Fair, "Actively learning to distinguish suspicious from innocuous anomalies in a batch of vehicle tracks," *Proc. SPIE*, vol. 9079, pp. 90790G-1–90790G-11, Jun. 2014.
[4] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symp. Secur. Privacy*, Oakland, CA, USA, May 2010, pp. 305–316.
[5] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.
[6] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," *J. Mach. Learn. Res.*, vol. 2, pp. 139–154, Dec. 2001.
[7] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, Jul. 2008, pp. 208–215.
[8] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
[9] K. Das, J. Schneider, and D. B. Neill, "Anomaly pattern detection in categorical datasets," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Las Vegas, NV, USA, Aug. 2008, pp. 169–176.
[10] Z. Qiu, D. J. Miller, and G. Kesidis, "Detecting clusters of anomalies on low-dimensional feature subsets with application to network traffic flow data," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process.*, Boston, MA, USA, Sep. 2015, pp. 1–6.
[11] G. V. Trunk, "A problem of dimensionality: A simple example," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 3, pp. 306–307, Jul. 1979.
[12] J. He, "Rare category analysis," Ph.D. dissertation, School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2010.
[13] L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion detection with unlabeled data using clustering," in *Proc. 8th ACM CSS Workshop Data Mining Appl. Secur.*, Nov. 2001, pp. 5–8.
[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2001.
[15] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2004, pp. 529–536.
[16] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[17] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *Comput. J.*, vol. 41, no. 8, pp. 578–588, Jan. 1998.
[18] Z. Ghahramani and M. J. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, Dec. 1999, pp. 449–455.
[19] S. C. Markley and D. J. Miller, "Joint parsimonious modeling and model order selection for multivariate Gaussian mixtures," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 3, pp. 548–559, Jun. 2010.
[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.
[21] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 1999.
[23] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951.
[24] A. Fujino, N. Ueda, and K. Saito, "Semisupervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 424–437, Mar. 2008.
[25] R. Gomes, A. Krause, and P. Perona, "Discriminative clustering by regularized information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2010, pp. 766–774.
[26] A. Corduneanu and T. Jaakkola, "Data-dependent regularization," in *Semi-Supervised Learning*, O. Chappelle, B. Schölkopf, and A. Zien, Eds. Cambridge, MA, USA: MIT Press, 2006.
[27] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York, NY, USA: Springer, 2006.
[28] T. M. Hospedales, S. Gong, and T. Xiang, "A unifying theory of active discovery and learning," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 453–466.
[29] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: Active learning in imbalanced data classification," in *Proc. 16th ACM Conf. Inf. Knowl. Manage.*, Lisbon, Portugal, Nov. 2007, pp. 127–136.
[30] T. M. Hospedales, S. Gong, and T. Xiang, "Finding rare classes: Active learning with generative and discriminative models," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 2, pp. 374–386, Feb. 2013.
[31] T. S. F. Haines and T. Xiang, "Active learning using Dirichlet processes for rare class discovery and classification," in *Proc. 22nd Brit. Mach. Vis. Conf.*, Dundee, Scotland, Aug. 2011, pp. 1–11.
[32] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 4th Int. Conf. Mach. Learn.*, vol. 97. Nashville, TN, USA, Jul. 1997, pp. 179–186.
[33] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *Proc. 7th Eur. Conf. Principles Pract. Knowl. Discovery Databases*, Cavtat-Dubrovnik, Croatia, Sep. 2003, pp. 107–119.
[34] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 30–39, 2004.
[35] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
[36] D. J. Miller and H. S. Uyar, "A mixture of experts classifier with learning based on both labelled and unlabelled data," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, Dec. 1996, pp. 571–577.
[37] P. Vatturi and W.-K. Wong, "Category detection using hierarchical mean shift," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Paris, France, Jun. 2009, pp. 847–856.
[38] G. S. Mann and A. McCallum, "Simple, robust, scalable semi-supervised learning via expectation regularization," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, Jun. 2007, pp. 593–600.
[39] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

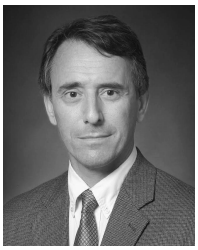
Zhicong Qiu received the B.S. and M.S. degrees in electrical engineering from the New York University Tandon School of Engineering, Brooklyn, NY, USA, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering with The Pennsylvania State University, University Park, PA, USA.

He was a Teaching and Laboratory Assistant with the NYU Wireless Laboratory, Brooklyn, from 2011 to 2012. He was also a Research Contractor with the Air Force Research Laboratory, Dayton, OH, USA, in 2015. He has been a Research Assistant with the Electrical Engineering Department, The Pennsylvania State University, since 2012. His current research interests include active learning, anomaly detection, machine learning, statistical pattern recognition, computer vision, and communication network.

David J. Miller (S'86–M'87–SM'07) received the B.S.E. degree from Princeton University, Princeton, NJ, USA, in 1987, the M.S.E. degree from the University of Pennsylvania, Philadelphia, PA, USA, in 1990, and the Ph.D. degree from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 1995, all in electrical engineering.

He was with General Atronics Corporation, Wyndmoor, PA, USA, from 1988 to 1990. He has been with the Electrical Engineering Department, The Pennsylvania State University, University Park, PA, USA, since 1995, where he has also been a Professor of Electrical Engineering since 2007. His current research interests include statistical pattern recognition, machine learning, source coding, bioinformatics, and network security.

Dr. Miller was a member of the Machine Learning for Signal Processing Technical Committee within the IEEE Signal Processing Society from 1997 to 2010, and was its Chair from 2007 to 2009. He received the National Science Foundation Career Award in 1996. He was the General Chair of the 2001 IEEE Workshop on Neural Networks for Signal Processing. From 2004 to 2007, he was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING.



George Kesidis received the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California at Berkeley, Berkeley, CA, USA, in 1990 and 1992, respectively.

He was a Professor of Electrical and Computer Engineering with the University of Waterloo, Waterloo, ON, Canada, for eight years. He has been a Professor of Electrical Engineering and Computer Science with The Pennsylvania State University, University Park, PA, USA, since 2000. His current research interests include many aspects of networking, cyber security, and machine learning, in particular, intrusion detection based on large-scale network datasets, and, more recently, energy efficiency and the impact of economic policy.