Active GNN reading notes

S2: An efficient graph based active learning algorithm with application to nonparametric classification

- Problem setup
  - active learning for binary label prediction on a graph
  - nonparametric active learning, S2 sequentially select vertices to be labeled
    - *cut-set*: $C = \{\{x, y\} \in E : f(x) \neq f(y)\}$
    - *boundary*: $\partial C = \{x \in V : \exists e \in C \text{ with } x \in e\}$ goal is to identify ∂C
    - algorithm assume a noiseless oracle that return label of a multiset of vertices, noisy oracle version algorithms can be transferred from noiseless
    - can be extended to multi-class
- Datasets
  - Digits:
    - Cedar Buffalo binary digits database
    - construct symmetrized 10-nearest-neighbor graph
  - Congressional Voting Records (CVR):
    - 380 vertices, boundary size of 234
  - Grid:
    - synthetic example of a 15x15, positive core in the center
- Methods
  - S2: Shortest Shortest Path

Algorithm 1 S^2: Shortest Shortest Path

Input Graph $G = (V, E)$, BUDGET $\leq n$

$L \leftarrow \emptyset$
while 1 **do**
  $x \leftarrow$ Randomly chosen unlabeled vertex
  do
    Add $(x, f(x))$ to L
    Remove from G all edges whose two ends have different labels.
    if $|L|$ = BUDGET **then**
      Return LABELCOMPLETION(G, L)
    end if
  while $x \leftarrow$ MSSP(G, L) exists
end while

    - LABELCOMPLETION: Any off-the-shelf graph prediction algorithms

        - **MSSP**: return midpoint on the shortest among all the shortest-paths that connect oppositely labeled vertices in L

Sub-routine 2 MSSP

Input Graph $G = (V, E)$, $L \subseteq V$

1: **for** each $v_i, v_j \in L$ such that $f(v_i) \neq f(v_j)$ **do**
2: $P_{ij} \leftarrow$ shortest path between v_i and v_j in G
3: $\ell_{ij} \leftarrow$ length of P_{ij} (∞ if no path exists)
4: **end for**
5: $(i^*, j^*) \leftarrow \arg\min_{v_i, v_j \in L: f(v_i) \neq f(v_j)} \ell_{ij}$
6: **if** (i^*, j^*) exists **then**
7: **Return** mid-point of $P_{i^* j^*}$ (break ties arbitrarily).
8: **else**
9: **Return** $\emptyset$
10: **end if**

    - Can be seen as: random sampling + aggressive search
        - aggressive search: like binary search to find the cut-edge, then unzip the cut-edge
- Baselines
    - measure query complexity
    - AFS On the complexity of finding an unknown cut via vertex queries
    - ZLG Combining active learning and semi- supervised learning using Gaussian fields and harmonic functions
    - BND Towards active learning on graphs: An error bound minimization approach

---

Active Learning for Networked Data

- Problem setup
    - classifying nodes (labels prediction)
        - node features
        - graph structure
        - features/labels of neighbor nodes
    - collective classification:
        - simultaneously predicting labels of all nodes
    - active learning
        - request labels, with goals of decreasing number of labels needed
        - pool-based setting:
            - initially provided with pool of unlabeled examples
            - each step select batch of instances, remove from pool, add to labeled corpus
        - task:
            - collective classification as base learner

        - train: active learning learn CC, CO
        - test: ICA + CC
- Methods
    1. cluster nodes based on graph structure: modularity clustering
    2. iterate:
        1. re-train CO, CC
        2. score clusters based on CO/CC disagreement, pick top k clusters
        3. label one of unlabeled node from each of the k clusters, remove them from pool
            - the node with greatest disagreement LD between CO, CC, majority is picked
        4. Semi-supervision and Dimensionality reduction
            1. semi-supervised collective classification method, use CO to predict unobserved neighbor
            2. PCA
    - note:
        - CC: $P(Y_i|X_i, \text{aggr}(N_i))$, consider neighbor labels
        - CO: $P(Y_i|X_i)$ local classifier with only node features
- Datasets
    - Cora & CiteSeer
        - citation network
        - ignore directions
        - cleaned up
- Baselines
    1. Semi-supervision and Dimensionality Reduction (Base Learner)
        1. CO
        2. CC
        3. CC+Semi-supervision
        4. CC+Semi-supervision+PCA
    2. ALFNET
        1. Random
        2. Uncertainty sampling
    3. Ablation
        1. disagreement: no cluster structure
        2. clustering: select cluster randomly