A Unifying Theory of Active Discovery and Learning

Timothy M. Hospedales, Shaogang Gong, and Tao Xiang

EECS, Queen Mary, University of London
{tmh,sgg,txiang}@eecs.qmul.ac.uk

Abstract. For learning problems where human supervision is expensive, active query selection methods are often exploited to maximise the return of each supervision. Two problems where this has been successfully applied are active discovery – where the aim is to discover at least one instance of each rare class with few supervisions; and active learning – where the aim is to maximise a classifier's performance with least supervision. Recently, there has been interest in optimising these tasks jointly, i.e., active learning with undiscovered classes, to support efficient interactive modelling of new domains. Mixtures of active discovery and learning and other schemes have been exploited, but perform poorly due to heuristic objectives. In this study, we show with systematic theoretical analysis how the previously disparate tasks of active discovery and learning can be cleanly unified into a single problem, and hence are able for the first time to develop a unified query algorithm to directly optimise this problem. The result is a model which consistently outperforms previous attempts at active learning in the presence of undiscovered classes, with no need to tune parameters for different datasets.

1 Introduction

Many real life learning problems start with relatively little prior knowledge about a domain, and require both the space of classes in the domain to be discovered as well as building classifiers to discriminate among said classes. Moreover, it is often the case that the distribution of class frequencies is highly uneven, and prior knowledge is limited to the majority background class, while the classes of most interest for discovery and discrimination are relatively rare. This is the task of learning to classify in the presence of undiscovered rare classes [1,2].

This problem is common in a wide range of scientific data analysis problems. For example, in astronomy [3], most sky survey content is well understood and only 0.001% may represent new phenomena of interest for study. In internet traffic analysis, the majority of data is due to regular activity, whereas types of network intrusions represent rare classes of interest [4]. In visual surveillance, most observed activities are ordinary behaviours, but rarely there may be a dangerous or malicious activity of interest to security services [5]. Finally, in learning from large scale vision data [6], there is typically a long tailed distribution of classes (Fig. 1). In all these cases, labelling sufficient data to cover all classes and model them well would be prohibitively costly since labelling each instance

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part V, LNCS 7576, pp. 453–466, 2012.

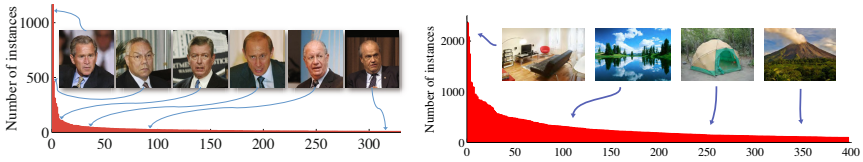


Fig. 1. Large scale visual learning problems such as face [6] and scene [7] recognition typically contain highly imbalanced class distributions

requires significant time from a human expert. For this reason, there has been recent interest in exploring a joint active discovery and active learning paradigm.

Active Learning. methods aim to iteratively select instances for human annotation such that classifier performance (on a known space of classes) is maximised with few supervisions. This is a large field, so we point the reader to the survey in [8] and only highlight a few relevant studies here. The most common approaches query instances exhibiting maximum uncertainty [9,10]. More theoretically appealing approaches minimise the generalisation error in expectation [11,12,13,14], or by an upper bound [15]. Active learning methods, however, are unsuitable for new domains as they typically require a known space of classes on which to operate. One potential solution is to first apply active discovery.

Active Discovery. methods aim to find at least one instance of each class in a dataset with few supervisions. Solutions to this problem have been varied, but involve very different query criteria to active learning. For example: outlier detection, low-likelihood in Gaussian mixtures [3], gradient [16], mean-shift hierarchical clustering [17] and nearest neighbour [18] criteria. This problem is typically treated in isolation from subsequent classification, and is seriously suboptimal if the ultimate aim is to detect or classify the discovered classes.

Active Discovery and Learning. To support interactive modelling of new domains, recent studies have tried to combine active discovery and learning to efficiently solve active classifier learning with undiscovered classes. Simple sequential or iterative application of separate discovery and learning criteria in fixed proportion was considered in [4]. However, this uses supervisions inefficiently, e.g., continually "wasting" discovery queries once all classes have been discovered. To address this issue, [1] proposes a more flexible approach which adapts between discovery and learning criteria based on their past success. This outperforms the non-adaptive approach [4] significantly, but is still heuristic and relies on careful selection of various free parameters. In contrast to [4] and [1], which still use independent discovery and query criteria, more recent studies [2] have made attempts at devising a single simple criterion which queries points that are likely to either reveal new classes, or to improve classification. This approach outperforms [1], but limited theoretical justification is provided.

