Graph-based Semi-Supervised Learning Using Harmonic Functions

Güneş Erkan

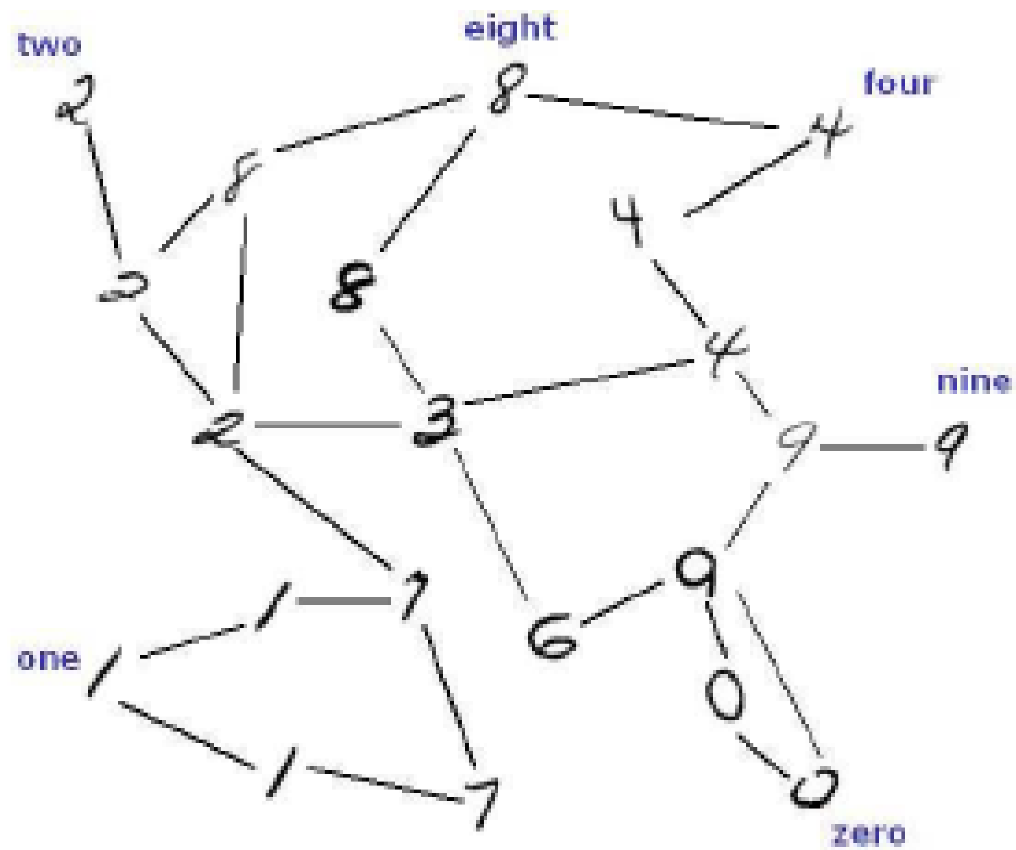
EECS 767 Advanced NLP

Winter 2006

Semi-Supervised Learning

- Combines labeled and unlabeled data.
- Labeling can be extremely costly
  - labeling single example for protein shape classification can take months
- Most semi-supervised learning algorithms exploit the *manifold structure.*
  - assumption: similar examples should belong to the same class.

Example Similarity Graph



Framework

- Edge weight function: $$w_{ij} = \exp\left(-\sum_{d=1}^{m} \frac{(x_{id} - x_{jd})^2}{\sigma_d^2}\right)$$
- Goal is to learn a function: $f : V \longrightarrow \mathbb{R}$
- For labeled examples we know that f(x) = 1 or 0.
  - so we'll *force* them to be that way.
- But how do we assign values to *unlabeled* nodes?

