Hierarchical Active Learning with Group Proportion Feedback

Zhipeng Luo and **Milos Hauskrecht**
Department of Computer Science, University of Pittsburgh, PA, USA
{zpluo, milos}@cs.pitt.edu

Abstract

Learning of classification models in practice often relies on nontrivial human annotation effort in which humans assign class labels to data instances. As this process can be very time consuming and costly, finding effective ways to reduce the annotation cost becomes critical for building such models. In this work we solve this problem by exploring a new approach that actively learns classification models from groups, which are subpopulations of instances, and human feedback on the groups. Each group is labeled with a number in [0,1] interval representing a human estimate of the proportion of instances with one of the class labels in this subpopulation. To form the groups to be annotated, we develop a hierarchical active learning framework that divides the whole population into smaller subpopulations, which allows us to gradually learn more refined models from the subpopulations and their class proportion labels. Our extensive experiments on numerous data sets show that our method is competitive and outperforms existing approaches for reducing the human annotation cost.

1 Introduction

Learning of classification models from real-world data often requires extensive human annotation effort on labeling data instances. As this annotation process is often time-consuming and costly, the key challenge is to find effective ways to reduce the annotation effort while guaranteeing that models built from limited labeled data are accurate enough to be applied in practice. One popular machine learning solution is active learning that sequentially selects examples to be labeled next. Active learning has been successfully applied in domains as diverse as computer vision, natural language processing and bio-medical data mining [Nguyen *et al.*, 2014; Valizadegan *et al.*, 2013].

Despite enormous progress in active learning research in recent years, the majority of current active learning solutions focus on *instance-based* methods that query and label individual data instances. Unfortunately, this may limit their applicability when targeting complex real-world classification tasks. There are two reasons for this. First, when the labeling budget is severely restricted, a small number of labeled data may not properly cover or represent the entire input space. In other words, the data selected by active learning are likely to suffer from *sampling bias* problem. To mitigate this issue [Dasgupta and Hsu, 2008] has developed a hierarchical active learning approach to sample instances in a more robust way which is driven by not only the currently sampled data, but also the underlying structure in the data.

Second, instance-based learning framework often assumes instances are easy to label for humans. But this is not always true, especially when each data instance has many features and some specific features are valued with high precision numbers which can be very intricate for human annotators to assess. For example, when a physician diagnoses a patient (e.g. for possible heart condition) he/she must review the patient record that consists of complex collections of results, symptoms and findings (such as *age, BMI, glucose levels, HbA1c blood test, blood pressure, heart rate* etc.). The review and the assessment of these records w.r.t. a specific condition may become extremely time-consuming as it often requires physicians to peruse through a large quantity of data.

In light of this, novel active learning methods based on group queries have been proposed: AGQ+ [Du and Ling, 2010] and RIQY [Rashidi and Cook, 2011]. The basic idea here is to (1) embody similar instances together as a group, (2) induce the most relevant rules which are conjunctive patterns of the input feature space to represent the group and (3) solicit a generic label on the group instead of on any specific instance. The group label is a number in [0,1] interval which represents a human estimate of the proportion of instances that belongs to one of the classes in the subpopulation, or equivalently, the probability with which an instance with that class label is drawn from the subpopulation. This line of work especially RIQY has shown empirically that active learning with generic group proportion feedback works more efficiently than instance-based active learning.

To solve the original annotation cost problem from a new perspective, in this paper we present a novel group learning framework called HALG (**H**ierarchical **A**ctive **L**earning with **G**roup proportion feedback) that combines the strengths of Dasgupta's hierarchical solution and the group queries introduced in RIQY. More specifically, by assuming a pool of unlabeled data instances at hand, our framework starts with a hierarchical clustering on the pool which will generate a hierarchy of groups. Then proceeding from the top levels in the hierarchy to lower ones, we actively select the most *influential* groups to be labeled next. The influence is measured by how much the groups will update the base classification model once they get labeled. This active selection strategy borrows the merit of *maximum model change* criterion [Roy and McCallum, 2001; Freytag *et al.*, 2014] used in instance-based active learning. In terms of model learning, we introduce a simple, yet effective, algorithm based on sampling that is able to learn any instance-level classification models from labeled groups. Our empirical study shows our HALG framework can estimate very well the underlying model parameters in a very few queries.

In the next section, let us review in more detail the related work that has inspired ours and also, on the other hand, see how our method can improve upon theirs.

