An Analysis of Graph Cut Size for Transductive Learning

Steve Hanneke

Machine Learning Department

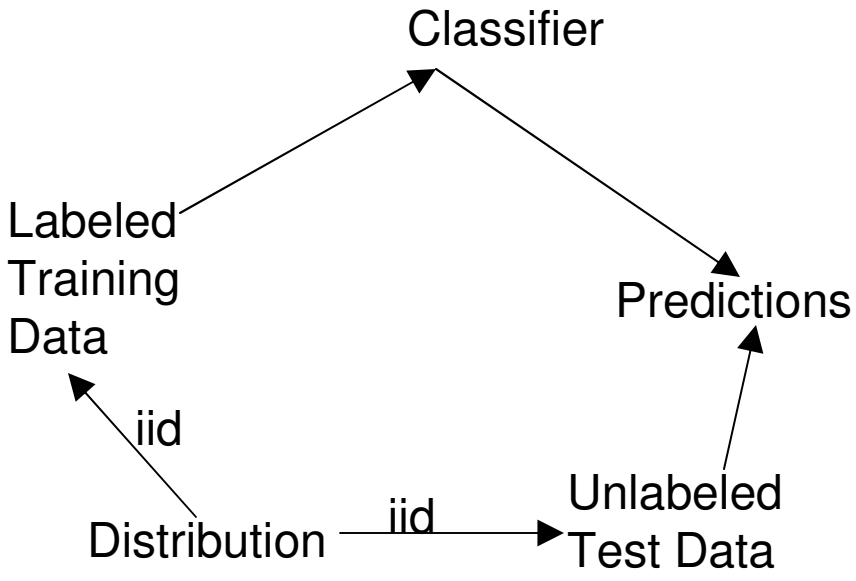
Carnegie Mellon University

Carnegie Mellon

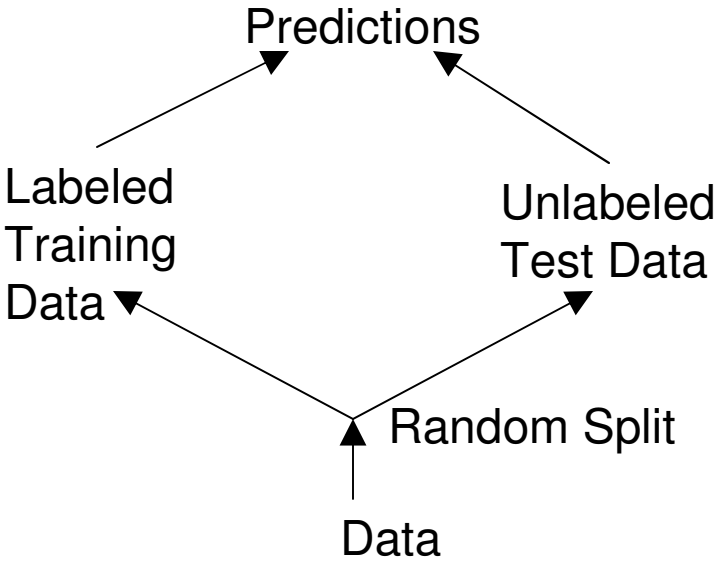
Outline

- Transductive Learning with Graphs
- Error Bounds for Transductive Learning
- Error Bounds Based on Cut Size

