Stream-based Joint Exploration-Exploitation Active Learning

Chen Change Loy[1], Timothy M. Hospedales[2], Tao Xiang[2], and Shaogang Gong[2]
[1]Vision Semantics Ltd, UK
[2]School of EECS, Queen Mary University of London, UK
ccloy@visionsemantics.com {tmh,txiang,sgg}@eecs.qmul.ac.uk

Abstract

Learning from streams of evolving and unbounded data is an important problem, for example in visual surveillance or internet scale data. For such large and evolving real-world data, exhaustive supervision is impractical, particularly so when the full space of classes is not known in advance therefore joint class discovery (exploration) and boundary learning (exploitation) becomes critical. Active learning has shown promise in jointly optimising exploration-exploitation with minimal human supervision. However, existing active learning methods either rely on heuristic multi-criteria weighting or are limited to batch processing. In this paper, we present a new unified framework for joint exploration-exploitation active learning in streams without any heuristic weighting. Extensive evaluation on classification of various image and surveillance video datasets demonstrates the superiority of our framework over existing methods.

1. Introduction

We consider a learning problem defined as follows. There are potentially unlimited instances streamed sequentially, but limited at any given time. Typically, a learner receives an instance at a time and cannot store or re-process all the past instances due to constraints such as memory limitation. Importantly, the overall class frequency exhibits a power-law or Zipf's law distribution [32, 20] – most instances in the data stream belong to several large classes, whilst many other classes only contribute a small portion in the whole data stream (see Fig. 1). As such, many minor classes are often unknown due to their rarity and unpredictability.

This learning problem is frequently encountered in many real-world scenarios [13, 8, 20]. For example, in visual surveillance applications, interesting and potentially dangerous activities are rare compared to typical behaviour, and occur in new and unanticipated forms which need be discovered and modelled. In such applications, data arrive in a stream over an extended period of time but not all of them can be stored [28]. Another example is large scale learning tasks, e.g. crowdsourcing, where content such as images or videos with potentially unknown classes are presented sequentially to human annotators. In such a case even a single scan over all the unlabelled data is not scalable, so stream-based data processing is required to support on-the-fly interactive labelling [29]. The common characteristics of these learning problems are *stream-based learning, unknown class discovery, ambiguity reasoning*, and *imbalanced class distributions*. Yet existing learning strategies fail to address these jointly (see Sec. 2).

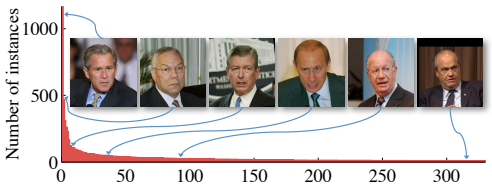

Figure 1. In many vision problems (e.g. classification of face images on the internet), the class frequency is distributed in a heavy-tailed power-law fashion [32].

Our goal is to formulate a stream-based learning framework capable of performing *active joint exploration-exploitation*: discovering *a-priori* unknown and rare classes (exploration) and learning the concept boundary (exploitation) by searching for instances with ambiguous class membership. The learner should return unknown and ambiguous instances to be labelled by a human on-the-fly, such that the labelling efforts needed to improve the model is minimal. In addition to that, the learner must be able to handle realistic imbalanced class distributions.

Specifically, we formulate a unified criterion based on the principle of committee consensus [22], using a Bayesian nonparametric framework to balance the two competing goals of exploration and exploitation without heuristic weighting. To handle imbalanced class distributions, we define a nonparametric Bayesian prior on known and unknown classes using the Pitman-Yor Processes (PYP) [17], which can produce heavy-tailed power-law distributions.

We show in our extensive experiments that the proposed active learning framework reduces labelling cost compared to passive random labelling and outperforms other state-of-the-art stream-based active learning methods. We also show that the PYP prior is superior in handling imbalanced class distributions compared to the Dirichlet Processes (DP) prior [27].

2. Related Work

Various strategies have been proposed to reduce human supervision in classifier learning, e.g. semi-supervised learning, weakly supervised learning, active learning, and transfer learning. Active learning [21] is useful in lowering labelling cost by requesting human labeling of only informative instances [4, 14, 10, 13, 6, 9, 29, 11]. In this paper, we focus on stream-based active learning, particularly in the hard case, where the full set of classes is not known in advance and the overall class distribution is imbalanced.

Stream-based active learning: Most existing methods consider a *pool-based* setting, where each query selection is performed via exhaustive search in an unlabelled data pool [19, 10, 6, 9, 29, 11]. In contrast, a *stream-based* learner makes immediate query decisions at each instance during a single scan of the data stream. A stream-based learning process has a clear advantage over pool-based: no expensive search in the data pool is needed. It is therefore feasible for applications that demand on-the-fly interactive labelling such as crowdsourcing [29] and visual surveillance tasks [13]. Stream-based learning, however, is harder than the pool-based learning since the learner lacks complete knowledge on the underlying data distribution [8].