2 Related Work

Cluster-Based Active Learning Our hierarchical learning framework is motivated by the work of [Dasgupta and Hsu, 2008] that leverages a pre-compiled hierarchical cluster tree to drive the instance selection procedure. They start learning from a few coarse-grained clusters and gradually split clusters that are impure (w.r.t. the known class labels) to smaller clusters such that the label entropy is reduced. In terms of model learning, not only the labeled instances but also the instances with predicted labels that fall to *sufficiently pure* clusters are used for learning. While their approach is able to reduce the sampling bias, learning with predicted labeled data can be risky especially when the class distribution is severely unbalanced. As a result it may take numerous queries before the first instance from the minority class is sampled. In our work, we directly query clusters regarding their class proportion labels and use them to learn classification models. As the proportion labels reflect the minority class proportion, our approach is friendlier to learning with highly unbalanced class distribution as opposed to Dasgputa's method.

Active Learning from Group Feedback AGQ+ [Du and Ling, 2010] and RIQY [Rashidi and Cook, 2011] are the first attempts to actively learn classification models from group rather than instance based feedback. As mentioned in introduction, we share the same motivation that in many practical domains annotators prefer to work with group-level queries which are shorter (in terms of feature space), less confusing and more intuitive. Here we borrow from RIQY's paper and their example in the heart disease domain:

An instance class-label query in heart disease domain could be formulated as: *"Assess whether the patient with (sex=female) & (age=39) & (chest pain type=3) & (fasting blood sugar=150 mg/dL) ... (20 more features) suffers from a heart disease?"*. The answer to this query is binary, reflecting the agreement or disagreement of an annotator with the instance falling to the heart disease class.

On the contrary, **a group query** which uses conjunctive patterns in the feature space can represent a generic population. For instance: *"What proportion of patients who are (sex=female) & (40<age<50) & (chest pain type=3) and (fasting blood sugar within [130,150] mg/dL) ... (not necessarily using all the features) suffer from a heart disease?"*. The answer to this query is an empirical assessment of the proportion of heart disease patients in the population, say *about 75% patients within this population have heart disease.*

As for their methodology, each active learning cycle starts by aggregating multiple similar instances together as a group which is centered at the most uncertain instance. RIQY represent this group by using conjunctive patterns and presents it to a human to ask for its proportion label. The returned label is propagated to all instances within the group and each instance is finally assigned a binary label based on some confidence. Therefore, any standard instance-level learning algorithm can be applied to learn a classification model. AGQ+ follows a similar process except that they assign soft labels to instances and adopt a weighted instance learning algorithm.

Both AGQ+ and RIQY have shown that active learning with group proportion feedback can find the true model faster than active learning with instance based feedback. Despite the good empirical results, their methodology can be further refined. One potential improvement is to devise new ways to form the groups. RIQY forms a group by picking one data instance first and then building a group around that instance. This is somewhat ad-hoc and it might not pursue the most meaningful groups. In our work, we form groups through hierarchical clustering which is a systematic way to discover informative groups that can better cover the input feature space.

Learning from Group Proportion Feedback Multiple works [Quadrianto *et al.*, 2009; Kück and de Freitas, 2012; Patrini *et al.*, 2014; Yu *et al.*, 2013] study the problem of learning instance-level classifiers from apriori given groups and their class proportion statistics. The problem formulation fits scenarios like political election, online purchasing or spam filtering. For example, we can easily obtain the percentage of voting results on election in each county and use these group statistics to predict individual's voting preference. In terms of model learning, their algorithm either uses the proportion labels to estimate the sufficient statistics required by the final likelihood function [Quadrianto *et al.*, 2009; Patrini *et al.*, 2014], or develops models that generate instance labels that are consistent with the proportions [Kück and de Freitas, 2012; Yu *et al.*, 2013]. There are two minor limitations of their approaches. First, their optimization procedures can be time-consuming and hence they may not work well in combination with active learning where models are updated frequently. Second, their algorithms are restricted to a limited number of models. Because of that, in our work, we resort to a very simple learning algorithm based on instance sampling which can efficiently learn many popular probabilistic parametric models. Finally, we note, that this line of work does not deal with are the problems of group generation, group representation and active querying of groups that are essential for our active learning methodology.