A Unified View. In this paper, we make four key contributions. (i) We show for the first time how active learning with undiscovered classes can be unified

into a single problem with clear theoretical motivation. Our approach is based on the expected utility of a classifier, exploiting a new Dirichlet process mixture approximation. This allows the potential gain of classifier improvement vs class discovery to be quantified in the same units, thus enabling direct optimisation of classifier performance. (ii) We show how various existing models including [2] are approximations to our full unified model. Similarly to other expected generalisation error approaches [12,13], a naive implementation of our framework is computationally expensive. To overcome this (iii) we show how to perform efficient incremental computations and approximations, without loosing the generality of our model or introducing data-dependent parameters. Finally, (iv) we conduct the largest empirical evaluation of active learning and discovery to date, including for the first time various large scale vision datasets, and show that our expected accuracy method significantly outperforms previous approaches.

2 Formulation

We will start with reviewing a general formalisation of the classifier learning task; then show how this relates various existing active learning criteria and finally derive query criteria suitable for use in the presence of undiscovered classes.

In classification, the goal is to learn a model $p_\theta(y|x)$, parameterized by θ, which generalises across a space of input $x \in \mathcal{X}$ and classes $y \in \mathcal{Y}$. The performance of the classifier for each input x is measured by its utility $\mathcal{U}$, which is some function of the true class y and the estimated class distribution at each point, $p_\theta(y|x)$. The expected generalisation utility of classifier θ over the domain is then:

$$\mathcal{U}(\theta) = \int_x \sum_y \mathcal{U}\left(y, p_\theta(y|x)\right) p(y|x)p(x)dx. \tag{1}$$

Note that equivalent converse formulations involving expected risk or loss are also common [12,13,8]. In practice, one typically approximates Eq. (1) by summing over a labelled training set $D_l = \{(x_i, y_i)\}_{i=1}^{N_l}$:

$$\tilde{\mathcal{U}}_{D_l}(\theta) = \frac{1}{N_l} \sum_{i \in L} \mathcal{U}(y_i, p_\theta(y|x_i)). \tag{2}$$

For both classifier training and active learning, different definitions of utility induce slightly different objectives. For example, one common definition of utility [12] is $0/1$ accuracy:

$$\mathcal{U}(y, p_\theta(y|x)) = \begin{cases} 1 & \text{if } y = \text{argmax}_{y'} p_\theta(y'|x) \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where optimising Eq. (3) corresponds to maximising the number of correct classifications in the training set. Alternatively, a common soft measure [13,14] is the probability of making the correct prediction $\mathcal{U}(y_i, p_\theta(y|x_i)) = p_\theta(y_i|x_i)$ where:

$$\tilde{\mathcal{U}}_{D_l}(\theta) = \frac{1}{N_l} \sum_{i \in L} p_\theta(y_i|x_i). \tag{4}$$

With this general formulation, we can see classifier learning as choosing the best parameters $\hat{\theta} = \text{argmax}_\theta \tilde{\mathcal{U}}_{D_l}(\theta)$, possibly with an added regularisation term on θ to avoid over-fitting, and active learning as choosing the best dataset $\hat{D}_l = \text{argmax}_{D_l} \text{max}_\theta \tilde{\mathcal{U}}_{D_l}(\theta)$.

Finally, if we are not confident about the true labels (e.g., because of missing data, or active learning [12,13,14]), then as in Eq. (1), we can include the expectation over the true label distribution $p(y_i|x_i)$, so Eq. (4) becomes:

$$\tilde{\mathcal{U}}_{D_u}(\theta) = \frac{1}{N_u} \sum_{i \in U} \sum_{y_i} p(y_i|x_i) p_\theta(y_i|x_i), \tag{5}$$

where U now indexes an unlabelled dataset. Next, we use the above formalisation to address issues in active learning and discovery.

2.1 Active Learning

In active learning, we have data $D = \{D_l, D_u\}$ with both labeled $D_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and unlabelled $D_u = \{x_i\}_{i=1}^{N_u}$ subsets indexed by L and U respectively. In first order active learning, we iteratively: (i) select the "best" element i^* of the unlabelled set U for labelling; (ii) remove x_{i^*} from D_u and add (x_{i^*}, y_{i^*}) to D_l and (iii) update classifier parameters θ based on D_l. A principled objective for active querying is to pick at each iteration the element which maximises expected utility, $i^* = \text{argmax}_i \tilde{\mathcal{U}}_{D^i}(\theta)$, where D^i denotes the dataset with i labeled $D^i = \{D_l \cup (x_i, y_i), D_u \backslash (x_i)\}$. To predict the utility of observing each i in advance, we take expectation over its true label y_i:

$$\tilde{\mathcal{U}}_{D^i}(\theta) = \sum_{y_i} p(y_i|x_i) \tilde{\mathcal{U}}_{D_l \cup (x_i, y_i), D_u \backslash x_i}(\theta), \tag{6}$$

where $\mathcal{U}_{D_l, D_u}(\theta)$ indicates the expected utility based on both the available labeled D_l and unlabelled data D_u. If we use the probabilistic utility functions of Eqs. (4) and (5) for labeled and unlabelled data respectively, then we obtain the complete expression for the expected utility of querying point i:

$$\tilde{\mathcal{U}}_{D^i}(\theta) = \sum_{y_i} p(y_i|x_i) \frac{1}{\overline{N}} \left(\sum_{j \in L \cup i} p_{\theta_{+i}}(y_j|x_j) + \sum_{j \in U \backslash i} \sum_{y_j} p(y_j|x_j) p_{\theta_{+i}}(y_j|x_j) \right) \tag{7}$$

The term θ_{+i} reflects the updated classifier parameters after adding i to the training set with its putative label y_i. Note that we do not know the true distribution of the queried point $p(y_i|x_i)$, or the unlabelled data ($p(y_j|x_j)$ in Eq. (7) above). We approximate these with the current classifier distribution $p(y|x) \approx p_\theta(y|x)$ [12,13,14]. The total dataset size is $\overline{N} = N_l + N_u$.

Explanation of Existing Criteria. The formulation of Eq. (7) is similar to to [13,14] except that we do not constrain ourselves to binary classification. Our approach is also a generalisation of [12] in that we use probabilistic accuracy for utility and evaluate the expectation on all the data instead of only unlabelled data. To understand this, note that if we approximate the utility Eq. (1) on only the unlabelled set, and use the 0/1-accuracy Eq. (6), or the log-loss utility function, we obtain criteria Eqs. (8) and (9) respectively as defined in [12]:

$$\tilde{\mathcal{U}}^{01}_{D^i}(\theta) \propto \sum_{y_i} p_\theta(y_i|x_i) \sum_{j \in U} \max_{y'} p_{\theta_{+i}}(y'|x_j), \tag{8}$$

$$\tilde{\mathcal{U}}^{LL}_{D^i}(\theta) \propto \sum_{y_i} p_\theta(y_i|x_i) \sum_{j \in U} p_{\theta_{+i}}(y_j|x_j) \log p_{\theta_{+i}}(y_j|x_j). \tag{9}$$

If we further assume that the classifier is not updated, so $\theta_{+i} \approx \theta$, then optimising utility gain $\tilde{U}_{D^i}(\theta) - \tilde{U}_D(\theta)$ for i reduces Eqs. (8) and (9) to the most two commonly used [8] variants of uncertainty sampling: minimum certainty Eq. (10) [10] and maximum entropy Eq. (11) [8]:

$$i^*_{01} = \text{argmin}_i p_\theta(\hat{y}_i|x_i), \tag{10}$$

$$i^*_{ll} = \text{argmax}_i - \sum_{y_i} p_\theta(y_i|x_i) \log p_\theta(y_i|x_i). \tag{11}$$

This analysis reveals that commonly used heuristic uncertainty sampling criteria can be seen as weak approximations to expected utility sampling.

2.2 Active Learning and Discovery

The focus of this paper is unifying active learning and discovery. We wish to actively select a dataset from which we can learn to classify a space $\mathcal{X} :\to \mathcal{Y}$ where the set of classes $\mathcal{Y}$ is not known in advance. However, Eqs. (6) and (7) for expected accuracy based querying no longer apply: due to the sum over unknown $y \in \mathcal{Y}$; and importantly because approximating the true class distributions with the current classifier $p(y|x) \approx p_\theta(y|x)$ would now be senseless as the classifier distribution $p_\theta(y|x)$ has a support of known classes L while the true distribution $p(y|x)$ may focus on an unseen class.

Our key idea is to model the true distribution under the Dirichlet process (DP) assumption [19,20] to account for unseen classes. In particular, we use the DP marginal over partitions – the Chinese Restaurant Process (CRP) – as a prior on the true class distribution:

$$p(y_k|y_{1:k-1}) = \begin{cases} n^{y_k}/(k-1+\alpha) & \text{known } y \\ \alpha/(k-1+\alpha) & \text{novel } y \end{cases}, \tag{12}$$

where α is the DP concentration parameter and n^{y_k} is the number of instances of class y_k seen so far. Compared to the general case of learning an *unsupervised*