Transductive Learning



Inductive Learning

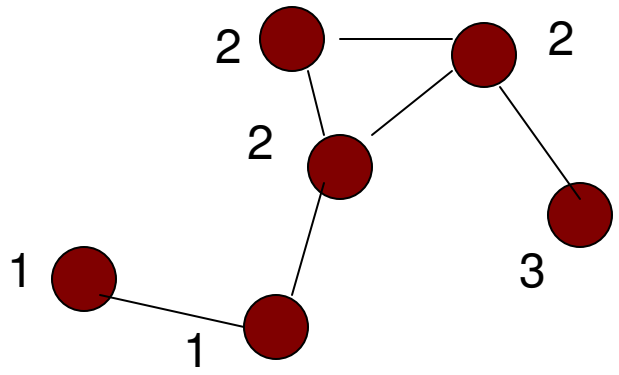


Transductive Learning

Vertex Labeling in Graphs

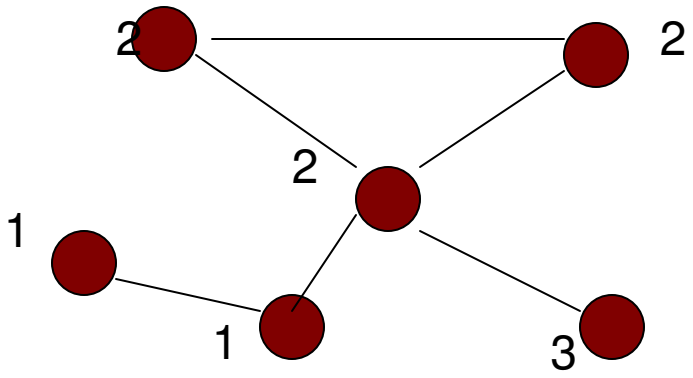
- G=(V,E) connected unweighted undirected graph. |V|=n. (see the paper for weighted graphs).
- Each vertex is assigned to exactly one of k classes {1,2,…,k} (target labels).
- The labels of some (random) subset of n_ℓ vertices are revealed to us. (training set)
- Task: Label the remaining (test) vertices to (mostly) agree with the target labels.

Example: Data with Similarity



- Vertices are examples in an instance space and edges exist between similar examples.
- Several clustering algorithms use this representation.
- Useful for digit recognition, document classification, several UCI datasets,…

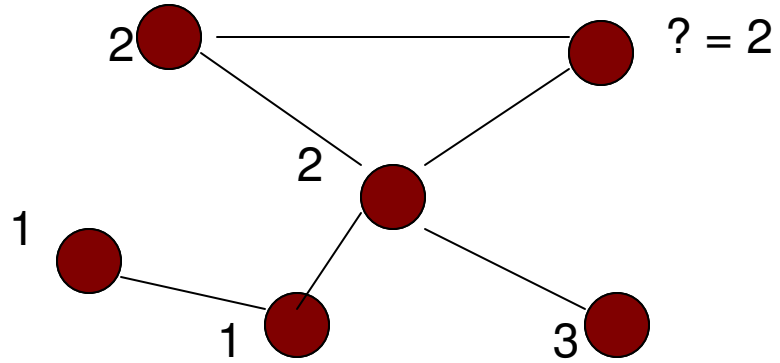
Example: Social Networks



- Vertices are high school students, edges represent friendship, labels represent which after-school activity the student participates in (1=football, 2=band, 3=math club, …).

Adjacency

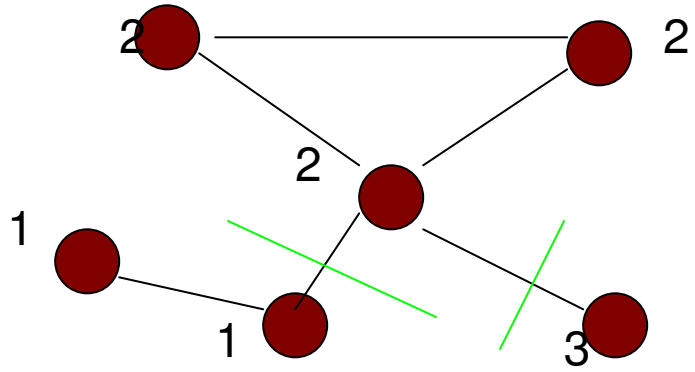
- Observation: Friends tend to be in the same after-school activities.



- More generally, it is often reasonable to believe adjacent vertices are usually classified the same.
- This leads naturally to a learning bias.