Energy Function

- We want nearby (similar) points (labeled or unlabeled) to have similar labels. So define the *energy function*:

$$E(f) = \frac{1}{2} \sum_{i,j} w_{ij} \left(f(i) - f(j) \right)^2$$

- So we try to minimize this energy function.
- The energy function is *real*. Remember $f : V \longrightarrow \mathbb{R}$

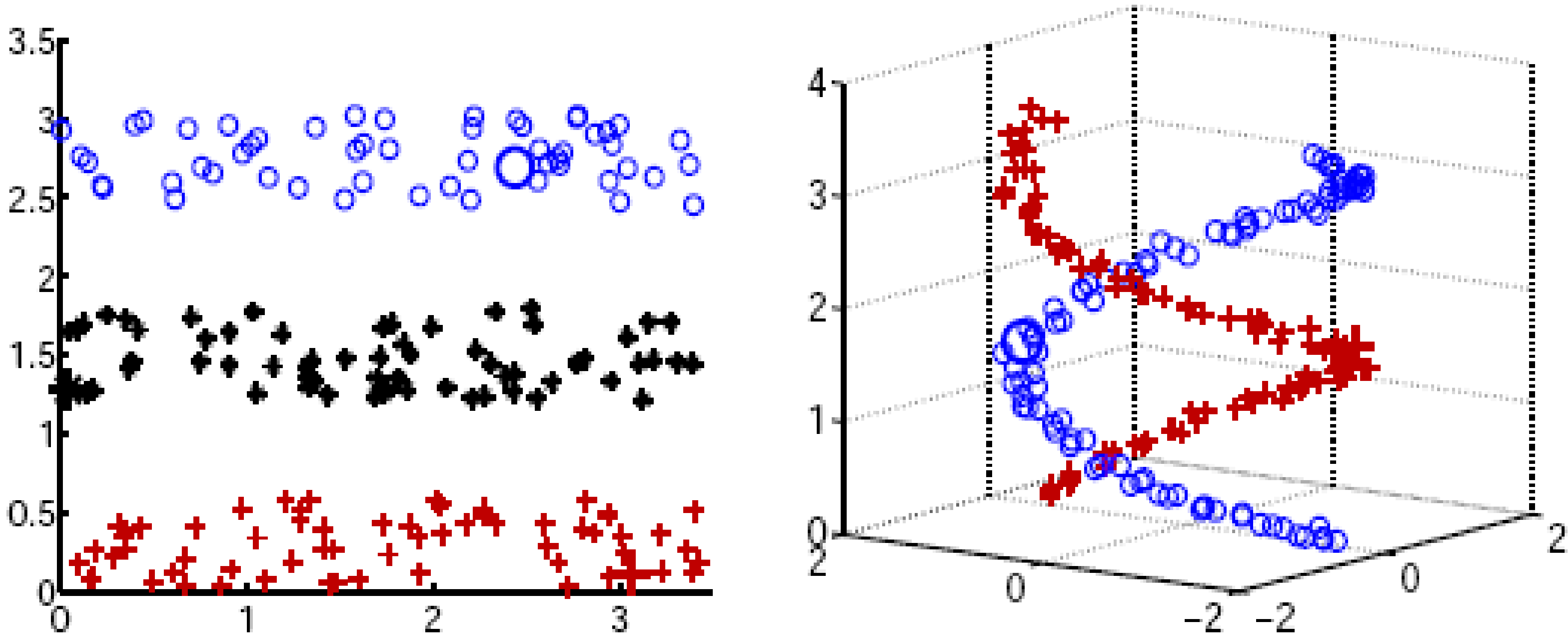
Harmonic Function

- It turns out that the solution to *f* that minimizes the energy function is a *harmonic* function.
- Harmonic property means that

$$f(j) = \frac{1}{d_j} \sum_{i \sim j} w_{ij} f(i), \text{ for } j = l+1, \ldots, l+u$$

- Keep labeled nodes' values either 1 or 0.
- An unlabeled node's value is equal to the average of its neighbors.

Example



Connections / Interpretations

- Random walks:
  - Given a node *i* , start a random walk from that node.
  - Then *f(i)* is the probability that it we will hit a node with label 1 (before a node with label 0).

Connections / Interpretations

- Random walks:
  - Given a node *i* , start a random walk from that node.
  - Then *f(i)* is the probability that it we will hit a node with label 1 (before a node with label 0).
- Electric Networks:
  - Imagine the graph is an electric network with resistors with conductance *w*
  - Connect nodes labeled 1 to a positive voltage source. Connect nodes labeled 0 to the ground.
  - Then *f(i)* is the voltage at each node.