DP mixture [20], our situation is simpler as we are independently predicting the label posterior of each point i solely based on D_l, the *labelled* data so far. We model the "true" and classifier distributions as Eqs. (13) and (14):

$$p_\theta^{DP}(y|x) \propto \begin{cases} p_\theta(x|y)n^y & \text{known } y \\ p_\theta(x)\alpha & \text{novel } y \end{cases}, \tag{13}$$

$$p_\theta(y|x) \propto p_\theta(x|y)n^y, \tag{14}$$

where the individual class conditional likelihoods $p_\theta(x|y)$ are learned from D_l, and $p_\theta(x)$ is the unconditional density of the entire dataset. Using Eqs. (6) and (13)-(14), we can now generalise the criterion Eq. (7) to active learning with undiscovered classes:

$$\tilde{\mathcal{U}}_{D^i}^{EA}(\theta) \approx \sum_{y_i} p_\theta^{DP}(y_i|x_i)\tilde{\mathcal{U}}_{D\cup(x_i,y_i),U\setminus x_i}^{EA}(\theta) \tag{15}$$

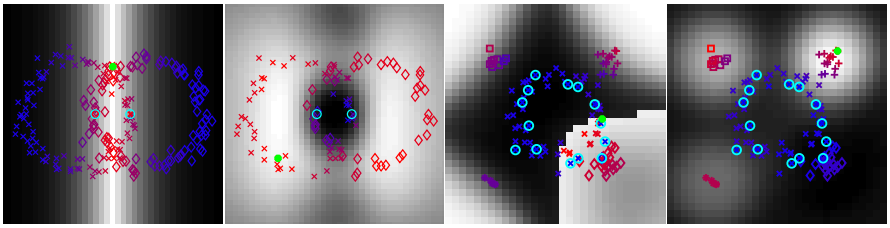
$$= \sum_{y_i} p_\theta^{DP}(y_i|x_i)\frac{1}{N}\left(\sum_{j\in L\cup i} p_{\theta_{+i}}(y_j|x_j) + \sum_{j\in U\setminus i}\sum_{y_j} p_{\theta_{+i}}^{DP}(y_j|x_j)p_{\theta_{+i}}(y_j|x_j)\right) \tag{16}$$

The classifier distribution $p_\theta(y|x)$ ranges over seen classes; but the approximate true class distribution $p_\theta^{DP}(y_j|x_j)$ also accounts for unknown classes. There are therefore two ways for expected accuracy to be penalised – either if a point j is uncertain as before, or if it is likely to represent a new class. In principle, the two sums over y in Eq. (16) cover the full (unknown) set of classes $\mathcal{Y}$. However, since only matches to known classes can have a positive contribution to the utility, the range of these is tractable.

The first sum over the putative labelling y_i of point i is over all classes so far in L plus a new class. In the new class case, the retrained classifier $p_{\theta+i}$ includes one more slot. The second sum over labels y_j of unknown points j is over all classes in $L \cup i$, plus another new class. Importantly, this means that hypothesising a new class for point i can potentially "explain away" any nearby unlabelled js which would otherwise have have low expected accuracy. We name this criterion Dirichlet Process Expected Accuracy (DPEA).

Explanation of pWrong. We investigate the connection of our DPEA criterion to the most effective learning and discovery criterion in the literature, "pWrong" [2]. The pWrong criterion queries the point most likely to be wrong, where a DP posterior is used to model the possibility of (automatically) being wrong due to a given point belonging to an unseen class:

$$p(y_i\, is\, wrong|x_i) = 1 - p_\theta^{DP}(\hat{y}_i|x_i),$$
$$\hat{y}_i = \underset{i\in L}{\text{argmax}} p_\theta(y_i|x_i). \tag{17}$$



(a) Min. Cert. (b) Gen. Error (c) pWrong (d) DPEA

Fig. 2. Illustrative examples of criteria preference for (a) and (b) active learning and (c) and (d) active learning with undiscovered classes. Circles indicate observed points, other symbols indicate classes. Background and symbol shading indicate criteria preferences. Green points are queried. Best viewed in color.

In fact, pWrong approximates our DPEA criterion (Eq. (16)) in an analogous way to the minimum certainty (Eq. (10)) criterion's approximation to the more general expected utility (Eq. (6), Section 2.1). Specifically, if we maximise expected utility gain of querying i with DPEA (Eq. (15)), but under the simpler assumptions of: 0/1 accuracy utility measure (Eq. (3)); evaluating only unseen data; and without classifier updates $\theta \approx \theta_{+i}$, then we have:

$$\begin{aligned} \tilde{\mathcal{U}}^{01}_{D^i} - \tilde{\mathcal{U}}^{01}_{D} \propto & \left(\sum_{j \in U \setminus i} \sum_{y_j} p^{DP}_{\theta}(y_j|x_j) \delta(y, \text{argmax}_{y'} p_{\theta+i}(y'|x_j)) \right) \\ & - \left(\sum_{j \in U} \sum_{y_j} p^{DP}_{\theta}(y_j|x_j) \delta(y, \text{argmax}_{y'} p_{\theta}(y'|x_j)) \right), \\ = & - \sum_{y_i} p^{DP}_{\theta}(y_i|x_i) \delta(y, \text{argmax}_{y'} p_{\theta}(y'|x_i)). \end{aligned} \tag{18}$$

Choosing i^* to maximise Eq. (18) is the same as choosing it to maximise Eq. (17). Therefore pWrong [2] is a rough approximation to our DPEA criterion. Moreover, it can be seen as the discovery and learning problem analogue of the popular minimum certainty criterion for vanilla active learning (Eq. (10), [10]).