3 Our Framework

Our HALG learning framework, summarized in Algorithm 1, is designed to learn a binary classification model from group

Algorithm 1: Our HALG Framework

```
Input: An unlabeled data pool $\mathcal{U}$; A labeling budget
Output: A binary classification model $P(y|\boldsymbol{x};\hat{\boldsymbol{\theta}})$
 1: $T \leftarrow$ Perform hierarchical clustering on $\mathcal{U}$
 2: $T_G \leftarrow$ Adjust $T$ to form a new tree of groups
 3: Query $(T_G)'s\ root$;
 4: Fringe $F^{(1)} \leftarrow \{(T_G)'s\ root\}$;
 5: Active learning time $t \leftarrow 1$;
 6: repeat
 7:     Learn $P(y|\boldsymbol{x};\hat{\boldsymbol{\theta}}^{(t)})$ from current $F^{(t)}$
 8:     Split a group $G_*$ in $F^{(t)}$ based on $P(y|\boldsymbol{x};\hat{\boldsymbol{\theta}}^{(t)})$
 9:     Query the labels of $G_*$'s children from labelers
10:     $F^{(t+\Delta t)} \leftarrow \{F^{(t)} - G_*\} \cup \{G_*\text{'s children}\}$
11:     $t \leftarrow t + \Delta t$ ($\Delta t$ = # of $G_*$'s children)
12: until the labeling budget runs out
13: return $P(y|\boldsymbol{x};\hat{\boldsymbol{\theta}}^{(t)})$
```

queries and group proportion feedback. Our framework starts with a hierarchical clustering performed on all unlabeled data (line 1) to identify initial groups and their hierarchy. The hierarchy is then adjusted such that only groups that can be compactly described via conjunctive patterns (line 2) are retained. This adjustment returns the tree of final groups. These groups are annotated via active learning in the top-down fashion. During the active learning process we always maintain a fringe of groups (line 4) which is a complete partition of the all the data, and perform active querying and learning based on this fringe (Line 6-12). In each active learning cycle, we use the *maximum model change* strategy to (1) select the most influential group in the fringe to split, (2) assess its child groups by human annotators and (3) replace the group with its children in the fringe. Every time the fringe is refined, the base classification model is updated and used to calculate the model change in the next cycle. In the following sections we will describe the details of our framework. Section 4 introduces the groups concepts. Section 5 presents our efficient solution for learning the instance-level classification model from group proportion feedback. Finally section 6 explains our active group selection procedure.

4 The Concept of Groups

In our work, we use the term *group* to denote a set of instances in the data set. We assume the data set is defined by a matrix of real numbers $\mathcal{U}_{n\times m}$ consisting of n m-dimensional instances. We start to build the groups by applying standard hierarchical clustering (using the ward linkage [Ward Jr, 1963]) on $\mathcal{U}$ to output a tree T of our initial groups.

Compact Group Description These initial groups, however, are hard to present to humans for assessment since they would require us to enumerate all instances that fall into the group. To present groups to annotators in a human-friendly way, we compactly describe them using conjunctive patterns over the input space features. As we have seen earlier, a group of patient instances may be described as: *(gender=male) & (heart rate 80-100) & (temperature 100-110F)* etc. This representation matches the hypercube definition of regions of a typical decision tree algorithm. Hence, we employ a C4.5 [Quinlan, 2014] classifier to automatically learn a more compact description (idea originally introduced by RIQY) of the group of instances found via clustering. More specifically, if we want to learn the description of a group G_i, we mark all instances in G_i as **1** and the rest of data instances $(\mathcal{U} - G_i)$ as **0**. Then a C4.5 classifier will output a set of hypercubes $\mathcal{C}(G_i)$ that could potentially describe G_i. The match (or fit) of each hypercube $c \in \mathcal{C}(G_i)$ to G_i can be assessed in terms of: (1) $precision$, which measures the proportion of data in c that are actually coming from G_i; and (2) $recall$, which measures the proportion of data in G_i that c can capture. As both metrics are important, we adopt $F1score = \frac{2\times precision \times recall}{precision + recall}$ to be the final quality metric to select the hypercube that best fits the description of group G_i. That is, the description of G_i is a hypercube which is $\arg\max_{c\in\mathcal{C}(G_i)} F1score(c)$.

The Final Tree of Hypercube-like Groups When the above C4.5-based group matching is performed on the clusters of the original hierarchical tree T, there may be some clusters (groups) that are not matched well by hypercube-shaped regions. In such a case, its best hypercube match comes with intolerably low $F1score$s. To mitigate this issue, we prune the original tree structure T to form a new tree T_G such that only well-fitted hypercube-like groups are preserved in T_G. More formally, we say that a cluster is *hypercube-like* if it can be approximated by hypercube with a minimum $precision(\geq 0.5)$ and $recall(\geq 0.5)$. Our goal is to preserve and approximate only hypercube-like regions in the original tree. We implement this idea by starting from the root of the tree T and by checking in the top-down fashion if the descendant clusters are hypercube-like. If a descendant cluster is not hypercube-like we exclude it from the tree by directly reconnecting its parent with its children clusters. As a result, the original binary tree T becomes a multi-nary tree T_G in which the clusters are all hypercube-like. We use the tree T_G to form the groups for the subsequent active learning process.

