Finding Rare Classes: Adapting Generative and Discriminative Models in Active Learning

Timothy Hospedales, Shaogang Gong and Tao Xiang

Queen Mary University, London, UK, N4 2YD
{tmh,sgg,txiang}@eecs.qmul.ac.uk

Abstract. Discovering rare categories and classifying new instances of them is an important data mining issue in many fields, but fully supervised learning of a rare class classifier is prohibitively costly. There has therefore been increasing interest both in active discovery: to identify new classes quickly, and active learning: to train classifiers with minimal supervision. Very few studies have attempted to jointly solve these two inter-related tasks which occur together in practice. Optimizing both rare class discovery and classification simultaneously with active learning is challenging because discovery and classification have conflicting requirements in query criteria. In this paper we address these issues with two contributions: a unified active learning model to jointly discover new categories and learn to classify them; and a classifier combination algorithm that switches generative and discriminative classifiers as learning progresses. Extensive evaluation on several standard datasets demonstrates the superiority of our approach over existing methods.

1 Introduction

Many real life problems are characterized by data distributed between vast yet uninteresting background classes, and small rare classes of interesting instances which should be identified. In astronomy, the vast majority of sky survey image content is due to well understood phenomena, and only 0.001% of data is of interest for astronomers to study [12]. In financial transaction monitoring, most are ordinary but a few unusual ones indicate fraud and regulators would like to find future instances. Computer network intrusion detection exhibits vast amounts of normal user traffic, and a very few examples of malicious attacks [16]. Finally, in computer vision based security surveillance of public spaces, observed activities are almost always people going about everyday behaviours, but very rarely may be a dangerous or malicious activity of interest [19]. All of these classification problems share two interesting properties: highly unbalanced frequencies – the vast majority of data occurs in one or more background classes, while the instances of interest for classification are much rarer; and unbalanced prior supervision – the majority classes are typically known *a priori*, while the rare classes are not. Classifying rare event instances rather than merely detecting *any* rare event is crucial because different classes may warrant different responses, for example due to different severity levels. In order to discover and learn to

classify the interesting rare classes, exhaustive labeling of a large dataset would be required to ensure sufficient rare class coverage. However this is prohibitively expensive when generating each label requires significant time of a human expert. Active learning strategies might be used to discover or train a classifier with minimal label cost, but this is complicated by the dependence of classifier learning on discovery: one needs examples of each class to train a classifier.

The problem of joint discovery and classification has received little attention despite its importance and broad relevance. The only existing attempt to address this is based on simply applying schemes for discovery and classifier learning sequentially or in fixed iteration [16]. Methods which treat discovery and classification independently perform poorly due to making inefficient use of data, (e.g., spending time on classifier learning is useless if the right classes have not been discovered and vice-versa). Achieving the optimal balance is critical, but non-trivial given the conflict between discovery and learning criteria. To address this, we build a generative-discriminative model pair [11,4] for computing discovery and learning query criteria, and adaptively balance their use based on joint discovery and classification performance. Depending on the actual supervision cost and sparsity of rare class examples, the quantity of labeled data varies. Given the nature of data dependence in generative and discriminative models [11], the ideal classifier also varies. As a second contribution, we therefore address robustness to label quantity and introduce a classifier switching algorithm to optimize performance as data is accumulated. The result is a framework which significantly and consistently outperforms existing methods at the important task of discovery and classification of rare classes.

Related Work A common unsupervised approach to rare class detection is outlier detection: building an unconditional model of the data and flagging unlikely instances. This has a few serious limitations: it does not classify; it fails with non-separable data, where interesting classes are embedded in the majority distribution; and it does not exploit any supervision about flagged outliers, limiting its accuracy – especially in distinguishing rare classes from noise.

Iterative active learning approaches are often used to learn a classifier with minimal supervision [14]. Much of the active learning literature is concerned with the relative merits of different query criteria. For example, querying points that: are most uncertain [14]; reduce the version space [17]; or reduce direct approximations of the generalization error [13]. Different criteria may be suited to different datasets, e.g, uncertainty criteria are good to refine decision boundaries, but can be fatal if the classes are non-separable (the most uncertain points may be hopeless) or highly multi-modal. This has led to attempts to select dataset specific criteria online [2]. All these approaches rely on classifiers, and do not generally apply to scenarios in which the target classes are themselves unknown.