Illustrative Example. Let us first provide some insight into the criteria discussed by way of synthetic examples. Figs. 2 (a) and (b) contrast minimum certainty and generalisation error active learning (Section 2.1) for a simple but non-separable dataset of two classes (crosses and diamonds). One point from each class is initially labeled (overlaid circles). The lightness of the background (and red-blue shading of the data points) indicate the preference of each criterion, and the starred point indicates the selected point. Minimum certainty (Fig. 2(a)) depends only on the two labeled points and simply prefers points at

the current decision boundary. Generalisation error (Fig. 2(d)) makes a more holistic assessment based on the full pool of data where, in noticeable contrast to minimum certainty, it avoids the overlapped decision boundary region (for which any label outcome is of limited use), and instead focuses on the outer region where a label could usefully disambiguate numerous points.

Figs. 2(c) and (d) contrast pWrong and DPEA discovery and learning criteria for a simple dataset with a large majority class (crosses) and four smaller minority classes (other symbols). A number of majority class points and one minority class point are labeled (circles). In this example pWrong (Fig. 2(c)) focuses primarily (bright background, bright red symbols) on the decision boundary between the existing labeled classes: again a very myopic preference literally about which classification is likely to be wrong. Obtaining labels in regions of high overlap may not be the most efficient way to improve performance. At the same time, the bright corners illustrate the tendency of pWrong to blindly query outliers (since anything with low conditional density is likely wrong). In contrast, DPEA (Fig. 2(d)) ignores (for now) the decision boundary between the known classes in favour of the undiscovered regions: because obtaining a label here would disambiguate more points and increase expected accuracy more quickly.

2.3 Implementation Details

Various density models could be used for the class conditional likelihoods $p_\theta(x|y)$ in our framework. A clean approach would be to use nested DP mixtures, where each likelihood is itself modelled as an (unsupervised) Dirichlet process mixture; however this would be prohibitively expensive. For computational efficiency, and for clearer comparison to prior work, we therefore model the class likelihoods using a mixture of Gaussians learned by the constant time incremental approximation described in [21] and used for active learning and discovery in [2,1].

Naively, the computational complexity of the DPEA criterion is $\mathcal{O}\left(N^2C^2KD^3\right)$ where N is the pool size, C is the number of classes, K is the number of Gaussians in the likelihood mixture and D is the dimension of the data. However, the GMM framework used permits a number of incremental computations: the kernel distance matrix only needs to be rebuilt for definite additions rather than putative additions (16); by caching the likelihood of the pool for each class, the data likelihood only needs to be evaluated under the hypothesised class; by caching the response of the data to each Gaussian component, re-evaluation is only needed against a single kernel rather than the whole mixture after each update; and the Cholskey decomposition of each component's covariance can be cached and only recomputed after an update. This results in complexity $\mathcal{O}\left(N^2C(C+D^2+K)\right)$. Finally, for the largest datasets, we use the (similarly optimised) $\mathcal{O}\left(N(C+D^2+K)\right)$ pWrong to filter a subset of P points to be considered by the full algorithm at each iteration[1], and we only consider the top Q most likely hypothetical class labels in approximating the outer sum in (16). The final result is therefore also complexity $\mathcal{O}\left(N(C+D^2+K)\right)$.

[1] Since pWrong is an approximation to DPEA.

Algorithm 1. Active learning with undiscovered classes

Input: Initial labeled L and unlabeled U samples.

1. Build unconditional GMM $p(\mathbf{x})$ from $L \cup U$
2. Estimate σ by LOO cross-validation on $p_\theta(\mathbf{x})$

Repeat as training budget allows:

1. Train classifier $p_\theta(y|x)$ on L
2. For each point $i \in U$ (16)
   (a) Infer class posterior $p_\theta^{DP}(y_i|x_i)$
   (b) For each potential class y_i
      i. Update classifier θ with (x_i, y_i)
      ii. Evaluate expected accuracy with $\theta_{+(x_i,y_i)}$
   (c) Compute overall expected accuracy $\tilde{\mathcal{U}}_{D^i}(\theta)$
3. Query $i^* = \text{argmax}_i \tilde{\mathcal{U}}_{D^i}(\theta)$

Algorithm 1 summarises the procedural steps required to realise our DPEA framework.

