Hierarchical Active Learning with Group Proportion Feedback

Zhipeng 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*.

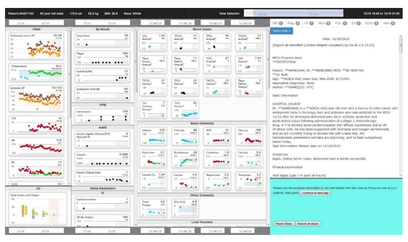
MOTIVATION

- **INSTANCE ANNOTATION**

Sometimes easy: **images**

Dog ?

Sometimes hard: **Electronic Health Records**

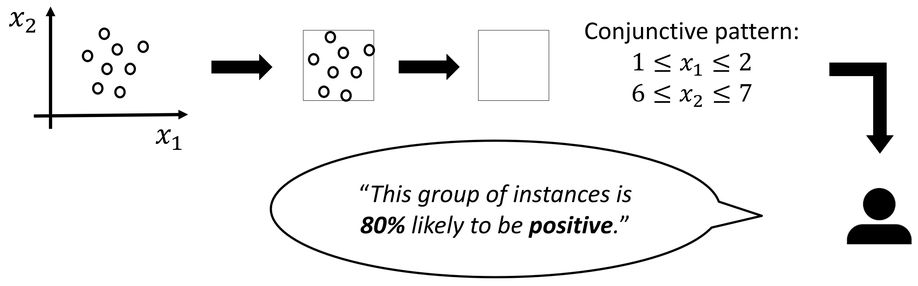


Heart disease ?

- ❖ **Electronic Health Records (EHRs)**
  - 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=female) & (40<age<50) & (chest pain type=3) & (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_1 suffer from a heart disease."* [4]

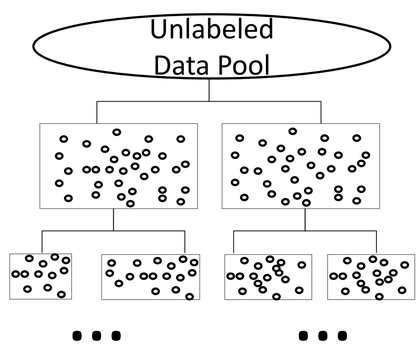
GROUP CONCEPT

- **GROUP**: A group is 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.



HIERARCHICAL GROUP FORMATION

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



ACTIVE LEARNING FRAMEWORK - HALG

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

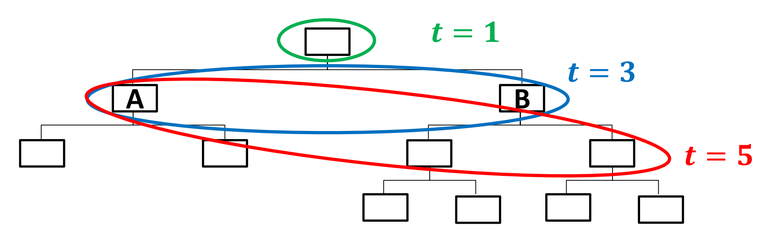
LEARNING A MODEL FROM GROUPS

- **GOAL**: Learn an instance-based classification model $P(y|\boldsymbol{x}, \boldsymbol{\theta})$ 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: $\{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}$
  - ❖ The label of G_i is a *positive* class proportion label $\mu_i \in [0,1]$ in binary setting.
  - ❖ Thus, $G_i = (\{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}, \mu_i)$
- 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].
  - ❖ Limitations: the optimization procedure is slow; restricted to specific models.
- **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 boostrap sample $S = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^K$ from the N labeled groups $\{G_i\}_{i=1}^N$:
    - $\boldsymbol{x}_k$ is uniformly sampled with replacement from all the instances in $\{G_i\}_{i=1}^N$;
    - y_k is sampled by a Bernoulli process with $p = \mu_i$, the label of G_i that $\boldsymbol{x}_k$ is in.
  - ❖ The model parameter of $P(y|\boldsymbol{x}, \boldsymbol{\theta})$ is then estimated from S as $\widehat{\boldsymbol{\theta}}$ through MLE.
    - Note that $\widehat{\boldsymbol{\theta}}$ is random, but it asymptotically follows a Normal $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ [6]
    - $\boldsymbol{\theta} = \mathrm{E}(\widehat{\boldsymbol{\theta}})$ is the converged parameter when $K \to \infty$;
    - $\boldsymbol{\Sigma}$ is the variance of $\widehat{\boldsymbol{\theta}}$ determined by a finite K. It is approximated by S as $\widehat{\boldsymbol{\Sigma}}$.
    - Thus, $\widehat{\boldsymbol{\theta}}$ asymptotically approximately follows a Normal $\mathcal{N}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}})$.