Group Proportion Label The human assessment of each group (approximated well by a hypercube-like region) is made via a proportion label, which is an estimate of the proportion of one of the classes in the group. For example, people could say that *"90% of instances in a certain group are positive"*. Or alternatively, we can interpret the proportion label as an instance-level likelihood. For instance, *"Instances in such group are 90% likely to be positive"*. To assess the label of each group, annotators will only need to review the description of the best hypercube-like region matching it, and thus, they do not have to explore each data instance that falls in the group individually.

The Fringe of Groups After we form the group hierarchy T_G, the top-down active learning process begins. In each active learning cycle, we maintain a fringe F of labeled groups which is a complete partition of all unlabeled data. Initially, we make one query to obtain the label of (T_G)'s root which can be interpreted as the prior probability of classes, and put the labeled root into $F^{(t)}$ at $t = 1$. Here t is the active learning time-step, basically counting the number of group queries made so far. As t increases, finer and finer groups and their proportion labels will replace their parents in $F^{(t)}$.

In the following, we explain how to learn a model from labeled groups in $F^{(t)}$ and how the model will assist us in choosing the group that should be split next (Section 6).

5 Learning a Model from Labeled Groups

Suppose at the active learning time t there are N labeled groups in the current fringe: $F^{(t)} = \{(G_i, \mu_i)\}_{i=1}^{N}$, where G_i denotes a group and μ_i its proportion label. Each group $G_i = \{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}$ contains n_i instances and its label $\mu_i \in [0, 1]$ is assumed to represent the positive class proportion in binary classification setting. Our goal is to learn a base model $P(y|\boldsymbol{x}; \boldsymbol{\theta})$ which is an instance-level discriminative classifier from $F^{(t)}$. Our learning algorithm does this by (1) sampling sufficiently many labeled instances from the labeled groups and (2) feeding them to the instance-level classification learning procedure.

Sampling We create a bootstrap sample $S^{(t)} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{K}$ of K labeled instances from $F^{(t)}$. Each $\boldsymbol{x}_k$ is uniformly sampled with replacement from the unlabeled data pool $\mathcal{U}$, and y_k is sampled from an independent Bernoulli process with parameter equal to μ_i which is the proportion label of group G_i that $\boldsymbol{x}_k$ resides in. When K is sufficiently large the randomness in the $\boldsymbol{x}$ part in $S^{(t)}$ is not of interest and we only focus on the randomness of y in $S^{(t)}$.

Learning With the sample $S^{(t)}$, we can estimate the model parameter vector as $\hat{\boldsymbol{\theta}}^{(t)}$ through maximum likelihood estimation (MLE). Although $\hat{\boldsymbol{\theta}}^{(t)}$ may vary because of the randomness in $S^{(t)}$, from the theoretic standpoint, when moderate MLE assumptions required by the Central Limit Theorem are satisfied, $\hat{\boldsymbol{\theta}}^{(t)}$ asymptotically follows a Normal distribution conditioned on $\{\boldsymbol{x}_k\}^{(t)}$ in $S^{(t)}$ [Bickel and Doksum, 2015]:

$$\hat{\boldsymbol{\theta}}^{(t)} \overset{asymp.}{\sim} \mathcal{N}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\Sigma}^{(t)}) \tag{1}$$

Here $\boldsymbol{\theta}^{(t)} = \mathbb{E}[\hat{\boldsymbol{\theta}}^{(t)}]$ is the converged parameter when $K \to \infty$, and the variance $\boldsymbol{\Sigma}^{(t)}$ is the inverse of Fisher information matrix $\mathcal{I}_K(\boldsymbol{\theta}^{(t)})$ depending on the actual finite sample size K. So the randomness of $\hat{\boldsymbol{\theta}}^{(t)}$ is bounded by this asymptotic Normal distribution and the larger K is, the smaller the variance would be.

In practice, however, $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\Sigma}^{(t)}$ are unknown. Yet usually $\boldsymbol{\Sigma}^{(t)}$ is approximated by the the MLE estimator as $\hat{\boldsymbol{\Sigma}}^{(t)}$ (for example when calculating the confidence interval of $\hat{\boldsymbol{\theta}}^{(t)}$). Further when K is large, the difference between $\hat{\boldsymbol{\theta}}^{(t)}$ and $\boldsymbol{\theta}^{(t)}$ is sufficiently small (or equivalently, the confidence interval of $\hat{\boldsymbol{\theta}}^{(t)}$ is very tight). So as another approximation, we replace $\boldsymbol{\theta}^{(t)}$ by $\hat{\boldsymbol{\theta}}^{(t)}$ in the normal distribution. That is, it is appropriate to assume the following asymptotic distribution holds when K is sufficiently large:

$$\hat{\boldsymbol{\theta}}^{(t)} \overset{asymp.}{\sim} \mathcal{N}(\hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)}) \tag{2}$$

In our experiments each label is sampled from 5 to 10 times depending on data sets and this gives us $K \sim 10^4$ magnitude which is large enough to provide a very small $\hat{\boldsymbol{\Sigma}}^{(t)}$.