Connections / Interpretations

- Mincuts:
  - Remember the energy function:

$$E(f) = \frac{1}{2}\sum_{i,j} w_{ij}\left(f(i) - f(j)\right)^2$$

Connections / Interpretations

- Mincuts:
  - Remember the energy function:

$$E(f) = \frac{1}{2}\sum_{i,j} w_{ij}\left(f(i) - f(j)\right)^2$$

  - This is mincut's energy function:

$$E_1(f) \;=\; \tfrac{1}{2}\sum_{i,j} w_{ij}\,|f(i) - f(j)|$$

Differences from Mincut

- Mincut's $f : V \rightarrow \{-1, +1\}$
- The harmonic function is a *real relaxation* of it.
- Mincut is not unique, but minimum-energy harmonic function is.
- Mincut is NP-hard, but harmonic function can be efficiently computed.

Incorporating Class Prior Knowledge

- Now we have *f*. How do we do classification?
- When $f(x) > 0.5$, then x is labeled 1.

Incorporating Class Prior Knowledge

- Now we have *f*. How do we do classification?
- When *f(x)* > 0.5, then x is labeled 1 ?
- Look at the *prior* distribution in the labeled data. Or use the domain knowledge.

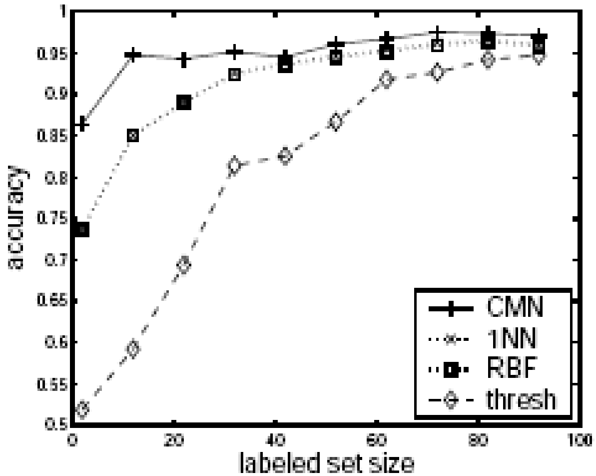
$$q \frac{f_u(i)}{\sum_i f_u(i)} > (1-q) \frac{1 - f_u(i)}{\sum_i (1 - f_u(i))}$$

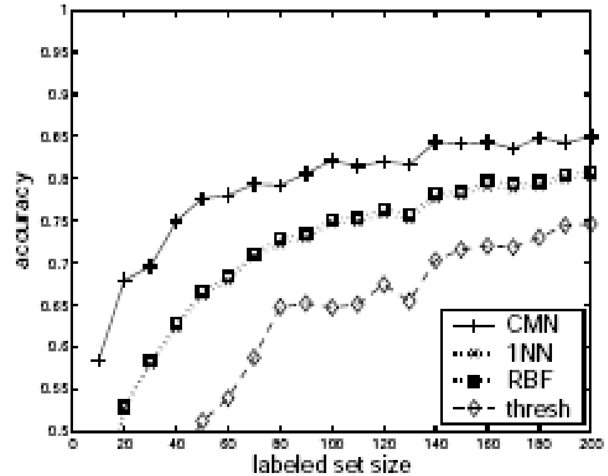
- Remark: no such problem in mincut since it is formulated discrete.