Different stream-based approaches have been proposed [4, 1, 8, 13], most of which require one to set an arbitrary threshold on the query criterion to decide whether to query or discard an incoming instance. An exception is the Query-by-Committee (QBC) algorithm [22], which exploits the consensus of an ensemble of committee members to achieve threshold-free query selection. We follow the QBC intuition, inheriting the appealing threshold-free property. However, the original QBC algorithm is not designed for discovering unknown classes. In addition, previous QBC method [1] is only applicable to discrete data with binomial and multinomial likelihoods. In this study, we extend the QBC paradigm to discover unknown classes and to handle multivariate normal likelihoods, commonly encountered in vision problems [18].

Query criteria and unknown class discovery: The uncertainty criterion [10, 29, 11] is widely used to identify instances with ambiguous class membership. It is usually quantified by entropy of class posterior [11] or distance from the decision hyper-plane [29], with the assumption that all the classes are known *a priori*. The assumption is invalid in many stream-based learning problems with unknown and rare classes. To relax this assumption, active discovery of unknown classes has been studied [16, 7], e.g. by selecting instance that minimise the likelihood given the model. Recent studies [24, 13, 6, 9] have attempted to combine uncertainty and low-likelihood criteria for achieving joint exploitation-exploration. These methods, however, have two drawbacks: (1) they are either limited to pool-based learning [24, 6, 9], or (2) rely on heuristic switching between criteria controlled by some ad-hoc parameters [24, 13, 9].

Imbalanced class distribution handling: Previous work [6] discovers unknown classes using DP class priors, which is not ideal to handle imbalanced class distribution. In this paper, we generalise the previously proposed DP to the PYP to address the shortcoming. In contrast to DP, the PYP produces power-law distributions more closely resembling those seen in real-world problems, such as natural language modelling [26] and natural scene segmentation [23], which have shown significant improvement by using PYP over DP, even though PYP was not considered nor formulated for active learning.

The contributions of this paper are three-fold: (1) we show for the first time how stream-based joint exploration-exploitation can be achieved using a unified active learning criterion. The proposed method makes immediate query decisions at each instance, and is thus computationally suitable for streaming data and large-scale learning tasks that demand on-the-fly interactive labelling; (2) our method is formulated as a nonparametric Bayesian model that adapts its complexity and exploration-exploitation balance to the data, without any heuristics or manual tuning of parameters. Its performance is thus reliable and stable across many datasets; (3) we leverage the PYP as a class prior in active learning, which gives improved performance on real-world long-tailed problems for which existing methods are weak.

Extensive evaluation is performed on numerous benchmark and visual databases, including public surveillance video, human gait, handwriting, faces, and natural images.

3. Stream-based Active Learning

The proposed model aims to minimise human labelling effort required to learn a model for classes $c \in \mathcal{C}$, where the full class set $\mathcal{C}$ is not known in advance. Conventional stream-based learning methods [1, 8] do not apply in this more general case. Calculating the probability that an instance belongs to an unknown class is non-trivial, as nothing is currently known about the class. We present a solution to this problem based on the Pitman-Yor Processes (PYP).

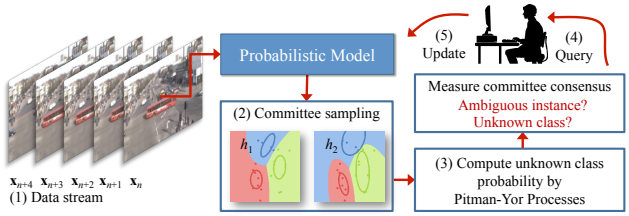

Figure 2. Overview of the proposed approach.

3.1. Model Overview

We first give a brief overview of our approach before going into detail in subsequent sections. We consider a generative classifier of the form $p(c|\mathbf{x}) \propto p(\mathbf{x}|c)p(c)$. Undiscovered classes are accounted for by assuming a PYP prior model $p(c)$. The likelihood for each class, $p(\mathbf{x}|c)$ is itself modelled by a DP mixture of Gaussians, allowing for classes of arbitrary complexity.

For stream-based active learning (see Fig. 2) we iteratively: (1) receive an instance $\mathbf{x}_n \in \mathbb{R}^d$ from the data stream; (2) draw two random hypotheses h_1 and h_2 from the model posterior to form a committee. (3) for each hypothesis, compute the posterior $p(c_n|\mathbf{x}_n)$ under the PYP assumption; (4) query the instance if the two hypotheses disagree on its classification, or they both assign the instance to an unknown class; (5) include the labelled $\{(\mathbf{x}_n, c_n)\}$ in the training set $\mathcal{L}$ to refine the classifier; The iteration stops when a criterion is met, e.g.the query budget is exhausted.

3.2. Sampling a Committee

Let us first explain the mechanism for forming a committee. A committee consists of multiple random hypotheses sampled from the posterior distribution over the model parameters, $p(\theta|\mathcal{L})$ conditioned by the training set $\mathcal{L}$ labelled so far. How to sample $p(\theta|\mathcal{L})$ depends on the parametric form of the model. Previous QBC studies [1, 14, 13] only cover multinomial likelihoods: an unrealistic constraint for real-world data, especially in vision. Here, we discuss the sampling process for models with multivariate normal likelihoods.