Sampling Bias Sampling bias is a problem faced by active learning methods, which is caused by a shortage of labeled data to learn the models. It may lead to samples that do not reflect properly the true underlying distribution generating the data. In Equation (1), $\boldsymbol{\theta}^{(t)}$ reflects what one can learn after t queries are made, and this estimate certainly differs from the underlying true model parameter (denoted by $\boldsymbol{\theta}^{(\infty)}$) that we aim to learn. One of the reasons is that each label y_k in the training data $S^{(t)}$ is sampled using its Bernoulli parameter equal to its group proportion label among all the instances, rather than on $\boldsymbol{x}_k$. Hence, the difference between $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}^{(\infty)}$ is due to the variability induced by the label assignments for instances in each individual group. To understand this issue, we define the sampling bias for each labeled group (G_i, μ_i) as:

$$bias(G_i) = \mathbb{E}[w_i] \tag{3}$$

where $\mathbb{E}[w_i]$ is the expected number of wrong labels if we randomly sample the labels of all the n_i instances in G_i w.r.t. its proportion label μ_i. We calculate $\mathbb{E}[w_i]$ through the following procedure:

1. For each instance in G_i, sample its label via an independent Bernoulli process with the parameter $= \mu_i$. This creates n_i sampled labels;
2. Derive the distribution of w_i, i.e. the number of mismatches between the sampled labels and the true labels. Although the true labels are unknown, each true label similarly follows an independent Bernoulli distribution with the same parameter μ_i. Therefore, the probability of mismatch for each instance also follows an independent Bernoulli distribution with parameter $= P(mismatch) = P[false\ positive] + P[false\ negative] = 2\mu_i(1 - \mu_i)$. Then apparently w_i follows a Binomial distribution $Bin(n_i, 2\mu_i(1 - \mu_i))$;
3. Lastly, $bias(G_i) = \mathbb{E}(w_i) = 2\mu_i(1 - \mu_i)n_i$.

The definition of $bias(G_i)$ clearly shows that larger n_i or more uncertain (more extreme) μ_i accounts for more sampling bias introduced by G_i.

6 Active Refinement of Groups

In this section, we explain our active learning strategy that refines the fringe $F^{(t)}$ at time t to generate $F^{(t+\Delta t)}$, where Δt is the number of new queries made. The gist of our approach is to split the most influential group $G_* \in F^{(t)}$ w.r.t. updating the base model, query the labels of its child groups, and then replace G_* with its child groups in the fringe.

Intuitively, according to the definition of $bias(G_i)$, we want to split a large and/or impure group such that the label uncertainty of large groups can be reduced. Moreover, we would also like the current model $\boldsymbol{\theta}^{(t)}$ to have a big update after the split such that it converges quickly to $\boldsymbol{\theta}^{(\infty)}$. Therefore, we adopt *maximum model change* criterion as our active learning strategy. The key idea is to split the group such that the model distribution $\mathcal{N}(\hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)})$ can be updated most from time t to $(t+\Delta t)$. To achieve this goal, we need to *guess* what would happen after we split each group. More specifically, we need to (1) infer the most probable labels of each G_i's child groups and (2) estimate how much the model distribution will change if we replace G_i with its child groups (with inferred labels) in the fringe for learning.

Label Inference One reasonable way to infer the label of a child group is to use the empirical mean statistics of all the instances inferred by the base model. Formally, let us suppose each group G_i in the current fringe $F^{(t)}$ has C_i child groups and each child group $G_{ic} = \{\boldsymbol{x}_{(ic)j}\}_{j=1}^{n_{ic}}$ has n_{ic} many instances for $c \in [1, C_i]$. The label of each child group $\hat{\mu}_{ic}$ can be inferred as:

$$\hat{\mu}_{ic} = \frac{1}{n_{ic}} \sum_{j \in [1, n_{ic}]} P(y_{(ic)j} = 1 | \boldsymbol{x}_{(ic)j}; \hat{\boldsymbol{\theta}}^{(t)}) \tag{4}$$

