Active learning by Greedy Split and Label Exploration

Alyssa Herbst
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061
alyssa2@vt.edu

Bert Huang
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061
bhuang@vt.edu

CoRR 2019

Background

Active learning approaches aim to acquire labels for data most useful for training specific model families.This goal can introduce significant bias. For example, methods such as uncertainty sampling prefer labeling points close to a model's decision boundary.so the distribution of labeled points will be highly dependent on the model family being trained. The acquired labels may not be as useful for training other model families.

A better approach is to use the structure of the data to determine which example to label.

Background

Hierarchical Sampling for Active Learning

- It constructs a hierarchical clustering and adaptively determine how to prune the clustering.

- The idea behind HSAL is that if examples sampled from a cluster exhibit high label uniformity, i.e., are mostly the same label, then it can be inferred that the rest of the cluster is likely to have that majority label.

Motivation

- The effectiveness of HSAL thus relies on the quality of the clustering and how well it aligns with the true labels of the classification task. In many settings, a feature-based clustering can have low label uniformity, resulting in negligible gains when using HSAL.

- Using some acquired data to guide the partitioning can drastically improve the uniformity of the partitions.

Active learning by Greedy Split and Label Exploration

The algorithm performs a greedy search of the full dataset for uniform subsets—subsets for which it is believed that most true labels are the same. In the greedy search, we allow two types of operations:

1.query the oracle to label a data point

2. split a dataset into subsets if we believe that a dataset is not uniform.

Algorithm

Variable	Definition
$\tilde{f}_v$	Lower bound of correct labels in dataset v
B	Budget of oracle calls
$\hat{y}_v$	Majority label of all empirically sampled data in dataset v
n_v	Sample size of dataset v
m_v	The number of samples of majority label $\hat{y}$ in a single dataset v
N_v	The total number of data points in dataset v, including labeled and unlabeled data
p_v	The true proportion of data points in dataset v with label $\hat{y}_v$
p'_v	The estimated proportion of data points in dataset v with majority label $\hat{y}_v$, based on the sampled data points seen so far
t_v	$p'_v - p_v$, the "buffer size"
k	The number of unique label classes
a	Possible actions, $a \in \{split, label\}$
v	Node, or subset of the full dataset $\{x_1, x_2, ...x_N\}$
ρ	Probability of using labeled data to train a supervised splitting function (as opposed to for calculating bound $\tilde{f}$)
T	Set of leaf nodes in tree

Algorithm

Algorithm 1 Pseudocode for Greedy Split and Label Exploration (GSAL)

Require: Dataset $X = \{x_1, x_2, ...x_N\}$, $Y = \varnothing$, budget B, training ratio ρ, quality threshold q, splitting function `split`, and oracle labeler `oracle`

Initialize a leaf node with all data; $v_r \leftarrow \{x_1, x_2, ...x_N\}$
Create a tree with root v_r; $T \leftarrow \{v_r\}$
while $B > 0$ **do**
  Estimate expected correct labels for all leaves; $S \leftarrow \left[\tilde{f}(v_i) \text{ for leaf } v_i \in T\right]$
  Estimate expected correct labels for all leaves if label; $S_{\text{label}} \leftarrow \left[\tilde{f}_{\text{label}}(v_i) \text{ for leaf } v_i \in T\right]$
  Estimate expected correct labels for all leaves if split; $S_{\text{split}} \leftarrow \left[\tilde{f}_{\text{split}}(v_i) \text{ for leaf } v_i \in T\right]$
  Choose best action $a*$ and node $v*$ corresponding to

$$a*, v* \leftarrow \underset{a \in \{\text{split, label}\}, v \in T}{\arg\max} S_a[v] - S[v]$$