A d-dimensional multivariate normal distribution has parameters: mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The conjugate prior for the multivariate normal is the *normal inverse Wishart distribution* ($\mathcal{N}$IW). The $\mathcal{N}$IW depends on parameters ν_0, κ_0, $\boldsymbol{\mu}_0$, Λ_0, where ν_0 and κ_0 are positive scalars, $\boldsymbol{\mu}_0$ is a $d \times 1$ vector and Λ_0 is a $d \times d$ matrix. After observing n points, we obtain the $\mathcal{N}$IW posterior (see [15]), from which we can sample specific committee members $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$:

$$\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} \sim \mathcal{N}\text{IW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\nu_n, \kappa_n, \boldsymbol{\mu}_n, \Lambda_n) \quad (1)$$

$$\text{where} \quad \boldsymbol{\Sigma} \sim \text{IW}_{\nu_n}(\Lambda_n^{-1}) \quad (2)$$

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/\kappa_n). \quad (3)$$

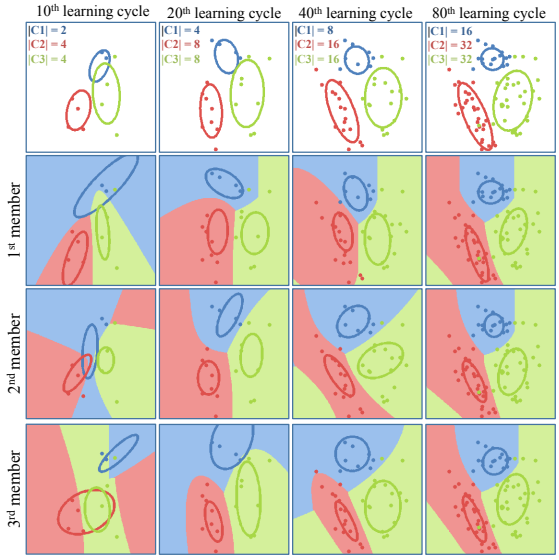

Figure 3. Committee members with different decision boundaries can be formed by sampling different $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ estimates from the posterior probability distribution over the model parameters. Sampling the posterior produces members whose parameters estimate differ most when the number of data n is low and tend to agree when n is high.

In a committee, each member is a standalone classifier modelling K classes, each with a Gaussian $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ or mixture of Gaussians $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ likelihoods (Eq. 1). Figure 3 illustrates the states of three members during active learning. Each member models three classes, each of which is parameterised by a single Gaussian $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. One can observe that when the number of observed data is low (e.g.$n = 10$), the variance of these estimates is large. Therefore, the decision boundaries vary and members tend to disagree among each other, yielding a large uncertain region in the decision space. Members tend to agree after observing more data (e.g. $n = 80$) when the variance of $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ estimates becomes smaller. Crucially, by querying instances lying in the disagreement space, the QBC framework not only helps to ascertain ambiguous class membership, but also implicitly reduces the model variance.

3.3. Unknown Class Probability

The classic QBC algorithm solely addresses inter-class uncertainty among known classes (see Sec. 3.4). In this section, we generalise QBC to account for uncertainty due to unknown classes in unexplored space using non-parametric Bayes.

We assume the data sequence to be drawn i.i.d. from a random probability distribution G, which is PYP distributed with base distribution G_0, a discount parameter $0 \leq \beta < 1$, and a concentration parameter $\alpha > -\beta$., written as

$$G|G_0, \beta, \alpha \sim \text{PY}(\beta, \alpha, G_0). \quad (4)$$

The base distribution G_0 can be understood as the mean of the probability distribution G, whilst β and α are parameters that control the amount of variability of G from G_0.

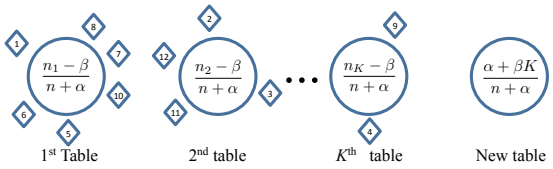

Figure 4. Pitman-Yor Processes with (α, β) seating plan, where numbered diamonds represent customers, and the large circles represent tables.

In PYP, the number of classes (partitions) is assumed to be infinite; while a prior is specified over the classes such that only a finite are observed. The prior can be explained using the Chinese restaurant process metaphor (see Fig. 4). Specifically, a restaurant has an infinite number of tables with K already occupied, and n_k customers at the kth table. The total of customers in the restaurant, including a new customer is $n = \sum_k^K n_k$. A new customer prefers popular tables, going to occupied table k with probability $(n_k - \beta)/(n - 1 + \alpha)$; or a new table with probability $(\alpha + \beta K)/(n - 1 + \alpha)$.