Then we can create a new fringe $F_{[i]}^{(t)} = \{F^{(t)} - (G_i, \mu_i)\} \cup \{(G_{ic}, \hat{\mu}_{ic})\}_{c=1}^{C_i}$, feed it to our parameter optimization algorithm and obtain a new model distribution, denoted by $\mathcal{N}(\hat{\boldsymbol{\theta}}_{[i]}^{(t)}, \hat{\boldsymbol{\Sigma}}_{[i]}^{(t)})$. So this $\mathcal{N}(\hat{\boldsymbol{\theta}}_{[i]}^{(t)}, \hat{\boldsymbol{\Sigma}}_{[i]}^{(t)})$ represents what the new model distribution would look like if we *were* to split the group G_i. Before we compare $\mathcal{N}(\hat{\boldsymbol{\theta}}_{[i]}^{(t)}, \hat{\boldsymbol{\Sigma}}_{[i]}^{(t)})$ to $\mathcal{N}(\hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)})$, one important note is that we should fix $\{\boldsymbol{x}_k\}$ in the sample $S^{(t)}$ for re-learning, as the two asymptotic normal distributions are comparable only if they are learned conditioned on the same $\{\boldsymbol{x}_k\}$. So when learning $\mathcal{N}(\hat{\boldsymbol{\theta}}_{[i]}^{(t)}, \hat{\boldsymbol{\Sigma}}_{[i]}^{(t)})$, only the labels in group G_i are re-sampled.

Model Change In terms of estimating the model change, we use KL-divergence to measure the distribution change from the current model distribution $\mathcal{N}(\hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)})$ to $\mathcal{N}(\hat{\boldsymbol{\theta}}_{[i]}^{(t)}, \hat{\boldsymbol{\Sigma}}_{[i]}^{(t)})$ for each group G_i. Finally, we select the group G_* with the maximum change to split:

$$G_* = \arg\max_{G_i \in F^{(t)}} D_{KL}(\mathcal{N}(\hat{\boldsymbol{\theta}}_{[i]}^{(t)}, \hat{\boldsymbol{\Sigma}}_{[i]}^{(t)}) \,||\, \mathcal{N}(\hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{\Sigma}}^{(t)}))$$

After the split, G_*'s children $\{G_{*c}\}$ are sent for querying and a new fringe is updated as:

$$F^{(t+\Delta t)} = \{F^{(t)} - (G_*, \mu_*)\} \cup \{(G_{*c}, \mu_{*c})\}_{c=1}^{C_*}$$

A minor caveat when calculating the model change is that when the number of groups in $F^{(t)}$, i.e. N, is large, it may be time-consuming to solve the N optimization procedures sequentially. However, we note that as the these N procedures are independent, parallel processing can be appropriately deployed to reduce the total runtime.

Dataset	# of Data	# of features	Major Class	Feature Type
Seismic	2584	18	93%	N-O-C
Ozone	1847	72	93%	N
Pima	768	8	65%	N-C
Spam	4601	57	60%	N-O
Music	1059	68	53%	N
Wine	4898	11	67%	N
Wine$_{ub}$	1895	11	95%	N
Gamma	5000	10	65%	N
SUSY	5000	18	55%	N

Table 1: 9 UCI data sets. 'N', 'O'and 'C'stand for 'Numeric', 'Ordinal'and 'Categorical'respectively.

7 Experiments

We conduct an empirical study to evaluate our proposed approach on 9 general binary classification data sets collected from UCI machine learning repository [Asuncion and Newman, 2007]. The purpose of this study is to research how efficiently (in terms of the number of queries) our HALG framework can learn classification models in cost-sensitive tasks.

7.1 Data Sets

The 9 data sets come from a variety of real life applications:

- **Seismic**: Predict if seismic bumps are in hazardous state
- **Ozone**: Detect ozone level on some days
- **Pima**: Diagnose diabetes disease among Indian women
- **Spam**: Classify spam commercial emails
- **Music**: Find the geographical origin of music
- **Wine**: Predict wine quality
- **Gamma**: Detect γ particles in Cherenkov telescope
- **SUSY**: Distinguish a signal or background process

Table 1 summarizes the basic statistics of the data sets. Some have been widely used in the previous active learning work: *Wine, Pima* [Xue and Hauskrecht, 2017]; some have high-dimensional feature space: *Ozone, Spam, Music*; and some carry highly unbalanced class distribution: *Seismic, Ozone*, *Wine unbalance* (simulated from *Wine*).