Recently, active learning has been applied to *discovering* rare classes using e.g., likelihood [12] or gradient [9] criteria. Solving discovery and classification problems together with active learning is challenging because for a single dataset, good discovery and classification criteria are often completely different. Consider the toy scenarios in Figure 1. Here the color indicates the true class, and the

symbol indicates the estimated class based on two initial labeled points (large symbols). The black line indicates the initial decision boundary. In Figure 1(a) all classes are known but the decision boundary needs refining. Likelihood sampling (most unlikely point under the learned model) inefficiently builds a model of the whole space (choosing first the points labeled L), while uncertainty sampling selects points closest to the boundary (U symbols), leading to efficient refinement. In Figure 1(b) only two classes are known. Uncertainty inefficiently queries around the known decision boundary (choosing first the points U) without discovering the new classes above. In contrast, these are the first places queried by likelihood sampling (L symbols). Evidently, single-criterion approaches are insufficient. Moreover, multiple criteria may be necessary for a single dataset at different stages of learning, e.g., likelihood to detect new classes and uncertainty to learn to classify them. A simple but inefficient approach [16] is to simply iterate over criteria in fixed proportion. In contrast, our innovation is to adapt criteria online so as to select the right strategy at each stage of learning, which can dramatically increase efficiency. Typically, "exploration" is automatically preferred while there are easily discoverable classes, and "exploitation" to refine decision boundaries when most classes have been discovered. This ultimately results in better rare class detection performance than single objective, or non-adaptive methods [16].

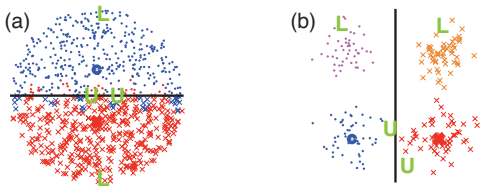

Fig. 1. Sample Problems.

Finally, there is the issue of what base classifier to use in the active learning algorithm of choice. One can categorize classifiers into two broad categories: generative and discriminative. Discriminative models directly learn $p(y|x)$ for class y and data x. Generative models learn $p(x, y)$ and compute $p(y|x)$ via Bayes rule. The importance of this for active learning is that for a given generative-discriminative *pair* (in the sense of equivalent parametric form – such as naive Bayes & logistic regression), generative classifiers typically perform better with few training examples, while discriminative models are better asymptotically [11]. *The ideal classifier is therefore likely to be completely different early and late in the active learning process.* An automatic way to select the right classifier online as more labels are obtained is therefore key. Existing active learning work focuses on single generative [13] or discriminative [17] classifiers. We introduce a novel algorithm to switch classifiers online as the active learning process progresses in order to get the best of both worlds.

2 Adaptive Active Learning

2.1 Active Learning

In this paper we deal with pool-based uncertainty sampling and likelihood sampling because of their computational efficiency and clearly complementary nature. Our method can nevertheless be easily generalized to other criteria. We consider a classification problem starting with many unlabeled instances $\mathcal{U} = (x_1, .., x_n)$ and a small set of labeled instances $\mathcal{L} = ((x_1, y_1), .., (x_m, y_m))$. $\mathcal{L}$ does not include the full set of possible labels $\mathcal{Y}$ in advance. We wish to learn the posterior conditional distribution $p(y|x)$ so as to accurately classify the data in $\mathcal{U}$. Active learning proceeds by iteratively: i) training a classifier $\mathcal{C}$ on $\mathcal{L}$; ii) using query function $\mathcal{Q}(\mathcal{C}, \mathcal{L}, \mathcal{U}) \rightarrow i^*$ to select unlabeled instances i^* to be labeled and iii) removing x_{i^*} from $\mathcal{U}$ and adding (x_{i^*}, y_{i^*}) to $\mathcal{L}$.