Cut Size

- For a labeling h of the vertices in G, define the *Cut Size*, denoted c(h), as the number of edges in G s.t. the incident vertices have different labels (according to h).



Example: Cut Size 2

Learning Algorithms

- Several existing transductive algorithms are based on the idea of minimizing cut size in a graph representation of data (in addition to number of training errors, and other factors).
- Mincut (Blum & Chawla, 2001)
- Spectral Graph Transducer (Joachims, 2003)
- Randomized Mincut (Blum et al., 2004)
- others

Mincut (Blum & Chawla, 2001)

- Find a labeling having smallest cut size of all labelings that respect the known labels of the training vertices.
- Can be solved by reduction to multi-terminal minimum cut graph partitioning
- Efficient for k=2.
- Hard for k>2, but have good approximation algorithms

Error Bounds

- For a labeling h, define $\hat{R}(h)$ and $R(h)$ the fractions of training vertices and test vertices h makes mistakes on, respectively. (training & test error)
- We would like a confidence bound of the form

$$\mathcal{P}r\{\forall h,\ R(h) \leq \ \text{Bound}(\hat{R}(h), \delta)\} \geq 1 - \delta$$

Bounding a Single Labeling

- Say a labeling h makes T total mistakes. The number of training mistakes is a hypergeometric random variable.

$$Hyp_T(n_\ell \hat{R}(h)) = \sum_{i=0}^{n_\ell \hat{R}(h)} \frac{\binom{T}{i}\binom{n-T}{n_\ell - i}}{\binom{n}{n_\ell}}$$

- For a given confidence parameter δ, we can "invert" the hypergeometric to get

$$R_{max}(\hat{R}(h), \delta) = \max\left\{ \frac{T - n_\ell \hat{R}(h)}{n - n_\ell} \middle| Hyp_T(n_\ell \hat{R}(h)) \geq \delta, T \in \mathbb{Z} \right\}$$

Bounding a Single Labeling

$$R_{max}(\hat{R}(h), \delta) = \max \left\{ \frac{T - n_\ell \hat{R}(h)}{n - n_\ell} \middle| Hyp_T(n_\ell \hat{R}(h)) \geq \delta, T \in \mathbb{Z} \right\}$$

- Single labeling bound:

$$\forall h, \delta, \; \mathcal{P}r\{R(h) \leq R_{max}(\hat{R}(h), \delta)\} \geq 1 - \delta$$

- We want a bound that holds simultaneously for all h.
- We want it close to the single labeling bound for labelings with small cut size.

The PAC-MDL Perspective

- Single labeling bound:

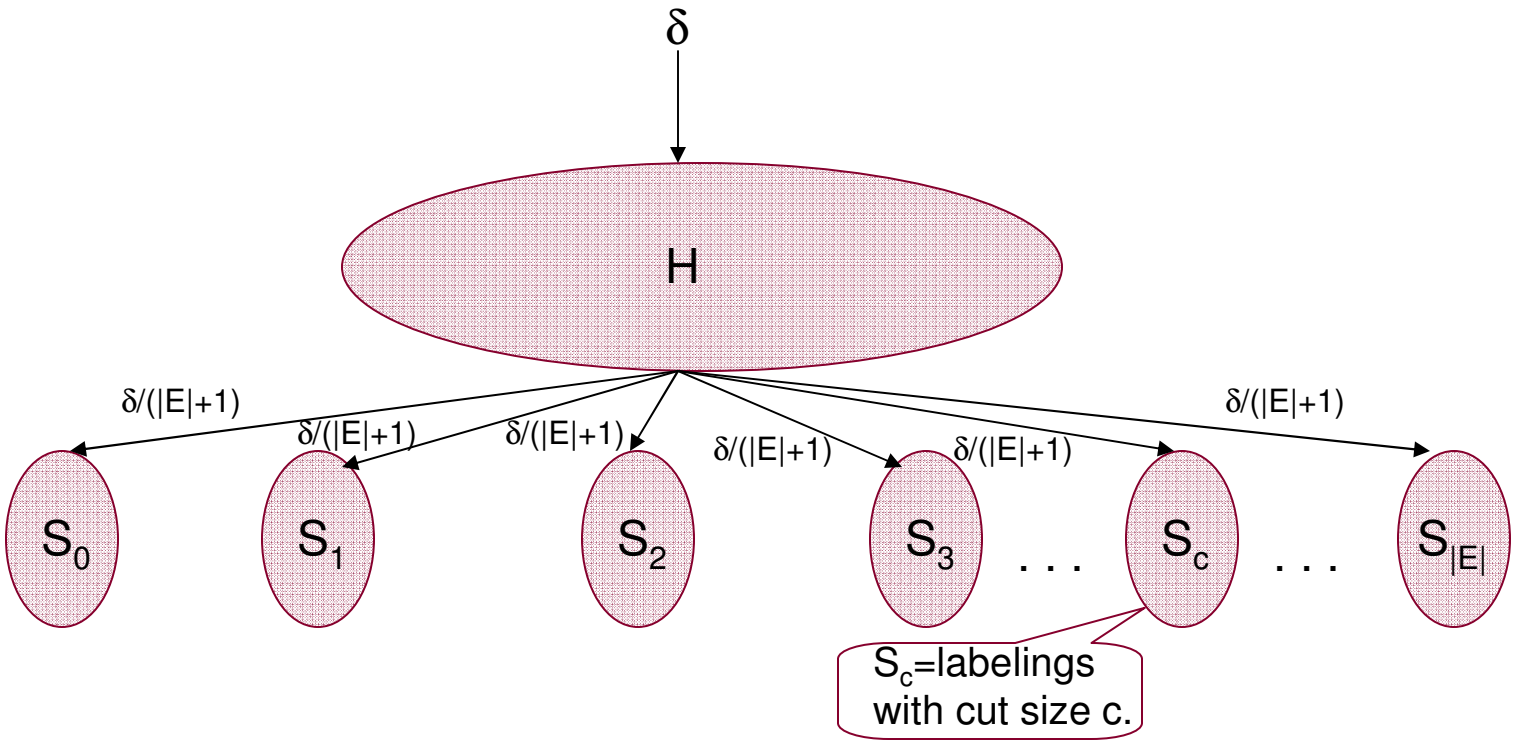
$$\forall h, \delta,\ \mathcal{P}r\{R(h) \leq R_{max}(\hat{R}(h), \delta)\} \geq 1 - \delta$$

- PAC-MDL (Blum & Langford, 2003):

$$\forall \delta,\ \mathcal{P}r\{\forall h, R(h) \leq R_{max}(\hat{R}(h), \delta p(h))\} \geq 1 - \delta$$