The PYP is exploited in our framework in that an unbounded number of discrete classes (tables) is considered, of which K have been seen so far. A new instance (customer) can be grouped into either an existing class $k \leq K$, or a new class $K + 1$. Formally, the existing and new class probabilities for a new instance $\mathbf{x}_n$ are:

$$p(c = k|\mathbf{x}_n) \propto \begin{cases} \frac{n_k-\beta}{n-1+\alpha} p(\mathbf{x}_n|c = k) & \text{if } k \leq K \\ \frac{\alpha+\beta K}{n-1+\alpha} p(\mathbf{x}_n) & \text{if } k = K+1 \end{cases}, \quad (5)$$

where n_k is the number of instances in the kth class. Normalising $p(c = k|\mathbf{x}_n)$ gives the probability of the new $\mathbf{x}_n$ belonging to each of the known classes as well as to an unknown class. More details on Eq. 5 are given as follows:

Pitman-Yor hyperparameters, α, β - Instead of fixing the values of α and β, we treat them as unknown parameters by putting hyper priors over them and infer their values following the sampling routine described in [25].

Obtaining $p(\mathbf{x}_n|c)$ - Recall that we sample a set of committee members from the (infinite) DP mixture of Gaussians representing each class c. Each member is therefore a (finite) Gaussian Mixture Model (GMM), for which $p(\mathbf{x}_n|c)$ is

$$p(\mathbf{x}_n|c) = \sum_{m=1}^{M_c} \pi_{c,m} \mathcal{N}(\boldsymbol{\mu}_{c,m}, \boldsymbol{\Sigma}_{c,m}), \quad (6)$$

where M_c is the number of components, and $0 \leq \pi_{c,m} \leq 1$ are mixing coefficient. More details are given in Sec. 3.5.

Obtaining $p(\mathbf{x}_n)$ - The unconditional density, $p(\mathbf{x})$ distinguishes useful samples from outliers [6]. Prior to active learning, we construct a Gaussian model of $p(\mathbf{x})$ from unlabelled samples randomly drawn from the stream.

Equation 5 encodes two main properties of the PYP [31]: (1) 'rich-gets-richer' clustering property – the class prior is proportional to the number of instances already assigned to it, so the PYP is likely to assign data to an existing class with a large number of samples. (2) 'stealing from the rich and giving to the poor' behaviour – PYP distributes some weight from large classes to potential new class through $\beta > 0$. That is, the more classes we observe, the more likely that data will be assigned to a new class. This produces a heavy-tailed power-law distribution that resembles those seen in real-world problems [26, 23]. Dirichlet Processes with $\beta = 0$ only exhibits the first but not the second property.

3.4. Measuring Committee Consensus

After sampling the committee members, for each member we can classify a new instance $\mathbf{x}_n$ by maximising $p(c = k|\mathbf{x}_n)$ in Eq. 5. The committee may have different opinions in assigning the instance to a new or existing class. Assuming two hypotheses h_1 and h_2, we measure their consensus as follows:

$$\{h_1(\mathbf{x}_n) \neq h_2(\mathbf{x}_n)\} \vee \{h_i(\mathbf{x}_n) = K+1 \mid \forall i\} \quad (7)$$

where K is the number of classes observed so far. This achieves (1) exploration when h_1 and h_2 assign the instance to a new class, that is $\{h_i(\mathbf{x}_n) = K+1 \mid \forall i\}$. (2) exploitation when both hypotheses disagree on the classification, that is $\{h_1(\mathbf{x}_n) \neq h_2(\mathbf{x}_n)\}$. Figure 5 illustrates the qualitative difference between the QBC and our approach in predictions relating to new class querying. As can be seen, the QBC solely addresses inter-class uncertainty among known classes, whilst our algorithm addresses uncertainty among known classes and unexplored space. Importantly, this threshold-free query strategy is in contrast to typical stream-based methods [8, 13], which require an explicit threshold to filter out instances with low query preference.

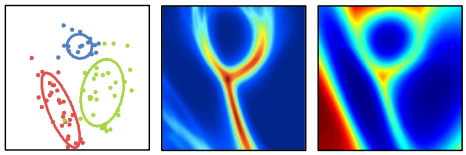

Figure 5. The same toy data example shown in Fig. 3 is used here. This figures shows the qualitative comparison of uncertainty regions addressed by both the QBC and the proposed approach. Red colour suggest high query preference, whilst blue colour suggests low query preference.

As we shall see in Sec. 4.2, exploration and exploitation are automatically balanced by the PYP partition prior, given

hyperparameters α and β. Specifically, the value of $\frac{\alpha+\beta K}{n-1+\alpha}$ in Eq. 5 tends to decrease, whilst $\frac{n_k-\beta}{n-1+\alpha}$ tends to increase when n grows larger. As a consequence, new class probability will decrease gradually after more points are observed. The exploration therefore dominates at the beginning and the active selection will slowly switch to exploitation. However, crucially, the above balancing behaviour is (1) free from heuristic parameters and (2) occurs at a data-driven rate via the learning of the PYP hyperparameters.

It is worth noting that one can extend QBC to multiple members and using different consensus metric such as vote entropy [1]. However, such extensions often require one to set an explicit threshold as a cut-off point on the consensus level. In addition, various studies [22, 1] have shown that adopting different committee sizes does not improve the performance significantly. We found that that using a two-member committee was sufficient in our experiments.