Query Criteria Perhaps the most commonly applied query criteria are uncertainty sampling and variants [14]. The intuition is that if the current classification of a point is highly uncertain, it should be informative to label. Uncertainty is typically quantified by posterior entropy, which for binary classification reduces to selecting the point whose posterior is closest to $p(y|x) = 0.5$. The posterior $p(y|x)$ of every point in $\mathcal{U}$ is evaluated and the uncertain points queried,

$$p_u(i) \propto \exp\left(\beta \sum_{y_i} p(y_i|x_i) \log p(y_i|x_i)\right). \tag{1}$$

Rather than selecting a single maxima, we exploit a normalized *degree of preference* $p_u(i)$ for every point i can be expressed by putting the entropy into a Gibbs function (1). For non-probabilistic SVM classifiers, an approximation to $p(y|x)$ can be derived from the distance to the margin from each point [14].

A complementary query criteria is that of low likelihood $p(x|y)$. Such points are badly explained by the current model, and should therefore be informative to label [12]. This may involve marginalizing over the class or selecting the maximum likelihood label,

$$p_l(i) \propto \exp\left(-\beta \max_{y_i} p(x_i|y_i)\right). \tag{2}$$

The uncertainty measure in (1) is in spirit *discriminative* (in focusing on decision boundaries), although $p(y|x)$ can obviously be realized by a generative classifier. In contrast, the likelihood measure in (2) is intrinsically *generative*, in that it requires a density model of each class y, rather than just the decision boundary. The uncertainty measure is generally unsuitable for finding new classes, as it focuses on known decision boundaries, and the likelihood measure is good at finding new classes, while being poorer at refining decision boundaries between known classes (Figure 1). Note that the likelihood measure can still be useful to improve known-class classification if the classes are multi-modal – it will explore different modes. Our adaptation method will allow it to be used in both ways. Next, we discuss specific parametric forms for our models.

2.2 Generative-Discriminative Model Pairs

We use a Gaussian mixture model (GMM) for the generative model and a support vector machine (SVM) for the discriminative model. These were chosen because they may both be incrementally trained (for active learning efficiency), and they are a complementary generative-discriminative *pair* in that (assuming a radial basis SVM kernel) they have equivalent classes of decision boundaries [4], but are optimized with very different criteria during learning.

Incremental GMM Estimation For online GMM learning, we use the incremental agglomerative algorithm from [15]. To summarize the procedure, for the first $n = 1..N$ training points observed with the same label y, $\{\mathbf{x}_n, y\}_n^N$, we incrementally build a model $p(\mathbf{x}|y)$ for y using kernel density estimation with Gaussian kernels $\mathcal{N}(\mathbf{x}_n, \Sigma)$ and weight $\omega_n = \frac{1}{n}$. d is the dimension of the data $\mathbf{x}$.

$$p(\mathbf{x}|y) = \frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}} \sum_{n=1}^{N} \omega_n \exp -\frac{1}{2}\left((\mathbf{x}-\mathbf{x}_n)^T \Sigma^{-1}(\mathbf{x}-\mathbf{x}_n)\right). \quad (3)$$

To bound the complexity, after some maximal number of Gaussians N_{max} is reached, merge two existing Gaussians i and j by moment matching [7].

$$\omega_{(i+j)} = \omega_i + \omega_j, \qquad \mu_{(i+j)} = \frac{\omega_i}{\omega_{(i+j)}}\mu_i + \frac{\omega_j}{\omega_{(i+j)}}\mu_j, \quad (4)$$

$$\begin{aligned}\Sigma_{(i+j)} = &\frac{\omega_i}{\omega_{(i+j)}}\left(\Sigma_i + (\mu_i - \mu_{(i+j)})(\mu_i - \mu_{(i+j)})^T\right) \\ &+\frac{\omega_j}{\omega_{(i+j)}}\left(\Sigma_j + (\mu_j - \mu_{(i+j)})(\mu_j - \mu_{(i+j)})^T\right). \quad (5)\end{aligned}$$

The components to merge are chosen by the selecting the pair of Gaussian kernels (G_i, G_j) whose replacement $G_{(i+j)}$ is most similar, in terms of the Kullback-Leibler divergence. Specifically, we minimize the cost C_{ij},

$$C_{ij} = \omega_i \mathcal{KL}(G_i||G_{(i+j)}) + \omega_j \mathcal{KL}(G_j||G_{(i+j)}). \quad (6)$$