ACTIVE GROUP SELECTION

- **ACTIVE STRATEGY**: *Maximal model change* principle [9,2].
- At each step time t, split the group G_* from current fringe $F^{(t)}$ such that if its children were labeled they would lead to *the greatest change* to current model $\widehat{\boldsymbol{\theta}}^{(t)}$.
- 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 $\widehat{\boldsymbol{\theta}}^{(t)}$ 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 change as follows:
    - Create a temporary fringe $F_{[i]}^{(t)} = \{F^{(t)} - G_i\} \cup \{G_i$'s children$\}$;
    - Learn a new model $\widehat{\boldsymbol{\theta}}_{[i]}^{(t)}$ following $\mathcal{N}(\widehat{\boldsymbol{\theta}}_{[i]}^{(t)}, \widehat{\boldsymbol{\Sigma}}_{[i]}^{(t)})$;
    - Use KL-Divergence to compare the model distribution change.
    - That is: $Model\ Change(G_i) = D_{KL}(\mathcal{N}\left(\widehat{\boldsymbol{\theta}}_{[i]}^{(t)}, \widehat{\boldsymbol{\Sigma}}_{[i]}^{(t)}\right) \parallel \mathcal{N}(\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\Sigma}}^{(t)}))$
  - ❖ Finally, select the group G_* with the largest change to split.

A top-down framework

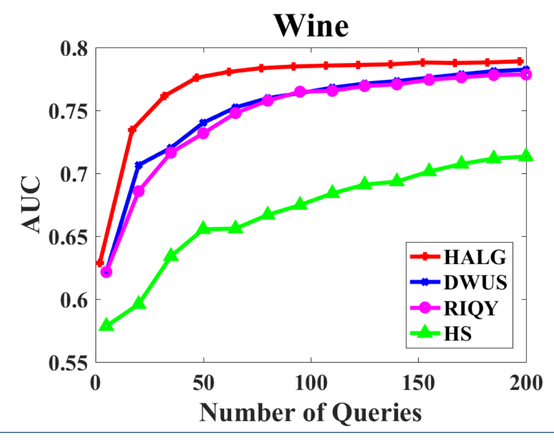


$Model\ Change\ (\boldsymbol{A}) = 7$
$Model\ Change\ (\boldsymbol{B}) = 9$

Split group $\boldsymbol{B}$ from $t = 3$ to $t = 5$

EXPERIMENTS

- Our HALG outperforms other active learning methods:
  - ❖ DWUS [5], RIQY [4], HS[1]
- 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*.



Contact Information

Zhipeng Luo and **Milos Hauskrecht**
Department of Computer Science
University of Pittsburgh
{zpluo, milos}@cs.pitt.edu
http://people.cs.pitt.edu/~zpluo/
http://people.cs.pitt.edu/~milos/

References

1. S. Dasgupta and D. Hsu. **Hierarchical sampling for active learning**. In Proceedings of the 25th international conference on Machine learning, 2008.
2. A. Freytag, et al. **Selecting influential examples: Active learning with expected model output changes**. In European Conference on Computer Vision, 2014.
3. N. Quadrianto, et al. **Estimating labels from label proportions**. Journal of Machine Learning Research, 2009.
4. P. Rashidi and D. Cook. **Ask me better questions: active learning queries based on rule induction**. In Proceedings of the 17th ACM SIGKDD, 2011.
5. B. Settles. **Active learning**. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2012.
6. Bickel, Peter J., and Kjell A. Doksum. **Mathematical statistics: basic ideas and selected topics**, volume I. Vol. 117. CRC Press, 2015.
7. Kuck, Hendrik, and Nando de Freitas. **Learning about individuals from group statistics.** arXiv preprint arXiv:1207.1393, 2012.
8. Patrini, Giorgio, et al. **(Almost) no label no cry.** Advances in Neural Information Processing Systems, 2014.
9. Roy, Nicholas, and A. McCallum. **Toward optimal active learning through monte carlo estimation of error reduction.** ICML, 2001.
10. Yu, Felix X., et al. ∝ **SVM for learning with label proportions**. arXiv preprint arXiv:1306.0886, 2013.

Acknowledgements

The work presented in this paper was supported in part by grants R01GM088224 and R01LM010019 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.