3.5. Incremental Dirichlet Process Mixture Model

Our framework can be easily adapted to any likelihood model, e.g. [9, 6]. In this study, we employ an incremental Dirichlet process mixture model (DPMM) [5] as our classifier, which we modified to cater for multi-class classification. Specifically, if K classes are observed, there will be K DP mixtures conditioned on a class variable.

The nonparametric Bayesian model offers two key advantages over traditional Bayesian (and non-Bayesian) models: (1) it allows an unbounded number of latent mixture components under each class, thus relaxing the component number restriction as in [6, 9]; and (2) it naturally models the posterior distribution over parameters that can be exploited directly in the committee sampling described in Sec. 3.2.

After each query, the labelled instance is used to update the DPMM incrementally using a memory-bounded variational inference strategy [5]. To obtain the class assignment of a test instance $\mathbf{x}^*$, we first compute the predictive distribution of each class $p(\mathbf{x}^*|\mathcal{L}_c)$ (see [12]-Eq. (7)), where $\mathcal{L}_c$ is the labelled data seen by c-th class. Assuming a uniform class prior, the most probable class is then $\hat{c}_n = \text{argmax}_{c\in\{1,\dots,K\}}\, p(\mathbf{x}^*|\mathcal{L}_c)$.

The proposed stream-based active learning method is summarised in Alg. 1.

3.6. Implementation Details

For learning the PYP hyperparameters, α and β, we fixed the hyper priors a_α, b_α, a_β, and b_β to 1; and we used 128 iterations for both the burn-in and posterior sample collection. For initialisation, α and β of previous learning cycle were used. The code of the proposed approach is available at http://www.eecs.qmul.ac.uk/~ccloy/.

Algorithm 1: Stream-based active learning for joint exploration-exploitation.

Input: Unlabelled data stream $\mathcal{U} = (\mathbf{x}_1, \dots, \mathbf{x}_n, \dots)$.
Output: A set of labelled samples $\mathcal{L}_n$ and a classifier $\mathcal{M}_n$ trained with $\mathcal{L}_n$.

repeat
  Receive $\mathbf{x}_n$;
  Sample two random hypotheses, h_1, h_2 from $p(\theta|\mathcal{L})$ [Eq. 1];
  Compute $p(c=k|\mathbf{x}_n)$ for h_1 and h_2, with $k \in \{1, \dots, K+1\}$ [Eq. 5];
  Compute $h_i(\mathbf{x}_n) = \text{argmax}_c\, p(c|\mathbf{x}_n)$, $i = 1, 2$;
  if $\{h_1(\mathbf{x}_n) \neq h_2(\mathbf{x}_n)\} \vee \{h_i(\mathbf{x}_n) = K+1 \mid \forall i\}$ **then**
    Request c_n and set $\mathcal{L}_n = \mathcal{L}_{n-1} \bigcup \{(\mathbf{x}_n, c_n)\}$;
    Obtain $\mathcal{M}_{n+1}$ by updating $\mathcal{M}_n$ with $\{(\mathbf{x}_n, c_n)\}$;
    Sample PYP hyper-parameters;
  else
    $\mathcal{L}_n = \mathcal{L}_{n-1}$;
  end
until *some stopping criterion*;

4. Results

We evaluated the proposed method on five UCI datasets [2] widely used in benchmarking active learning methods. In addition, we also included five vision datasets, i.e. CASIA gait database[1], MNIST handwritten digits dataset[2], QMUL public surveillance dataset [13], Labeled Yahoo! News face database[3], and Scene Understanding database[4]. These datasets are given abbreviations gait, digits, qmul, yahooface, and sun respectively. Details of each datasets are shown in Table 1. Note that all datasets contains naturally unbalanced class proportions, except gait and digits, which were subsampled to contain geometric class proportions as in [9, 6]. We applied similar preprocessing steps described in [9] on the UCI, gait, and digits datasets,. For qmul we use the same activity representation as in [13]. For yahooface and sun datasets we performed PCA to reduce the original high-dimensional descriptors[5] to 40 dimensions[6].

Data	N	d	N_c	$S\%$	$L\%$
pageblocks	5473	10	5	0.49	89.28
shuttle	20000	9	7	0.02	78.40
thyroid	7200	21	3	2.47	92.47
covertype	5000	10	7	3.56	24.36
kdd	33650	113	15	0.04	51.46
gait	2353	25	9	2.92	48.66
digits	13184	25	10	0.10	50.05
qmul	1800	8	6	0.22	59.76
yahooface	10390	40	330	0.10	11.23
sun	108754	40	397	0.09	2.17

Table 1. Dataset properties: N = number of instances, d = dimension of data; N_c = number of classes, $S\%$ and $L\%$ = proportions of smallest and largest classes.