The only parameter in our learning and discovery model is the DP concentration α, which reflects prior belief about the concentration of classes. We learn α at each iteration with the method [22]. We learn the GMM likelihood's prior scale parameter σ using leave-one-out cross-validation on the full unlabelled training dataset (as in [21,2,1]) prior to estimating the unconditional density $p(X)$. To constrain the computation, we set a GMM component cap of $K = 32$ [2,1].

3 Experiments

In this section we compare the performance of our criterion DPEA to a variety of alternatives. We followed [2,1] in experimental procedure: Evaluation was based on the average classification accuracy per class (ensuring the accuracy at classifying each rare class is weighted equally with the majority class) summarised over the area under the learning curve (AUC). Each dataset was evaluated by two-fold cross-validation – training on one fold starting from a single labeled example of the majority class and proceeding for 150 active learning iterations, while testing on the held-out fold. This procedure was repeated 25 times and the results averaged. We evaluated 15 public datasets. Eleven were UCI datasets chosen for containing naturally unbalanced class proportions. Three were vision datasets (digits, gait and letters) previously subsampled to contain geometric class proportions [2,1]. Finally, we studied two large scale vision datasets with naturally unbalanced proportions: Yahoo faces in the news [6], from which we pruned persons with less than 10 faces each and reduced the dimension of the provided features to 32 using PCA, and SUN scene recognition [7]. The largest datasets involve an order of magnitude more classes than previous active

Table 1. Dataset properties. (N) number of instances. (d) dimension of data. (N_c) number of classes. (S%/L%) proportions of smallest and largest classes.

Dataset	N	d	N_c	S%	L%	Dataset	N	d	N_c	S%	L%
Glass	214	10	6	4	36	Winequality	4893	11	6	.37	45
Ecoli	336	7	8	1.5	42	Letters	9276	16	26	.3	14
Yeast	1484	8	10	.27	31	Shuttle	20000	9	7	.01	78
Segment	635	18	7	1	48	KDD99	33650	23	15	.04	51
Pageblocks	5473	10	5	.5	90	CASIA Gait	2353	25	9	3	49
Covertype	5000	10	7	3.6	25	MNIST digits	13000	25	10	.1	50
Thyroid	7200	21	3	2.5	92	Faces [6]	10390	32	330	0.04	11
						SUN Scene [7]	108754	32	397	.1	2

discovery papers have considered [2,1,17,3,16], and we evaluated 1000 iterations. The properties of each dataset are detailed in Table 1.

For comparison, we consider the two previously best performing active discovery and learning models, Adapt [1,23] and pWrong [2]. We also include two new alternatives composed by running one of two state of the art active discovery models, NNDM [16] and RADAR [18], for 1/3 of the query budget, followed by minimum certainty (Eq (10)) querying for the remaining 2/3s.

3.1 Results

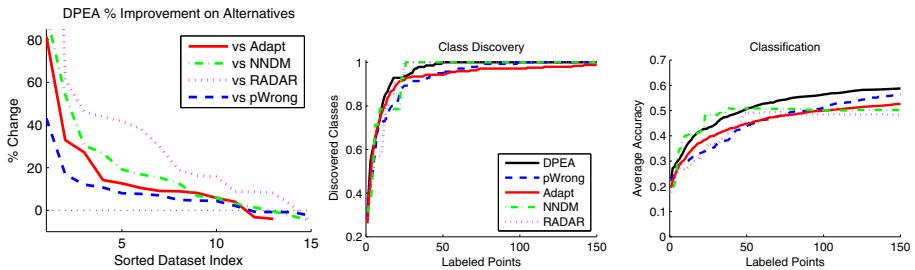
The classification AUC for all 15 datasets are reported in Table 2, along with the average over all datasets and the number of wins by each model. Our criterion, DPEA (16) scored the highest average and the most wins by a large margin, followed by pWrong [2] and the others. Note that the models in [1] and [2] have already been shown to decisively improve on random sampling and iterating discovery and learning criteria [3].

