Interactively Guiding Semi-Supervised Clustering via Attribute-based Explanations

Shrenik Lad and Devi Parikh
Virginia Tech

**MOTIVATION**

Unsupervised Clustering

Semi-Supervised Clustering (SSC)

Semi-Supervised Approaches
- Use pairwise constraints to guide clustering

Problems
- Thousands of constraints required for decent clustering
- Lot of redundancy between constraints

**OUR IDEA**

Semi-Supervised Clustering (SSC)

Active Clustering
- Solicit constraints on actively chosen pairs

Problems
- Each constraint is still a weak indication of desired clustering
- Specific to clustering algorithm

**ATTRIBUTES**

- Mid-level concepts like furry, shiny, natural, etc

**RESULTS**

Conclusions
- Our approach outperforms several strong baselines
- Actively selecting pairs performs better than randomly chosen pairs
- Using attribute-based explanations, clustering accuracy increases by 15-20% in 100 iterations
- Our approach is effective wrt human effort and time

- Both binary and relative attributes improve clustering performance

**APPRAOCH**

- Instead of simply saying must-link or cannot-link, also provide an attribute-based explanation
- Vocabulary of attributes is pre-defined
- Works with both binary as well as relative attributes

- Machine propagates human explanation to many unlabeled pairs
- Attributes are predicted automatically

**EXPERIMENTAL SETUP**

SSC Algorithms
1) Constrained K-Means
2) MPCK-Means
3) Spectral Clustering

Datasets
3 domains:
(N: # images, K: # clusters)
1) Scenes (Patterson et al.)
- SUN600: N = 600, K = 6
- 102 Binary Attributes
2) Shoes (Berg et al.)
- Shoes1000: N = 1000, K = 4
- 10 Relative Attributes
3) Faces (Kumar et al.)
- PubFig-Personalized: N = 570, K = 4
- 73 Binary Attributes

Experiment on MTurk
- We showed the desired clustering to workers using some image samples
- Worker is shown a pair of images and asked whether the two must-link or cannot-link
- Worker selects attribute(s) to explain the ML/CL label
- Avg time per question (with attributes): 17 sec
- Avg time per question (without attributes): 5 sec

Personalized Clustering
- We also experiment with user-preference based clustering
- Hand create two additional ground truth (GT) clusterings: Shoes-Personalized and PubFig-Personalized
- Results show our approach is amenable to personalized clustering

**APPROACH**

Which image pair to query the user?
- Use Entropy Minimization approach
- Query the pair whose label is most uncertain
- Let d_ij be the distance between image i and image j

\[ P_{ML} = P(\text{label} = \text{ML} | d_{ij}) \]
\[ P_{CL} = P(\text{label} = \text{CL} | d_{ij}) \]

\[ P(\text{label} = \text{ML} | d_{ij}) \propto P(d_{ij} | \text{label} = \text{ML}) \times P(\text{label} = \text{ML}) \]
\[ P(\text{label} = \text{CL} | d_{ij}) \propto P(d_{ij} | \text{label} = \text{CL}) \times P(\text{label} = \text{CL}) \]

Our actively chosen pair informative to the machine 40% of the times as compared to random pair, which is informative 25% of the times.

**BINARY**

- Attributes are semantic, and hence human understandable
- Attributes are visual and mid-level, and hence machine understandable
- We use attributes to guide a Semi-Supervised Clustering algorithm towards desired clustering

**RELATIVE**

(Software Ordering wrt Heel Height)