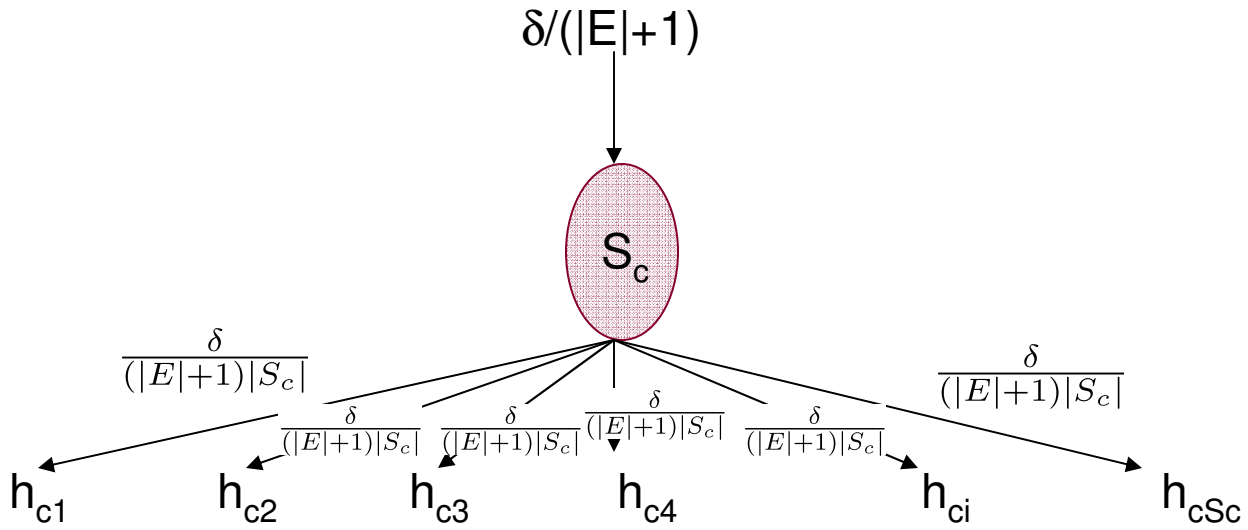
- where p(·) is a probability distribution on labelings. (the proof is basically a union bound)
- Call δp(h) the "tightness" allocated to h.

The Structural Risk Trick



Split the labelings into |E|+1 sets by cut size and allocate δ/(|E|+1) total "tightness" to each set.

The Structural Risk Trick



Within each set S_c, divide the $\delta/(|E|+1)$ tightness equally amongst the labelings. So each labeling receives tightness exactly $\frac{\delta}{(|E|+1)|S_c|}$. This is a valid $\delta p(h)$.

The Structural Risk Trick

- We can immediately plug this tightness into the PAC-MDL bound to get that with probability at least 1-δ, every labeling h satisfies

$$R(h) \leq R_{max}(\hat{R}(h), \frac{\delta}{(|E|+1)|S_{c(h)}|})$$

- This bound is fairly tight for small cut sizes.
- But we can't compute $|S_c|$. We can upper bound $|S_c|$, leading to a new bound that largely preserves the tightness for small cut sizes.

Bounding $|S_c|$

- Not many labelings have small cut size.
- At most n^2 edges, so

$$|S_c| \leq \binom{n^2}{c} k^{2c} \leq (kn)^{2c}$$

- But we can improve this with data-dependent quantities.

Minimum k-Cut Size

- Define *minimum k-cut size*, denoted C(G), as minimum number of edges whose removal separates G into at least k disjoint components.
- For a labeling h, with c=c(h), define the *relative* cut size of h

$$\rho(c) = \frac{c}{C(G)}$$

A Tighter Bound on $|S_c|$

- Lemma: For any non-negative integer c, $|S_c| \leq B(\rho(c))$, where for $\frac{1}{2} \leq \rho < n/(2k)$,

$$
\begin{aligned}
B(\rho) &= k^{\lfloor 2k\rho \rfloor} \binom{n}{2(k-1)\rho} \binom{\lfloor 2k\rho \rfloor}{2(k-1)\rho}^{-1} \\
&\leq \left(\frac{kne}{2(k-1)\rho} \right)^{2(k-1)\rho}
\end{aligned}
$$

- (see paper for the proof)
- This is roughly like $(kn)^{\rho(c)}$ instead of $(kn)^c$.

Error Bounds

- $|S_c| \leq B(\rho(c))$, so the "tightness" we allocate to any h with c(h)=c is at least

$$\frac{\delta}{(|E|+1)|S_c|} \geq \frac{\delta}{(|E|+1)B(\rho(c))}$$

- **Theorem 1** (main result): With probability at least 1-δ, every labeling h satisfies

$$R(h) \quad \leq \quad R_{max}\left(\hat{R}(h), \frac{\delta}{(|E|+1)B(\rho(c(h)))}\right)$$

(can be slightly improved: see the paper)