The absolute difference in the *mean* performance of DPEA and the alternatives is 4-11% (Table 2, average). However, these figures miss some important points: DPEA often performs similarly to pWrong, but it is rarely notably worse (max 1%) and often significantly better. To illustrate this, Fig. 3(a) plots the change in performance of DPEA relative to the alternatives as a percentage. Clearly, while performance is similar in many cases, it is frequently significantly better, (e.g., 43% improvement on pWrong for Gait, 82% improvement on Adaptive for Digits, 300% improvement vs RADAR on KDD, 90% improvement vs NNDM on KDD). The mean *improvement* on the alternatives is 8%, 16%, 18% and 44%. The improvement relative to pWrong can be understood because (as discussed in Section 2.2) by making very local decisions about individual points' uncertainty, pWrong risks focusing on "impossible" points in overlapped regions of space. DPEA in contrast, has greater robustness to such situations by considering the impact that labelling each point would have on the classification of the other points and hence is better at querying points which will have actual impact on performance. This difference in robustness explains the significant lead in number of wins by DPEA, made up of similar performance in some datasets,

Table 2. Area under classification curve for our model vs [1], [2], [16], [18]

Classify	Adapt[1]	NNDM[16]	RADAR[18]	pWrong[2]	DPEA
Glass	65	66	65	**71**	70
Ecoli	61	62	**63**	59	59
Yeast	42	41	35	45	**46**
Segment	67	74	62	74	**74**
P. blocks	59	53	58	60	**63**
C. type	46	48	44	46	**51**
Thyroid	**59**	44	40	54	56
Wine	23	24	22	23	**24**
Letters	28	24	23	**38**	37
Shuttle	42	36	47	43	**48**
KDD99	58	38	17	63	**74**
Gait	57	**61**	43	42	60
Digits	28	44	36	49	**51**
Faces	-	13	9	12	**14**
Scene	-	1.2	1.2	1.3	**1.4**
Average	42	42	38	45	**49**
Wins	1	1	1	2	**10**

and significant improvement in other datasets. We note also that an even greater improvement is made over Adapt [1], despite the more powerful discriminative SVM classifier available to the latter. Moreover, one further issue with strategies based on SVMs is that their typical $\mathcal{O}(C^2)$ training complexity renders them slow for large multi-class datasets [6,7] compared to the other $\mathcal{O}(1)$ generative models tested. Indeed [1] could not complete the full faces dataset due to the libsvm component failing with hundreds of classes. Our matlab implementation of DPEA proceeded at about 1sec per iteration for most of the datasets.



(a) DPEA Performance (b) Learning Example

Fig. 3. (a) Percentage improvement of AUC for DPEA over prior models [1,2,16,18]. (b) Illustrative learning and discovery curves.

Table 3. Area under discovery curve for our model vs [1], [2], [16], [18]

Discover	Adapt[1]	NNDM[16]	RADAR[18]	pWrong[2]	DPEA
Glass	95	95	94	**96**	95
Ecoli	91	86	86	93	**93**
Yeast	88	86	82	**92**	91
Segment	94	93	92	**96**	93
P. blocks	95	94	97	**99**	96
C. type	93	95	95	93	**96**
Thyroid	93	91	95	**96**	93
Wine	93	80	84	**94**	93
Letters	70	66	71	**85**	74
Shuttle	**90**	74	85	87	74
KDD99	78	65	60	83	**86**
Gait	94	95	95	89	**96**
Digits	48	**85**	76	79	78
Faces	-	53	**59**	58	58
Scene	-	54	**56**	49	50
Average	86	81	82	**86**	84
Wins	1	1	2	**7**	4

As an illustrative example, Fig. 3(b) shows the full learning curve for a vision dataset (covertype). Here all the methods are competitive at discovery, and NNDM is initially best at classification, however all the more myopic criteria are eventually outperformed by our DPEA. To provide additional insight, we also present the area under the discovery curve (how quickly all classes are discovered) in Table 3. We reiterate that the objective is actively learning to classify in the presence of unknown classes, not simply discovery per se. Interestingly, pWrong is the best at discovery despite being weaker at classification, while DPEA is weaker at discovery despite being best at classification. This highlights the important point that there are multiple ways to improve performance in the active learning with undiscovered classes context: by discovering new classes, or learning to classify existing classes better. Adapt [1] addresses this with explicit heuristics to balance these two sub-goals. pWrong aims to provide a single objective, but its myopic outlier-preferring approach (Sec. 2.2) is biased suboptimally in favour of discovery given its good performance there but poorer overall performance (Table 2). The bigger picture view of DPEA explains why it can under-perform at the discovery subtask, while still performing best in the overall classification task. NNDM and RADAR are sometimes worse and sometimes better at discovery (e.g., letters vs digits). This illustrates the serious issue with heuristic sequential combinations, i.e., knowing when to terminate the discovery phase and start the learning phase, and hence the value in direct optimisation of our unified parameter-free criterion.

