Hierarchical Active Learning with Group Proportion Feedback
Zhispeng Luo and Milos Hauskrecht
Department of Computer Science, University of Pittsburgh, USA

ABSTRACT

- CHALLENGE: Learning of classification models from data instances that are not initially labeled often requires non-trivial human annotation.
- OUR WORK: Actively learn classification models (instance-based) from groups (sets of instances) and their proportion labels.

GOAL: Learn an instance-based classification model \( P(y|x, \Theta) \) from labeled groups.

The labeled groups (say there are \( N \)) can be notated as \( \{G_i\}_{i=1}^N \), where:
- Each group \( G_i \) has \( n_i \) instances: \( \{x_{ij}\}_{j=1}^{n_i} \).
- The label of \( G_i \) is a positive class proportion label \( \mu_i \in [0,1] \) in binary setting.
- Thus, \( G_i = (\{x_{ij}\}_{j=1}^{n_i}, \mu_i) \).

There are two categories of learning algorithms developed in previous work:
- Use the proportion labels to estimate the sufficient statistics in the likelihood function of data instances [3,8].
- Train instance-based models such that they generate instance labels that are consistent with the proportion labels in the corresponding groups [7,10].

Our algorithm: sample sufficiently many instance-label pairs in labeled groups and feed them to any standard instance-level learning algorithm.

- Here we learn a probabilistic model by using Maximum Likelihood Estimation.
- Create a bootstrap sample \( S = \{ (x_{ij}, y_{ij}) \}_{j=1}^{n_i} \) from the \( N \) labeled groups \( \{G_i\}_{i=1}^N \); \( x_{ij} \) is uniformly sampled with replacement from all the instances in \( G_i \); \( y_{ij} \) is sampled by a Bernoulli process with \( p = \mu_i \), the label of \( G_i \) that \( x_{ij} \) is in.
- The model parameter of \( P(y|x, \Theta) \) is then estimated from \( S \) as \( \Theta \) through MLE.
- Note that \( \Theta \) is random, but it asymptotically follows a Normal \( N(\Theta, \Sigma) \) [6]
- \( \Theta = b(\hat{\Theta}) \) is the converged parameter when \( K \to \infty \).
- \( \Sigma \) is the variance of \( \hat{\Theta} \) determined by a finite \( K \). It is approximated by \( \hat{\Sigma} \).
- Thus, \( \hat{\Theta} \) asymptotically approximately follows a Normal \( N(\hat{\Theta}, \hat{\Sigma}) \).

ACTIVE GROUP SELECTION

ACTIVE STRATEGY: Maximal model change principle [9,2].
At each step time \( t \), split the group \( G_i \) from current fringe \( F(t) \) such that if its children were labeled they would lead to the greatest change to current model \( \Theta(t) \).

The INTUITION behind this strategy:
- Large groups should be split first as they represent a broader input feature space.
- Impure groups (in terms of class proportions) should be prioritized as well.
- The refinement of the above two types groups offers more labeling information.
- And thus they give rise to faster model change rate and model convergence.

How to calculate the Model Change value for each candidate group \( G_i \):
- They key idea is to predict how much the model \( \Theta(t) \) will be updated if \( G_i \) is split.
- Apparently we do not know the labels of \( G_i \)'s children, but we can infer them.
- With the inferred labels of \( G_i \)'s children, we can calculate the model change as follows:
  - Create a temporary fringe \( F(t+1) = (F(t) - G_i) \cup (G_i \text{'s children}) \).
  - Learn a new model \( \Theta(t+1) \) following \( N(\Theta(t+1); \Sigma) \).
  - Use KL-Divergence to compare the model distribution change.
  - That is: \( \text{Model Change}(G_i) = D_{KL}(\hat{\Theta(t+1)} ; \hat{\Theta(t)})) \).
  - Finally, select the group \( G_i \) with the largest change to split.