In the following experiments, unlike [13, 8], we do not

[1] http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp
[2] http://yann.lecun.com/exdb/mnist/
[3] http://lear.inrialpes.fr/data/
[4] http://groups.csail.mit.edu/vision/SUN/
[5] For the sun dataset, we used only the HOG2x2 descriptor.
[6] All processed data used in our experiments are available at http://www.eecs.qmul.ac.uk/~ccloy/

return any discarded instance back to the unlabelled data stream as we assume that a learner typically does not reuse past instances in a strict stream-based learning environment. The active learning may stop under two circumstances: (1) when a learner discards 100 instances consecutively, or (2) a given query budget is exhausted. A query budget of 150 was assigned to all datasets, except those with a large number of classes, i.e. the sun and yahooface databases, which were allocated a budget of 1000. For performance comparison, we use the *average class accuracy* [9], in which final accuracy is obtained by averaging classification accuracy for each class. As such, the final average accuracy is fairly penalised when there is mis-classifications in small classes. All results are averaged over 25 runs with two-fold cross-validation.

4.1. Pitman-Yor vs Dirichlet Processes

We first focus our attention on the performance comparison between the use of PYP and DP in stream-based active learning. In particular, we compare our proposed Alg. 1 and a variant that replaces PYP in step 4 with DP. We labelled them as qbc-pyp and qbc-dp respectively.

Note that a method may stop learning at anytime without finishing the given query budget, i.e. when they discard 100 instances consecutively. We observed close discovery and classification performance between the both qbc-dp and qbc-pyp when they terminated. Nonetheless, the qbc-pyp tends to use fewer budgets in most cases. To highlight the performance difference, we compared the ratio of average class accuracy to the number of instance queried at the point when the learning stopped. Note that a high ratio value can only be achieved by an effective learner that returns high accuracy by just observing a handful of instances. A table comparing the ratios across different datasets are presented in Table 2. Random sampling (rand), the 'passive' learning equivalent, was used as baseline.

Accuracy ratio	Data	pageblocks	shuttle	thyroid	covertype	kdd
	rand	0.3963	0.2846	0.3641	0.3837	0.2793
	qbc-dp	0.5902	**0.5042**	**0.3757**	0.3878	0.3098
	qbc-pyp	**0.5963**	0.4690	0.3695	**0.3897**	**0.3156**
	Data	gait	digits	qmul	yahooface	sun
	rand	0.4402	0.3263	0.2742	0.0187	0.00208
	qbc-dp	0.4434	0.3298	0.2756	0.0190	0.00207
	qbc-pyp	**0.4460**	**0.3351**	**0.2786**	**0.0191**	**0.00214**

Table 2. Comparing the performance of using the Pitman-Yor and Dirichlet process in Alg. 1. The performance metric is the ratio of average class accuracy (in percentage) to number of instance queried. A higher ratio indicates a better performance.

From Table 2, it is evident that both the qbc-dp and qbc-pyp consistently outperformed random sampling. Our proposed qbc-pyp outperformed the alternative qbc-dp in 8 out of 10 datasets. To provide insight into the performance difference between PYP and DP, we chose the yahooface database, and plotted the number of instances of individual classes vs. the associated rank on a log-log scale in Fig. 6. The yahooface database has a highly imbalanced class distribution (see Table 1) following a power-law scaling (see Fig. 1), where the 50 most common classes account for over 50% of the observed face images. From Fig. 6 and the results presented in Table 2, one can see that the PYP captures the power law statistics better than the DP, suggesting the PYP prior is a more sensible choice for problems with high bias in class distribution.

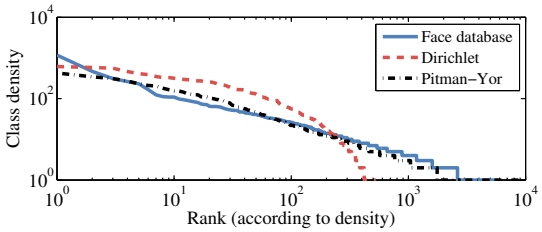

Figure 6. Using the *Labeled Yahoo! News* face database, we plot the empirical class density (number of instances in each class) vs. density rank across all classes in a log-log scale. We also plot the number of unique classes drawn from a Pitman-Yor processes (PYP) and Dirichlet processes (DP). A pure power law relationship would be a perfect straight line on a log-log scale.

4.2. Comparison to State-of-the-Art Methods

We compared the proposed method (Alg. 1) against the following state-of-the-art stream-based methods:

- *lowlik* - low-likelihood criterion specialised for quick unknown class discovery. The criterion is similar to that proposed in [16] but is modified for stream-based learning in [13].
- *qbc* - Query-by-Committee [1], state-of-the-art stream-based algorithm to search for ambiguous instances.
- *multi-kldiv* - a multi-criteria active learning method [13] that combines qbc and lowlik for joint exploration-exploitation. Different criteria are balanced through measuring individual impact on the model change over time.

Note that the limited number of stream-based active learning studies restricted the methods we could compare against. Alternative methods such as [9, 6, 29, 10, 7] are not suitable due to their pool-based nature. In addition, some methods are designed for active detection but not active classification [29], whilst other approaches are computationally intractable on our model [3]. We did not compare with [8] because it is conceptually similar to its generative version 'qbc'.