  if $a* =$ label **then**
    Select $x_{v*,i}$ randomly from $v*$
    $y_{v*,i} \leftarrow \texttt{oracle}(x_{v*,i})$
    Choose with probability $(1 - \rho)$ if adding to the label set.
    if label set **then**
      Update majority label $\hat{y}*$ for node $v*$ with new label $y_{v*,i}$
      $m_{v*} \leftarrow$ number of non-training labeled points in $v*$ of majority label $\hat{y}*$
      Add new label to set of labels; $Y \leftarrow Y \cup \{y_{v*,i}\}$
      $n_{v*} \leftarrow n_{v*} + 1$
      $B \leftarrow B - 1$ if x_i has never been labeled
    else
      Add $x_{v*,i}$, $y_{v*,i}$ to $v*$'s isolated training set.
    end if
  else
    Split parent node into set of child nodes; $U \leftarrow \texttt{split}(v*)$
    For each $u \in U$, set $n_u = 0$, $m_u = 0$ and empty isolated training set
    Remove parent node from tree ; $T \leftarrow T - v*$
    Add child nodes to tree; $T \leftarrow T \cup U$
  end if
end while
for $v \in T$ and $p'_v > q$ **do** Set $y_{v,i} = \hat{y}_v$ and add label to return set $Y \leftarrow Y \cup \{y_{v,i}\}$ **end for**
return Y

Derivation and Analysis of Lower Bound $\tilde{f}$

The bound is based on Hoeffding Inequality, a distribution-free concentration bound that becomes tighter to true values as more label are obtained. Hoeffding bounds are defined with a deviation t.

Hoeffding's Inequality provides this guarantee:

$$\Pr\left(p \leq \frac{m}{n} - t\right) \leq \exp\left(-2nt^2\right).$$

m: The number of points with the majority label.
n: The number of points with known labels
p: The true proportion of the majority label
m/n: Current estimate of the probability of the majority label

Derivation and Analysis of Lower Bound $\tilde{f}$

The expected number of correct labels:

$$
\begin{aligned}
f(m,n,N) = pN &\geq n + \underbrace{\left(1 - \Pr\left(p \leq \frac{m}{n} - t\right)\right)}_{\text{Probability that } p > \frac{m}{n} - t}(N-n)\left(\frac{m}{n} - t\right) \\
&\geq n + \left(1 - \exp\left(-2nt^2\right)\right)(N-n)\left(\frac{m}{n} - t\right) \\
&:= \tilde{f}(m,n,N,t),
\end{aligned}
$$

$$
\tilde{f}(m,n,N) = \max_{t \in [0,1]} \left(n_v + \left(1 - \exp\left(-2n_v t^2\right)\right)(N_v - n_v)\left(\frac{m_v}{n_v} - t\right)\right).
$$