Importantly for iterative active learning online, merging Gaussians and updating the cost matrix requires constant $\mathcal{O}(N_{max})$ computation every iteration once the initial cost matrix has been built. In contrast, learning a GMM with latent variables requires multiple expensive $O(n)$ expectation-maximization iterations [12]. The initial covariance parameter Σ is assumed uniform diagonal $\Sigma = \mathbf{I}\sigma^2$, and is estimated *a priori* by leave-one-out cross validation on the (large) unlabeled dataset $\mathcal{U}$:

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmax}} \left(\prod_{n \in \mathcal{U}} \sigma^{-\frac{d}{2}} \sum_{x \neq x_n} \exp -\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{x}_n)^2 \right). \quad (7)$$

Given the learned models $p(\mathbf{x}|y)$, we can classify $\hat{y} \leftarrow f_{gmm}(\mathbf{x})$, where

$$f_{gmm}(\mathbf{x}) = \underset{y}{\text{argmax}} p(y|\mathbf{x}), \qquad p(y|\mathbf{x}) \propto \sum_i w_i \mathcal{N}(\mathbf{x}; \mu_{i,y}, \Sigma_{i,y}) p(y). \quad (8)$$

SVM We use a standard SVM approach with RBF kernels, treating multi-class classification as a set of 1-vs-1 decisions, for which the decision rule [4] is given (by an equivalent form to (8)) as

$$f_{svm}(\mathbf{x}) = \underset{y}{\text{argmax}} \left(\sum_{\mathbf{v}_i \in SV_y} \alpha_{ki} \mathcal{N}(\mathbf{x}; \mathbf{v}_i) + \alpha_{k0} \right), \quad (9)$$

and $p(y|\mathbf{x})$ can be computed via an optimization based on the binary posterior estimates [18].

2.3 Combining Active Query Criteria

Given the generative GMM and discriminative SVM models defined in Section 2.2, and their respective likelihood and uncertainty query criteria defined in Section 2.1, our first concern is how to adaptively combine the query criteria online for discovery and classification. Our algorithm involves probabilistically *selecting* a query criteria Q_k according to some weights $\mathbf{w}$ ($k \sim \text{Multi}(\mathbf{w})$) and then sampling the query point from the distribution $i^* \sim p_k(i)$ ((1) or (2)).[1] The weights $\mathbf{w}$ will be adapted based on the discovery and classification performance ϕ of our active learner at each iteration. In an active learning context, [2] shows that because labels are few and biased, cross-validation is a poor way to assess classification performance, and suggest the unsupervised measure of binary *classification entropy* (CE) on the unlabeled set $\mathcal{U}$ instead. This is especially the case in the rare class context where there is often only one example of a given class, so cross-validation is not well defined. To overcome this problem, we generalize CE to multi-class entropy (MCE) of the classifier $f(\mathbf{x})$ and take it as our indication of classification performance,

$$H = -\sum_{y=1}^{n_y} \frac{\sum_i \mathrm{I}(f(\mathbf{x}_i) = y)}{|\mathcal{U}|} \log_{n_y} \frac{\sum_i \mathrm{I}(f(\mathbf{x}_i) = y)}{|\mathcal{U}|}. \quad (10)$$

Here I is the indicator function that returns 1 if its argument is true, and n_y is the number of classes observed so far. Importantly, we explicitly reward the discovery of new classes to jointly optimize classification and discovery. We define overall active learning performance $\phi_t(i)$ upon querying point i at time t as,

[1] We choose method because each criterion has very different "reasons" for its preference. An alternative is querying a product or mean [2] of the criteria. That risks querying a merely moderately unlikely and uncertain point – neither outlying nor on a decision boundary – which is useless for either classification or discovery.

$$\phi_t(i) = \alpha \mathrm{I}(y_i \notin \mathcal{L}) + (1-\alpha)\left((e^{H_t} - e^{H_{t-1}}) - (1-e)\right)/(2e-2). \quad (11)$$

