Hierarchical Active Learning with Group Proportion Feedback

Zhipeng Luo and Milos Hauskrecht
Department of Computer Science
University of Pittsburgh, USA
Motivation: Data Annotation Cost

Learning of classification models relies on labeled data instances

Problems:
- Large number of labeled instances may be needed for learning
- The actual cost or hardness of labeling of data instances may differ:
Objective

• How to reduce the annotation cost? How to learn an instance-level classification model, $f : x \rightarrow y$, more efficiently with a smaller number of queries?

• **Solution studied in this work:**
  
  Group Annotation + Active Learning with Groups

• **Why the groups and queries on groups:**
  
  • Many similar instances can be ‘labeled’ together as groups
  
  • Groups often have a more compact description that is easier to understand for a user, further reducing the annotation cost
The Group Concept

- **Group**: represents a (sub)set of instances $x$ in the data
The Group Concept

- **Group**: represents a (sub)set of instances \( x \) in the data
- **Region of the input space \( x \)**: defines a subpopulation of \( x \)
  - A group may be covered by a region
The Group Concept

- **Group**: represents a (sub)set of instances $x$ in the data

- **Region of the input space $x$**: defines a subpopulation of $x$
  - A group may be covered by a region

- **A region/group can be assessed with respect to target $y$ by using a proportion label**
  - A proportion of instances in the group that belong to one of the classes

60% of red (positives)
Active learning with groups algorithm: idea

Hierarchical Active Learning with Groups (HALG)
- Makes queries that assess proportion labels for the groups
- Learns instance-level classifier model from the groups and their labels
- Actively chooses queries to gradually improve the model

Basic questions:
- How to form the groups?
- How to describe the groups to the user/annotator?
- How to learn an instance level classifier from group-proportion feedback?
- How to pick the groups to be labeled next?
How to form the groups?

- **Hierarchical Group Formation**: based on hierarchical clustering
How to form the groups?
How to form the groups?

- Hierarchical Group Formation: based on hierarchical clustering
How to describe the groups to the user?

- Fit groups of instances obtained via hierarchical clustering to ‘rectangular’ input space regions formed by conjunctive patterns.

Conjunctive pattern:
\[ 1 \leq x_1 \leq 2 \]
\[ 6 \leq x_2 \leq 7 \]

“This group of instances is **80% likely to be positive**.”
“What proportion of patients with:

• sex=female,
• 40<age<50,
• chest pain type=3, and
• fasting blood sugar within [130,150] mg/dL

suffer from a heart disease?”

“60% of patients in this group suffer from a heart disease.”

Learning a Model from Labeled Groups

• **Existing algorithms** for learning classifiers from proportion labels:
  • Assume the groups and proportions are known apriori
  • **Algorithms**: either (1) define a loss based on the fit of the groups to the discriminative projection or (2) rely on sample-based algorithms

• We use a **sample-based** algorithm:
  • Sample instance labels according to their group proportion labels;
  • Feed them to any instance-based learning algorithms.
  • E.g. MLE estimates of the parameters $\hat{\theta}$ defining $P(y|x, \theta)$.

How to select the group to be labeled next?

- **Top-bottom group labelling process:** assigns proportion label to rectangular regions fitted to original clusters induced by the hierarchical clustering algorithm from more general to more specific groups.

```
     "65%+" (prior)
        /   \
   "40%+"   "85%+
  /     \  /     \
"15%+" "60%+" "95%+" "50%+
```

......

......
Active Group Selection

• How to select the group to query next?
  • Query A’s or B’s children next?
• Intuition: split the parent group that is the most influential for our model:
  • Impure: with an uncertain label;
  • Large: affects many instances.
• Solution:
  • Split the group which can potentially lead to the maximum model change to the current classifier.
Active Group Selection: Maximum Model Change

**Key idea:** estimate the model change after the group split and annotation

- Assume $\hat{\theta}_{\{A, B\}}$ are the parameters of the current model, trained with groups $A$ and $B$,
- For each group, $A$ or $B$ (say $A$ here):
  - Infer the labels of $A$’s children $\{A_1, A_2\}$;
  - Re-train the model with $A$’s children instead, to get $\hat{\theta}_{\{A_1, A_2, B\}}$;
  - See how much the model has changed by splitting $A$: $d(\hat{\theta}_{\{A, B\}}, \hat{\theta}_{\{A_1, A_2, B\}})$
- Finally, split the group and query its children that comes with the largest estimated change.
Experiments and Observations

- Our method **HALG** outperforms other active learning methods.
  - **Initially**: learning with groups is superior to learning with instances
  - **In the long run**: maximum model change principle leads to faster model convergence

Conclusions and Future Work

• We have developed an annotation cost-effective framework for learning instance-level classifiers, effective to situations where instance-labeling is expensive but group-labeling is more feasible and efficient.

• **Future work**
  • Dynamic and supervised clustering: check out our new paper @ ECML 2018
  • Handling of high dimensional inputs
  • Theoretic analysis of the sample complexity of learning from groups

Hierarchical Active Learning with Group Proportion Feedback

Thank you!