4 Conclusion

Summary In this study we proposed the first unified framework for the problem of active learning in the presence of undiscovered classes by querying points which maximise the expected probabilistic accuracy of a classifier – where the true averaging distribution is estimated based on a Dirichlet Process mixture. This is in contrast to most previous work which has focused on heuristic combinations of active discovery and active learning criteria. In the process we show explicitly, and to our knowledge for the first time, how various popular active learning criteria such as maximum entropy and minimum certainty are special cases of – or approximations to – a more general expected utility criterion. Notably, the best previous attempt at active learning and discovery, pWrong [2], turns out to be an approximation to our full framework. We also detailed the exploitation of incremental computation required to speed up our DPEA criterion to the same computational complexity as pWrong, albeit with larger constant. Finally, we performed the largest empirical evaluation of active learning with undiscovered classes to date. The value of our unified model is shown by consistently and often significantly outperforming alternatives with sub-optimal heuristic objectives and free parameters. DPEA is therefore of significant value for interactive modelling of new domains where supervision is expensive.

Future Work. Although DPEA often significantly outperforms pWrong, either can be recommended for the discovery and learning task depending on the relative expense of computer time and supervision in a given application. It may also be possible to develop models sensitive to both annotation and processing costs [24] which can automatically determine the ideal criterion to use. Another interesting question is customising the objective function to the goal of the task. Here we optimised overall expected accuracy (Eq. (1)), but any potential objective of interest could potentially be optimised (e.g., average accuracy per class, precision, recall, f-measure). A related unaddressed but important question is the significance of the common practice of optimising a particular (e.g., discriminative) query criterion, while the underlying models are trained with a different (e.g., generative maximum likelihood) criterion as in [12,10]. Better results may be obtained if the query criteria and underlying models can be aligned such that they are optimised with the same objective. A final avenue for future work is generalising our unified perspective to stream based discovery and learning [25].

References

1. Hospedales, T.M., Gong, S., Xiang, T.: Finding Rare Classes: Adapting Generative and Discriminative Models in Active Learning. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part II. LNCS, vol. 6635, pp. 296–308. Springer, Heidelberg (2011)
2. Haines, T., Xiang, T.: Active learning using dps for rare class discovery and classification. In: BMVC (2011)

3. Pelleg, D., Moore, A.: Active learning for anomaly and rare-category detection. In: NIPS (2004)
4. Stokes, J.W., Platt, J.C., Kravis, J., Shilman, M.: Aladin: Active learning of anomalies to detect intrusions. Technical Report 2008-24, MSR (2008)
5. Hospedales, T., Gong, S., Xiang, T.: Video behaviour mining using a dynamic topic model. International Journal of Computer Vision 98, 303–323 (2012)
6. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: CVPR (2009)
7. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR, pp. 3485–3492 (2010)
8. Settles, B.: Active learning literature survey. Technical Report 1648, University of wisconsin–Madison (2009)
9. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research 2, 45–66 (2001)
10. Jain, P., Kapoor, A.: Active learning for large multi-class problems. In: CVPR (2009)
11. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of Artificial Intelligence Research, 129–145 (1996)
12. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: ICML, pp. 441–448 (2001)
13. Kapoor, A., Horvitz, E., Basu, S.: Selective supervision: Guiding supervised learning with decision-theoretic active learning. In: IJCAI (2007)
14. Vijayanarasimhan, S., Grauman, K.: Multi-level active prediction of useful image annotations for recognition. In: NIPS (2008)
15. Beygelzimer, A., Hsu, D., Langford, J., Zhang, T.: Agnostic active learning without constraints. In: NIPS (2010)
16. He, J., Carbonell, J.: Nearest-neighbor-based active learning for rare category detection. In: NIPS (2007)
17. Vatturi, P., Wong, W.K.: Category detection using hierarchical mean shift. In: KDD, pp. 847–856 (2009)
18. Huang, H., He, Q., He, J., Ma, L.: RADAR: Rare Category Detection via Computation of Boundary Degree. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part II. LNCS, vol. 6635, pp. 258–269. Springer, Heidelberg (2011)
19. Antoniak, C.E.: Mixtures of dirichlet processes with applications to bayesian nonparametric problems. Annals of Statistics 2(6), 1152–1174 (1974)
20. Rasmussen, C.: The infinite gaussian mixture model. In: NIPS (2000)
21. Sillito, R., Fisher, R.: Incremental One-Class Learning with Bounded Computational Complexity. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007. LNCS, vol. 4668, pp. 58–67. Springer, Heidelberg (2007)
22. Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90(430), 577–588 (1995)
23. Hospedales, T., Gong, S., Xiang, T.: Finding rare classes: Active learning with generative and discriminative models. In: IEEE TKDE (preprint)
24. Vijayanarasimhan, S., Grauman, K.: Cost-sensitive active visual category learning. International Journal of Computer Vision 91, 24–44 (2011)
25. Loy, C.C., Hospedales, T.M., Xiang, T., Gong, S.: Stream-based joint exploration-exploitation active learning. In: CVPR (2012)