We first compared the class discovery performance of various approaches. The results are presented in Table 3. It is not surprising that qbc obtained poor discovery results in most datasets (e.g. pageblocks, shuttle, and kdd) since it is not designed for the discovery task. Minor improvements

Data	N_c	lowlik	qbc	multi-kldiv	qbc-pyp
pageblocks	5	4.72	3.42	3.30	**5.00**
shuttle	7	3.72	3.28	3.44	**5.00**
thyroid	3	2.92	2.68	2.56	**3.00**
covertype	7	7.00	7.00	7.00	7.00
kdd	15	**9.76**	3.32	4.12	8.71
gait	9	8.98	8.98	9.00	9.00
digits	10	8.84	8.84	8.52	7.76
qmul	6	5.10	**5.38**	5.10	5.10
yahooface	330	279.22	274.88	19.94	**279.24**
sun	397	325.44	318.52	325.98	**326.22**

Table 3. Number of classes discovered by different methods. N_c is the original number of classes. Best results are highlighted.

on some datasets were observed by using the multi-criteria method multi-kldiv. Nonetheless, due to the difficulty on tuning the criteria-weighting parameters, its heuristic balancing scheme failed to apply the right criterion at the right learning stage, therefore giving poor results on datasets such as pageblocks, shuttle, thyroid, and yahooface. Overall, our method qbc-pyp outperformed qbc and multi-kldiv, with comparable performance to lowlik that is specialised for quick unknown class discovery.

Data	supervised	lowlik	qbc	multi-kldiv	qbc-pyp
pageblocks	70.00	63.23	45.79	44.71	**71.72**
shuttle	62.17	45.46	38.87	42.39	**55.49**
thyroid	73.96	54.62	50.07	47.98	**55.42**
covertype	73.60	57.13	**58.68**	58.27	58.45
kdd	73.09	**51.21**	19.42	23.80	47.35
gait	75.13	66.03	64.81	**67.79**	66.90
digits	83.48	48.94	**58.04**	55.11	50.27
qmul	52.21	**42.31**	41.96	41.67	41.78
yahooface	35.84	18.65	18.58	1.64	**19.06**
sun	8.25	2.08	2.08	2.09	**2.14**

Table 4. Average class accuracy achieved using different methods. Best results are highlighted.

Next, we compared the average classification accuracy of each approach. Fully supervised learning is also included for reference. Despite the good results on some datasets, performance of qbc and multi-kldiv was generally unstable, and poor accuracy was observed on those datasets where they performed poorly in class discovery. The strong exploration capability of lowlik compensated its missing exploitation feature, it thus surprisingly performed rather well on several datasets. Nonetheless, the proposed qbc-pyp outperformed lowlik on 8 out of the 10 datasets.

To illustrate the significance of the results, Fig. 7 plots the percentage change in performance of qbc-pyp relative to lowlik [16], qbc [1], and multi-kldiv [13]. Clearly while qbc-pyp rarely performs noticeably worse than the others, the potential improvement is significant. Overall the results suggest that qbc-pyp an all-round and more stable and effective stream-based active learning framework compared to its state-of-the-art alternatives.

Finally, we give more insight into how the proposed approach balances the exploration and exploitation goals. Plots in Fig. 8 demonstrate how the level of interest in discovering unknown classes drops as the algorithm makes

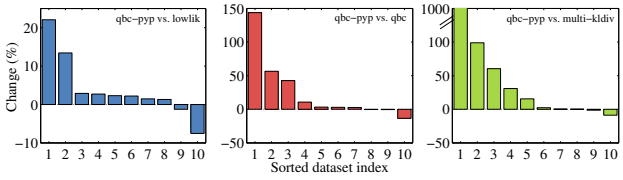

Figure 7. Percentage improvement of average classification accuracy for qbc-pyp over prior models [16, 1, 13] for all datasets.

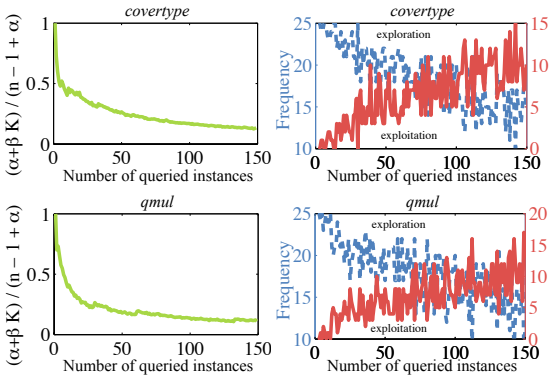

Figure 8. covertype and qmul datasets – (*Left figures*): Pitman-Yor prior on unknown class reflects the weight assigned to finding new classes. (*Right figures*): Automatic switching between exploration and exploitation.

more queries. In particular, the left plots show that the PYP prior of new class probability started at a high level at the initial stage of learning and gradually dropped as more data were observed. For illustration purposes, we decoupled the proposed criterion Eq. 7 into two parts, and calculated the frequency of $\{h_1(\mathbf{x}_n) \neq h_2(\mathbf{x}_n)\}$ and $\{h_i(\mathbf{x}_n) = K+1 \mid \forall i\}$ being invoked in 25 random runs. As can be seen from both the right plots, exploration dominated at the beginning when there were numerous new classes or unknown regions to be discovered; the learner eventually switched its tendency to searching for ambiguous instances when the exploratory learning was no longer fruitful. Importantly, this adaptation occurs in a principled and data-driven way (see Sec. 3.4).