$$
\tilde{f}_{\text{label}}(v) = \tilde{f}(m_v + 1, n_v + 1, N).
$$

$$
\tilde{f}_{\text{split}} = \sum_{u \in U} \tilde{f}(m_u, n_u, N_u).
$$

Experiment

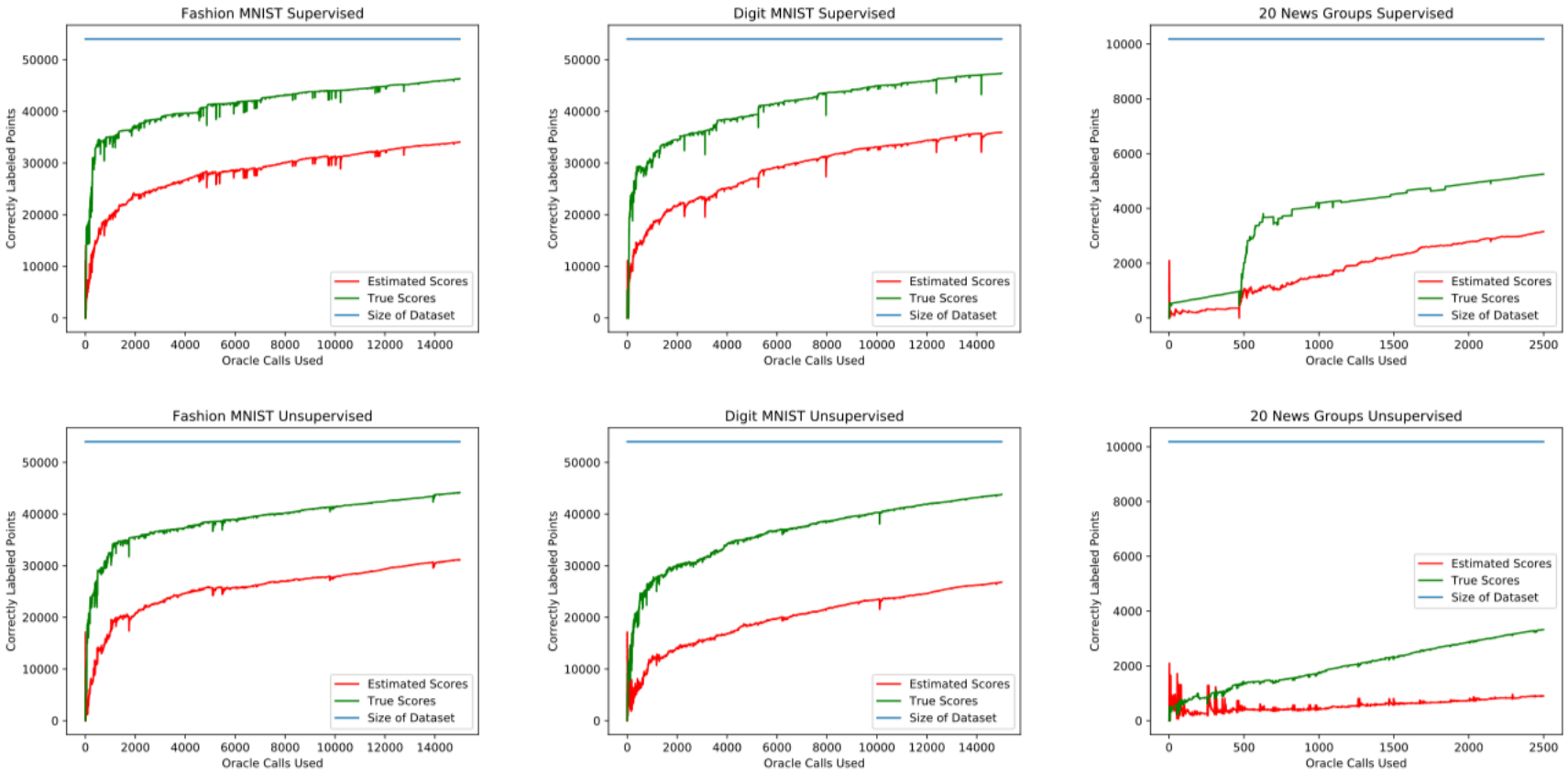

Figure 1: The number of correctly labeled points by each approach applied to the three datasets. GSAL (top row) uses label information along the way to refine its hierarchy. HSAL (bottom row) uses a fixed hierarchy, which avoids bias but is restricted to an initial hierarchical clustering. On the 20 Newsgroups data, GSAL (top right) makes a significant adjustment after receiving enough labels to infer a data split better aligned to the class labels than the clustering used by HSAL (bottom right).

Experiment

Table 2: GSAL (our method) compared to HSAL using a budget of 1,500 oracle queries. These results are averaged over three trials.

Method	Dataset	Size of Y	Label Accuracy	Model Accuracy
GSAL	Fashion-MNIST	$23{,}376 \pm 1{,}345$ (43%)	0.88 ± 0.00	0.69 ± 0.03
HSAL	Fashion-MNIST	$16{,}244 \pm 3{,}426$ (30%)	0.89 ± 0.03	0.65 ± 0.04
GSAL	MNIST (Digit)	$20{,}761 \pm 4{,}578$ (38%)	0.88 ± 0.00	0.74 ± 0.03
HSAL	MNIST (Digit)	$11{,}347 \pm 3{,}051$ (21%)	0.83 ± 0.01	0.66 ± 0.02
GSAL	20 Newsgroups	$1{,}623 \pm 130$ (15%)	0.98 ± 0.03	0.60 ± 0.03
HSAL	20 Newsgroups	$1{,}584 \pm 34$ (15%)	0.96 ± 0.02	0.60 ± 0.01

Thanks