The first right hand term above rewards discovery of a new class, and the second term rewards an increase in MCE (as an estimate of classification accuracy) after labeling point i at time t. The constants $(1-e)$ and $(2e-2)$ ensure the second term lies between 0 and 1. The parameter α is the mixing prior for discovery vs. classification. Given this performance measure, we define an update for the future weight w_{t+1} of each active criterion k,

$$w_{t+1,k}(q) \propto \lambda w_{t,k} + (1-\lambda)\phi_t(i)\frac{p_k(i)}{p(i)} + \epsilon. \quad (12)$$

Here we define an exponential decay (first term) of the weight in favor of (second term) the current performance ϕ weighted by how strongly criteria k recommended the chosen point i, compared to the joint recommendation $p(i) = \sum_k p_k(i)$. λ is the forgetting factor. The third term encourages exploration by diffusing the weights so every criterion is tried occasionally. In summary, this approach adaptively selects more frequently those criteria that have been successful at discovering new classes and/or increasing MCE, thereby optimizing both discovery and classification accuracy.

2.4 Adaptive Selection of Classifiers

As discussed in Section 1, although we broadly expect the generative GMM classifier to have better initial performance, and the discriminative SVM classifier to have better asymptotic performance, the ideal classifier will vary with dataset and active learning iteration. The remaining question is how to combine these classifiers [10] online for best performance given any specific supervision budget. Cross-validation to determine reliability is infeasible because of lack of data; however we can again resort to the MCE over the training set $\mathcal{U}$ (10). In our experience, MCE is indeed indicative of generalization performance, but relatively crudely and non-linearly so. This makes approaches based on MCE weighted posterior fusion unreliable. We therefore choose a simpler but more reliable approach which *switches* the final classifier at the end of each iteration to the one with higher MCE, aiming to perform as well as the better classifier for any label budget. Additionally, the process of multi-class posterior estimation for SVMs [18] requires cross-validation and is inaccurate with limited data. To compute the uncertainty criterion (1) at each iteration, we therefore use posterior of the classifier determined to be more reliable by MCE. This ensures that uncertainty sampling is as accurate as possible in both low and high data contexts.

Summary Algorithm 1 summarizes our approach. There are four parameters: Gibbs parameter β, discovery vs. classification prior α, forgetting rate λ and exploring rate ϵ. None of these were tuned; we set them all crudely to intuitive values for all experiments, $\beta = 100$, $\alpha = 0.5$, $\lambda = 0.9$ and $\epsilon = 0.01$. The GMM and SVM classifiers both have regularization hyperparameters N_{max} and (C, γ). These were not optimized, but set at standard values $N_{max} = 32$, $C = 1, \gamma = 1/d$.

Algorithm 1 Integrated Active Learning for Discovery and Classification

Active Learning

Input: Labeled $\mathcal{L}$ and unlabeled $\mathcal{U}$ data. Classifiers $\mathcal{C}$, query criteria $\mathcal{Q}_k$, weights $\mathbf{w}$.

1. Build unconditional GMM from $\mathcal{L} \cup \mathcal{U}$ (3)-(5)
2. Estimate σ by cross-validation (7)
3. Train initial GMM f_{gmm} and SVM f_{svm} classifiers on $\mathcal{L}$ using σ

Repeat as training budget allows:

1. Compute query criteria $p_u(i)$ (1) and $p_l(i)$ (2)
2. Sample query criteria to use $k \sim \text{Multi}(\mathbf{w})$
3. Query point $i^* \sim p_k(i)$, add (x_{i^*}, y_{i^*}) to $\mathcal{L}$
4. Update classifiers f_{gmm} and f_{svm} with point i^* (8) and (9)
5. Compute multi-class classification entropies H^{gmm} and H^{svm} (10)
6. Update query criteria weights $\mathbf{w}$ (11) and (12)
7. If $H^{gmm} > H^{svm}$: select classifier $f_{gmm}(\mathbf{x})$, Else: select $f_{svm}(\mathbf{x})$

Testing

Input: Testing samples $\mathcal{U}^*$, selected classifier c.