7.2 Methods Tested

We compare our HALG to 3 other methods:

1. **DWUS**: Density-Weighted Uncertainty Sampling which combines uncertainty sampling and the structure in data [Settles, 2012] to decide queries;
2. **RIQY**: the state-of-the-art active learning with group proportions [Rashidi and Cook, 2011];
3. **HS**: Hierarchical Sampling [Dasgupta and Hsu, 2008]

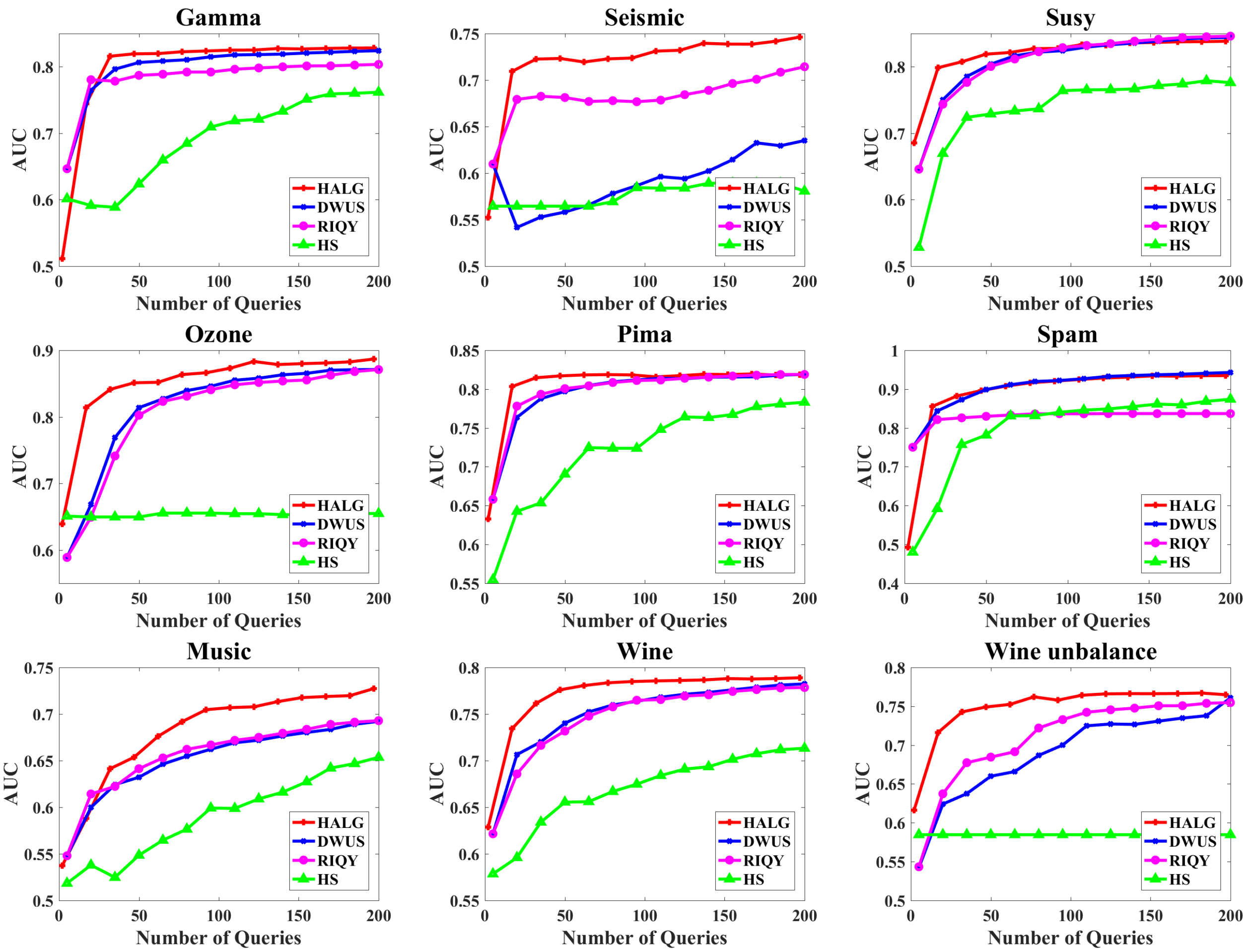

Figure 1: Performances of different methods on 9 UCI data sets.

7.3 Experimental Settings

Data Split To run the experiments, we split each data set into three disjoint subsets: the initial labeled data set (about 1%-2% of all available data), a test data set (about 25% of data) and a training data set $\mathcal{U}$ (the rest of the data) that is used for training and active learning. Please note only DWUS and RIQY require the initial labeled data to start training.

Group Proportion Label Feedback To simulate the labeling process and group proportion feedback we use counts of instances and their labels to estimate the class proportion represented by the group. We note the same method was applied to test RIQY in its original paper.