4.3. Computational Time

We measured the run time of our MATLAB implementation on a dual-core 3.3 GHz machine. On the yahooface dataset, our stream-based learning scheme of $\mathrm{O}(K)$ complexity required on average of 0.28 seconds to make a query decision, and an additional 0.10 seconds to train the DPMM incrementally and sample the PYP hyperparameters. In contrast, an pool-based scheme similar to that proposed in [19] has a complexity of $\mathrm{O}(n^2K)$, requiring approximately 8 minutes to make a decision, which is clearly infeasible given large databases.

5. Conclusion

We have presented a novel stream-based active learning framework for supervision-efficient learning in tasks without a pre-defined class-space – those requiring joint exploration-exploration. Our framework shows superior performance over three state-of-the-art stream-based methods in both unknown class discovery and classification. Importantly, the active learner automatically balances exploration and exploitation by learning the Pitman-Yor process hyperparameters of the Bayesian non-parametric model. Moreover, the use of the PYP prior improves performance on more challenging real-world datasets with power-law class distributions compared to recently studied DP-based priors [6]. Finally, via incremental learning our approach is well suited for real-time applications.

Our proposed qbc-pyp framework is well suited to address numerous increasingly common and important contemporary tasks requiring on-the-fly interactive learning from unbounded streams and very large scale data with evolving class composition: video surveillance, network security, social networking mining, etc. Future work will explore potential extension such as active learning from multiple noisy oracles [30] and combining active learning with semi-supervised learning.

References

[1] S. Argamon-Engelson and I. Dagan. Committee-based sample selection for probabilistic classifiers. *JAIR*, 11:335–360, 1999. 2, 3, 5, 6, 7

[2] A. Asuncion and D. Newman. UCI machine learning repository, 2007. 5

[3] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, pages 49–56, 2009. 6

[4] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. In *ML*, pages 133–168, 1997. 2

[5] R. Gomes, M. Welling, and P. Perona. Incremental learning of nonparametric Bayesian mixture models. In *CVPR*, pages 1–8, 2008. 5

[6] T. Haines and T. Xiang. Active learning using Dirichlet processes for rare class discovery. In *BMVC*, 2011. 2, 4, 5, 6, 8

[7] J. He and J. Carbonell. Nearest-neighbor-based active learning for rare category detection. In *NIPS*, 2007. 2, 6

[8] S.-S. Ho and H. Wechsler. Query by transduction. *TPAMI*, 30(9):1557–1571, 2008. 1, 2, 4, 5, 6

[9] T. M. Hospedales, S. Gong, and T. Xiang. Finding rare classes: Active learning with generative and discriminative models. *TKDE*, 2011. 2, 5, 6

[10] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Gaussian processes for object categorization. *IJCV*, 88(2):169–188, 2009. 2, 6

[11] A. Kovashka, S. Vijayanarasimhan, and K. Grauman. Actively selecting annotations among objects and attributes. In *ICCV*, 2011. 2

[12] K. Kurihara, M. Welling, and N. Vlassis. Accelerated variational dirichlet process mixtures. In *NIPS*, 2006. 5

[13] C. C. Loy, T. Xiang, and S. Gong. Stream-based active unusual event detection. In *ACCV*, 2010. 1, 2, 3, 4, 5, 6, 7

[14] A. K. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *ICML*, pages 350–358, 1998. 2, 3

[15] K. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, UBC, 2007. 3

[16] D. Pelleg and A. Moore. Active learning for anomaly and rare-category detection. In *NIPS*, pages 1073–1080, 2004. 2, 6, 7

[17] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997. 1

[18] S. J. Prince. *Computer Vision: Models, Learning and Inference*. Cambridge University Press, 2012. 2

[19] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *ICML*, pages 441–448, 2001. 2, 7

[20] R. Salakhutdinov, A. Torralba, and J. Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR*, 2011. 1

[21] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010. 2

[22] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT*, pages 287–294, 1992. 1, 2, 5

[23] A. Shyr, T. Darrell, M. Jordan, and R. Urtasun. Supervised hierarchical Pitman-Yor process for natural scene segmentation. In *CVPR*, pages 2281–2288, 2011. 2, 4

[24] J. W. Stokes, J. C. Platt, J. Kravis, and M. Shilman. ALADIN: Active learning of anomalies to detect intrusions. Technical report, Microsoft Research, 2008. 2

[25] Y. W. Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical report, NUS, 2006. 4

[26] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *ICCL*, 2006. 2, 4

[27] Y. W. Teh. Dirichlet processes. Encyclopedia of Machine Learning, 2010. 2

[28] M. Tubaishat and S. Madria. Sensor networks: an overview. *Potentials, IEEE*, 22(2):20–23, 2003. 1

[29] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. In *CVPR*, 2011. 1, 2, 6

[30] P. Welinder and P. Perona. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *CVPR*, pages 25–32, 2010. 8

[31] F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. Teh. The sequence memoizer. *Communications of the ACM*, 54(2):91–98, 2011. 4

[32] G. Zipf. The psychobiology of language. 1935. 1