1. Classify $x \in \mathcal{U}^*$ with $f_c(\mathbf{x})$ ((8) or (9))

3 Experiments

Evaluation Procedure We tested our method on 7 rare class datasets from the UCI repository [1] and on the CASIA gait dataset [20], for which we addressed the image viewpoint recognition problem. We unbalanced the CASIA dataset by sampling training classes in geometric proportion. In each case we labeled one point from the largest class and the goal was to discover and learn to classify the remaining classes. Table 1 summarizes the properties of each dataset. Performance was evaluated at each iteration by: i) the number of distinct classes discovered and ii) the average classification accuracy over all classes. This accuracy measure weights the ability to classify rare classes equally with the majority class despite the fewer rare class points. Moreover, it means that undiscovered rare classes automatically penalize accuracy. Accuracy was evaluated by 2-fold cross-validation, averaged over 25 runs from random initial conditions.

Comparative Evaluations We compared the following methods: **S**/**R**: A baseline SVM classifier making random queries. **G**/**G**: GMM classification with GMM likelihood criterion (2). **S**/**S**: SVM classifier with SVM uncertainty criterion (1). **S**/**GSmix:** SVM classifier alternating GMM likelihood and SVM uncertainty queries (corresponding to [16]). **S**/**GSonline**: SVM classifier fusing GMM likelihood & SVM uncertainty criteria by the method in [2]. **S**/**GSadapt**: SVM classification with our adaptive fusion of GMM likelihood & SVM uncertainty criteria (10)-(12). **GSsw**/**GSadapt**: Our full model including online switching of GMM and SVM classifiers, as detailed in Algorithm 1.

Shuttle (Figure 2(a)). Our methods S/GSadapt (cyan) and GSsw/GSadapt (red), exploit likelihood sampling early for fast discovery, and hence early classi-

fication accuracy. (We also outperform the gradient and EM based active discovery models in [9] and [12].) Our adaptive models switch to uncertainty sampling later on, and hence achieve higher asymptotic accuracy than the pure likelihood based G/G method. Figure 2(c) illustrates this process via the query criteria weighting (12) for a typical run. The likelihood criterion discovers a new class early, leading to higher weight (11) and rapid discovery of the remaining classes. After 50 iterations, with no new classes to discover, uncertainty criteria obtains greater reward (11) and dominates, efficiently refining classification performance.

Thyroid (Figure 2(b)). Our GSsw/GSadapt model (red) is the strongest overall classifier: it matches the initially superior performance of the G/G likelihood-based model (green), but later achieves the asymptotic performance of the SVM classifier based models. This is because of our classifier switching innovation (Section 2.4). Figure 2(d) illustrates switching via the average (training) classification entropy and (testing) classification accuracy of each of the classifiers composing GSsw/GSadapt. The GMM classifier entropy (black dots) is higher than the SVM entropy (blue dots) for the first 25 iterations. This is approximately the period over which the GMM classifier (black line) has better performance than the SVM classifier (blue line), so switching classifier on training entropy allows the classifier pair (green dashes) to always perform as well as the best classifier for each iteration.

Data	N	d	N_c	S%	L%
Ecoli	336	7	8	1.5%	42%
PageBlock	5473	10	5	.5%	90%
Glass	214	10	6	4%	36%
Covertype	10000	10	7	3.6%	25%
Shuttle	10000	9	7	.01%	78%
Thyroid	3772	22	3	2.5%	92%
KDD99	50000	23	15	.04%	51%
Gait view	2353	25	9	3%	49%

Table 1. Dataset properties. Number of items N, classes N_c, dimensions d. Smallest and largest class proportions S/L.

Data	G/G	S/GSmix	S/GSad	GSsw/GSad
EC	59	60	60	**62**
PB	53	57	58	**59**
GL	63	55	57	**64**
CT	41	39	43	**46**
SH	40	39	42	**43**
TH	50	55	56	**59**
KD	41	23	54	**59**
GA	38	31	49	**57**

Table 2. Classification performance summary in terms of area under classification curve.

Glass (Figure 2(e)). GSsw/GSadapt again performs best by switching to match the good initial performance of the GMM classifier and asymptotic performance of the SVM. Note the dramatic improvement over the SVM models in the first 50 iterations. **Pageblocks** (Figure 2(f)). The SVM-based models outperform G/G at most iterations. Our GSsw/GSadapt correctly selects the SVM classifier throughout. **Gait view** (Figure 2(g)). The majority class contains outliers, so likelihood criteria is unusually weak at discovery. Additionally for this data SVM performance is generally poor, especially in early iterations. GSsw/GSadapt adapts impressively to this dataset in two ways enabled by our contributions: exploiting uncertainty sampling criteria extensively and switching to predicting using the GMM classifier.