Evaluation Metrics We adopt Area Under the Receiver Operating Characteristic curve (AUC) to evaluate the quality of the learned classification model (in our case Logistic regression) on the test data. Our graphs plot the AUC scores sequentially as the number of queries gradually increases to 200. All results are averaged over 20 runs in different random splits. When generating the results we also assumed the different types of queries are equivalent in terms of their costs, although in reality different query types may carry different costs. For example, a group query may be easier for objects like medical records and patient data, but not for images.

7.4 Experiment Results

The main results are shown in Figure 1. Overall, our HALG approach (in red line) is able to outperform other methods on the majority of the data sets and is close to the best method on the remaining sets. It comes with two advantages. First, initially when the labeling budget is severely limited, learning with labeled groups is superior to learning with the same number of labeled instances, simply because generic group queries can provide richer class information than specific instance queries. Second, the initial steep slopes and early convergence in our learning curves lend great credence to our active learning strategy that it is capable of selecting the most influential group to split and consequently it can accelerate the convergence rate of the method.

Unbalanced Classes For data sets *Seismic, Ozone* and *Wine unbalance* with unbalanced class distribution, our method performs even better as it could capture properly the

Dataset	FRR	Dataset	FRR
Wine	42%	Spam	60%
Ozone	89%	Music	88%
Gamma	34%	SUSY	61%
Seismic	61%	Pima	40%

Table 2: The averaged feature reduction rate (FRR) of group queries

minority class information via proportion labels. In contrast, the instance-based methods may find these proportions very slowly. Also note that hierarchical sampling (HS) completely failed because it always determines the labels of unlabeled instances by the majority vote if they belong to *pure enough* (but not entirely pure) clusters, and hence it may miss to capture the minority class information.

Complexity of Group Queries Our last experiment aims to analyze the benefit our group queries in terms of the query complexity. We assess this complexity by using *feature reduction rate* which is defined as $1 - \frac{\#\ features\ to\ describe\ G}{\#\ all\ features}$. This definition clearly reflects the savings due to the description of the group G relative to the complexity of the full feature space. The results in Table 2 suggest that on average it only takes about one third to one half of features to distinguish one group from the other groups. This can considerably simplify the interaction with human annotators especially when data objects are high-dimensional, and when the active learning queries need to present only the features relevant for the group and its query.

8 Conclusions

We have developed and presented HALG - a new framework that can actively learn instance-based classifiers from group proportion feedback. The groups used in our framework are formed by hierarchical clustering and they can be refined actively based on the maximum model change criterion such that the model learned from these groups can converge quickly to a very good model. In terms of application, our framework is best suited for problems when instance labeling is hard due to high-dimensional objects we need to label, and when the group description is more compact and depends only on a limited number of features necessary to distinguish the different groups.

Acknowledgements

The work presented in this paper was supported by NIH grants R01GM088224 and R01LM010019. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of NIH.

References

[Asuncion and Newman, 2007] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[Bickel and Doksum, 2015] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volume I*, volume 117. CRC Press, 2015.

[Dasgupta and Hsu, 2008] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215. ACM, 2008.

[Du and Ling, 2010] Jun Du and Charles X Ling. Asking generalized queries to domain experts to improve learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(6):812–825, 2010.

[Freytag *et al.*, 2014] A. Freytag, E. Rodner, and J. Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision*, pages 562–577. Springer, 2014.

[Kück and de Freitas, 2012] Hendrik Kück and Nando de Freitas. Learning about individuals from group statistics. *CoRR*, abs/1207.1393, 2012.

[Nguyen *et al.*, 2014] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification models with soft-label information. *Journal of the American Medical Informatics Association*, 21(3):501–508, 2014.

[Patrini *et al.*, 2014] Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pages 190–198, 2014.

[Quadrianto *et al.*, 2009] Novi Quadrianto, Alex J Smola, Tiberio S Caetano, and Quoc V Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(Oct):2349–2374, 2009.

[Quinlan, 2014] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[Rashidi and Cook, 2011] Parisa Rashidi and Diane J Cook. Ask me better questions: active learning queries based on rule induction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 904–912. ACM, 2011.

[Roy and McCallum, 2001] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

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

[Valizadegan *et al.*, 2013] Hamed Valizadegan, Quang Nguyen, and Milos Hauskrecht. Learning classification models from multiple experts. *Journal of biomedical informatics*, 46(6):1125–1135, 2013.

[Ward Jr, 1963] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[Xue and Hauskrecht, 2017] Yanbing Xue and Milos Hauskrecht. Active learning of classification models with likert-scale feedback. In *SIAM Data Mining Conference, 2071*. SIAM, 2017.

[Yu *et al.*, 2013] Felix Yu, Dong Liu, Sanjiv Kumar, Jebara Tony, and Shih-Fu Chang. \proptosvm for learning with label proportions. In *ICML*, pages 504–512, 2013.