ACTIVE LEARNING FRAMEWORK - HALG

1. Form a hierarchy \( \mathcal{T} \) of unlabeled groups from an unlabeled data pool \( U \).
2. Maintain a fringe (a complete partition of \( \mathcal{T} \)): \( F(t) \) = (the labeled root of \( \mathcal{T} \)).
3. Set step time \( t \to t+1 \); Instantiate an instance-level classifier with parameters \( \Theta(t) \);
4. Repeat:
   - Actively select a group \( G_i \), \( E(t+1) \) based on \( \Theta(t) \);
   - Split \( G_i \), and query the proportion labels of \( G_i \)'s child groups;
   - Replace \( G_i \) with its children in the fringe;
   - \( F(t+1) = (F(t) - G_i) \cup (G_i \text{'s children}) \).
   - \( \Delta t = \) the number of \( G_i \)'s children.
   - Retrain the classifier with groups in \( F(t+1) \), and update \( t \to t + \Delta t \).

LEARNING A MODEL FROM GROUPS

- GROUP: A compact collection of similar instances.
- GROUP DESCRIPTION: Groups are described to annotators via conjunctive patterns over the input space features.
- GROUP LABEL: a proportion of instances in the group falling in one of the classes.
- INSTANCE ANNOTATION: Sometimes easy: images
- MOTIVATION: Electronic Health Records
  - Complex patient-specific information. Search for the right information to classify the EHR instances increases the annotation effort.

GROUP ANNOTATION: Can be easier if groups of data objects are described compactly over fewer dimensions.
- Groups of patients (EHRs) can be described more compactly as:
  - A group of patients: (sex=’male’ & [40<age<50] & [chest pain type=’X’] & [fasting blood sugar within [130,150] mg/dl]) [4]

PROBLEM: groups may include instances with the different labels.
- SOLUTION: annotate each group with a proportion label, reflecting the proportion of instances in the group falling in one of the classes:
  - “60% of patients in Group \( G_i \) suffer from a heart disease.” [4]

GROUP CONCEPT

- ISSUE: Grouped data are often not available.
- SOLUTION: Use hierarchical clustering to form a hierarchy of groups.
- The description of each group in conjunctive patterns can be automatically learned by a standard decision tree [4] algorithm.

HIERARCHICAL GROUP FORMATION

- Unlabeled Data Pool
  - A top-down framework
  - Model Change (A) = 7
  - Model Change (B) = 9
  - Split group \( B \) from \( t = 3 \) to \( t = 5 \)

ACTIVE LEARNING FRAMEWORK - HALG

1. Form a hierarchy \( \mathcal{T} \) of unlabeled groups from an unlabeled data pool \( U \).
2. Maintain a fringe (a complete partition of \( \mathcal{T} \)): \( F(t) \) = (the labeled root of \( \mathcal{T} \)).
3. Set step time \( t \to t+1 \); Instantiate an instance-level classifier with parameters \( \Theta(t) \);
4. Repeat:
   i. Actively select a group \( G_i \), \( E(t+1) \) based on \( \Theta(t) \);
   ii. Split \( G_i \), and query the proportion labels of \( G_i \)'s child groups;
   iii. Replace \( G_i \) with its children in the fringe;
   iv. \( F(t+1) = (F(t) - G_i) \cup (G_i \text{'s children}) \).
   v. Retrain the classifier with groups in \( F(t+1) \), and update \( t \to t + \Delta t \).

EXPERIMENTS

- Our HALG outperforms other active learning methods:
  - DUSL [5], RIQY [4], HS1
- Two strengths of HALG:
  - Initially, learning groups can boost the model performance more than learning with instances.
  - Later, our active learning strategy leads to faster model convergence.

Hierarchical Active Learning with Group Proportion Feedback
Zhispeng Luo and Milos Hauskrecht
Department of Computer Science, University of Pittsburgh, USA

REFERENCES


Acknowledgements

The work presented in this paper was supported in part by grants R01GM088224 and R01LM01019 from the NIH. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

CONTACT INFORMATION

Zhispeng Luo and Milos Hauskrecht
Department of Computer Science, University of Pittsburgh
(pkuo, milos)@cs.pitt.edu
http://people.cs.pitt.edu/~pkuo/
http://people.cs.pitt.edu/~milos/