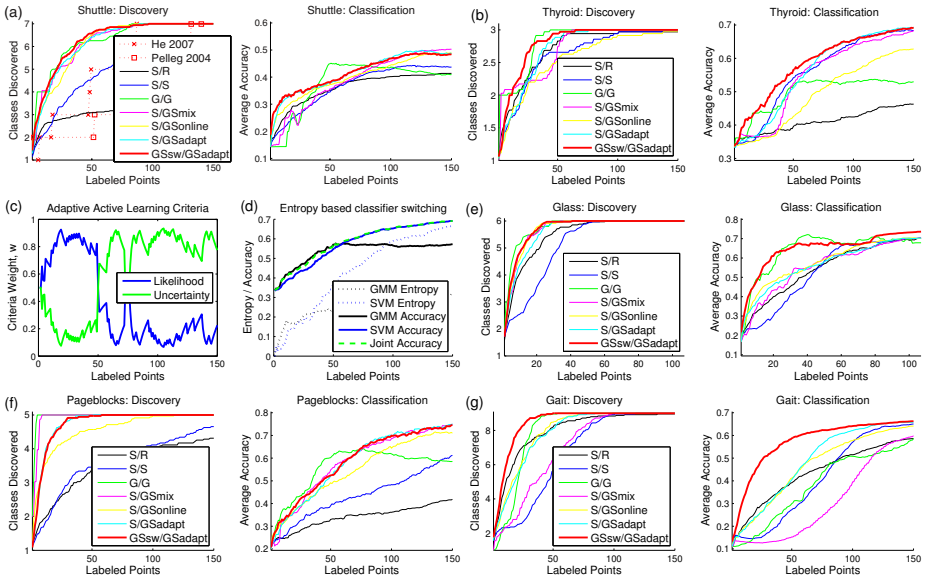

Fig. 2. (a) Shuttle and (b) Thyroid dataset performance. (c) Shuttle criteria adaptation, (d) Thyroid entropy based classifier switching. (e) Glass, (f) Pageblocks and (g) Gait view dataset performance.

In summary the G/G method (likelihood criterion) was usually the most efficient at discovering classes as expected. However, it was usually asymptotically weaker at classifying new instances. This is because generative model misspecification tends to cost more with increasing amounts of data [11]. S/S, (uncertainty criterion), was general poor at discovery (and hence classification). Alternating between likelihood and uncertainty sampling, S/GSmix (corresponding to [16]) did a fair job of both discovery and classification, but under-performed our adaptive models due to its inflexibility. S/GSonline (corresponding to [2]) was better than random or S/S, but was not the quickest learner. Our first model S/GSadapt, which solely adapted the multiple active query criteria, was competitive at discovery, but sometimes not the best at classification in early phases with very little data – due to exclusively using the discriminative SVM classifier. Finally, by exploiting generative-discriminative classifier switching, our complete GSsw/GSadapt model was generally the best classifier over all stages of learning. Table 2 quantitatively summarizes the performance of the most competitive models for all datasets in terms of area under the classification curve.

4 Conclusion

Summary We highlighted active classifier learning with *a priori* unknown rare classes as an under-studied but broadly relevant and important problem. To solve joint rare class discovery and classification, we proposed a new framework

to adapt both active query criteria and classifier. To adaptively switch generative and discriminative classifiers online we introduced MCE; and to adapt query criteria we exploited a joint reward signal of new class discovery and MCE. In adapting to each dataset and online as data is obtained, our model significantly outperformed contemporary alternatives on eight standard datasets. Our approach will be of great practical value for many problems.