Error Bounds

- **Theorem 2**: With probability at least 1-δ, every h with ½ <ρ(h)<n/(2k) satisfies

$$R(h) \leq R_{max}\left(\hat{R}(h), \frac{\delta B(\rho(h))^{-1}}{|E|+1}\right)$$

$$\leq \hat{R}(h) + \sqrt{\left(\frac{n}{n_u}\right)\left(\frac{n_u+1}{n_u}\right)\left(\frac{2(k-1)\rho(h)\ln\left(\frac{kne}{2(k-1)\rho(h)}\right) + \ln\frac{|E|+1}{\delta}}{2n_\ell}\right)}$$

(overloading ρ(h)=ρ(c(h)))

Something like training error $+\sqrt{\frac{\rho(h)}{n_\ell}}$

Proof uses result by Derbeko, et al.

Visualizing the Bounds

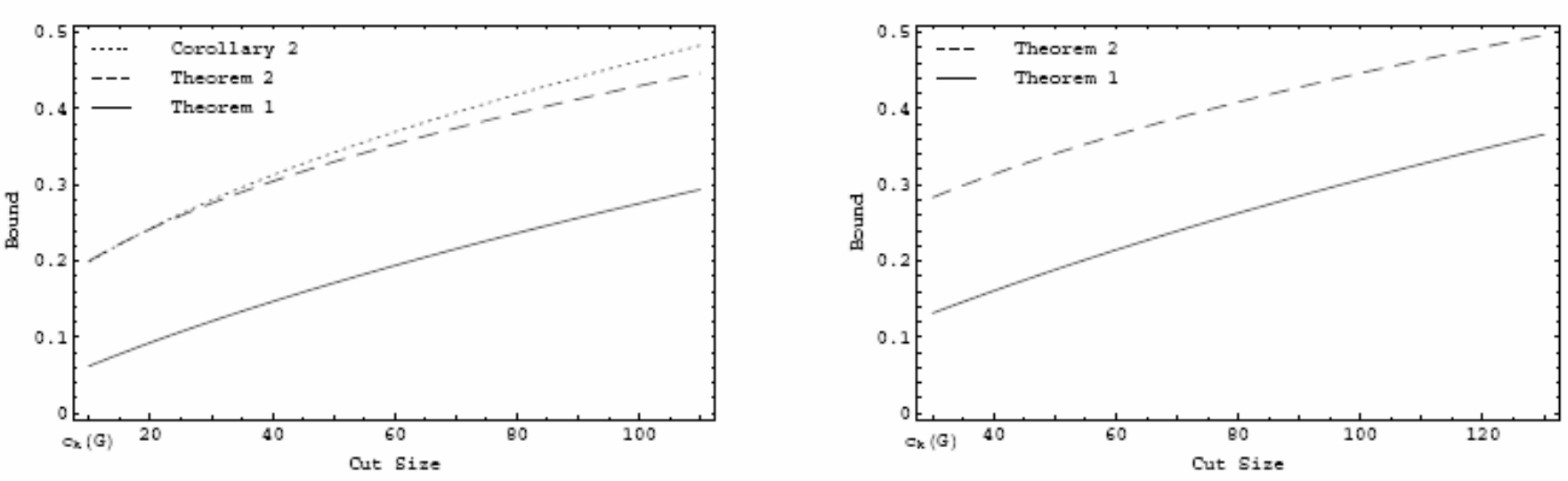


Figure 1. Comparison of cut bounds for $k = 2$ (left) and $k = 4$ (right).

n=10,000; n_ℓ=500; |E|=1,000,000; C(G)=10(k-1); δ=.01; no training errors.

- Overall shapes are the same, so the loose bound can give some intuition.

Conclusions & Open Problems

- This bound is not difficult to compute, it's Free, and gives a nice guarantee for any algorithm that takes a graph representation as input and outputs a labeling of the vertices.
- Can we extend this analysis to include information about class frequencies to specialize the bound for the Spectral Graph Transducer (Joachims, 2003)?

Questions?