Discussion A related area of research to our present work is that of learning from imbalanced data [8] which aims to learn classifiers for classes with imbalanced distributions, while avoiding the pitfall of simply classifying everything as the majority class. One strategy to achieve this is uncertainty based active learning [6], which works because the distribution around the class boundaries is less imbalanced than the whole dataset. Our task is also an imbalanced learning problem, but more general in that the rare classes must also be discovered. We succeed in learning from imbalanced distributions via our use of uncertainty sampling, so in that sense our method generalizes [6]. Although our approach lacks the theoretical bounds of the fusion method in [2], we find it more compelling for various reasons: it addresses a very practical and previously unaddressed problem of learning to discover new classes and find new instances of them by jointly optimizing searching for new classes and refining their decision boundaries. It adapts based on the current state of the learning process, i.e., early on, class finding via likelihood may be more appropriate, and later on boundary refinement via uncertainty. In contrast [2] solely optimises classification accuracy and is not directly applicable to discovery. [5] and [3] address the fusion of uncertainty and density (to avoid outliers) criteria for classifier learning (not discovery). [5] adapts between density weighted and unweighted uncertainty sampling based on their expected future error. This is different to our situation because there is no meaningful notion of future error when an unknown number of classes remain to be discovered. [3] samples from a weighted sum of density and uncertainty criteria. This is less powerful than our approach because it does not adapt online based on the performance of each criteria. Importantly, both [5] and [3] prefer high density points; while for rare class discovery we require the opposite – low likelihood. In comparison to other active rare class discovery work, our framework generalizes [12], (which exclusively uses generative models and likelihood criteria) to using more criteria and adapting between them. [9] focuses on a different active discovery intuition, using local gradient to discover non-separable rare classes. We derived an analogous query criterion based on GMM local gradient. It was generally weaker than likelihood-based discovery (and was hence adapted downward in our framework) for our datasets, so we do not report on it here. Unlike our work here, [5,12,9] all also rely on the very strong assumption that the user at least specifies the *number* of classes in advance. Finally, the only other work of which we are aware which addresses both discovery and classification is [16]. This uses a fixed classifier and non-adaptively iterates between discovery and uncertainty criteria (corresponding to our S/GSmix condition). In contrast, our results have shown that our switching classifier and adaptive query criteria provide compelling benefit for discovery and classification.

Future Work There are various interesting questions for future research including and how to create tighter coupling between the generative and discriminative components [4], and generalizing our ideas to stream based active learning, which is a more natural setting for some practical problems.

Acknowledgment. This research was funded by the EU FP7 project SAMURAI with grant no. 217899.

References

1. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/ml/`
2. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms. Journal of Machine Learning Research 5, 255–291 (2004)
3. Cebron, N., Berthold, M.R.: Active learning for object classification: from exploration to exploitation. Data Min. Knowl. Discov. 18(2), 283–299 (2009)
4. Deselaers, T., Heigold, G., Ney, H.: SVMs, gaussian mixtures, and their generative/discriminative fusion. In: ICPR (2008)
5. Donmez, P., Carbonell, J.G., Bennett, P.N.: Dual strategy active learning. In: ECML (2007)
6. Ertekin, S., Huang, J., Bottou, L., Giles, L.: Learning on the border: active learning in imbalanced data classification. In: CIKM (2007)
7. Goldberger, J., Roweis, S.: Hierarchical clustering of a mixture model. In: NIPS (2004)
8. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263 –1284 (2009)
9. He, J., Carbonell, J.: Nearest-neighbor-based active learning for rare category detection. In: NIPS (2007)
10. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
11. Ng, A., Jordan, M.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: NIPS (2001)
12. Pelleg, D., Moore, A.: Active learning for anomaly and rare-category detection. In: NIPS (2004)
13. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: ICML. pp. 441–448 (2001)
14. Settles, B.: Active learning literature survey. Tech. Rep. 1648, University of wisconsin–Madison (2009)
15. Sillito, R., Fisher, R.: Incremental one-class learning with bounded computational complexity. In: ICANN (2007)
16. Stokes, J.W., Platt, J.C., Kravis, J., Shilman, M.: Aladin: Active learning of anomalies to detect intrusions. Tech. Rep. 2008-24, MSR (2008)
17. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: ICML (2000)
18. Wu, T.F., Lin, C.J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. Journal of Machine Learning Research 5, 975–1005 (2004)
19. Xiang, T., Gong, S.: Video behavior profiling for anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(5), 893–908 (2008)
20. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: ICPR (2006)