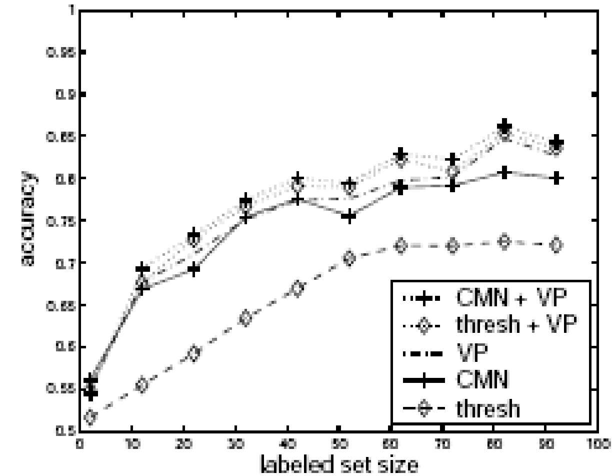
Incorporating External Classifiers

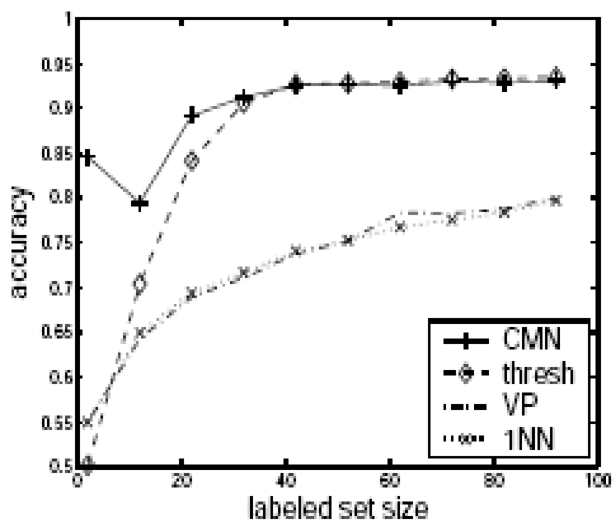
- Suppose you have a baseline classifier, and you want to use it as a starting point for the unlabeled nodes.
- Compute *h,* the baseline function.
- Attach a "dongle" node to each unlabeled node *i* with value *h(i)*.
- Let the transition probability from *i* to its dongle be *d,* and discount all other transitions from *i* by *1-d*.

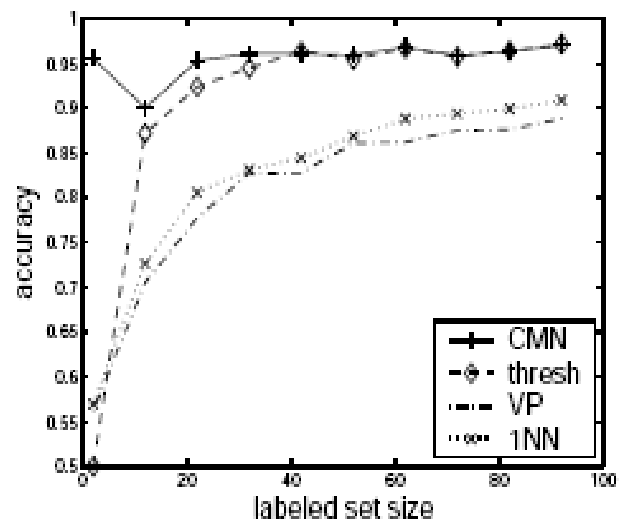
Results

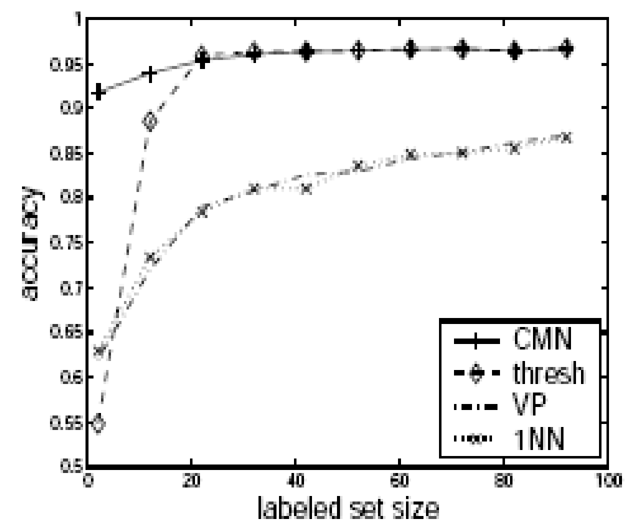

accuracy
labeled set size
CMN
1NN
RBF
thresh


accuracy
labeled set size
CMN
1NN
RBF
thresh


accuracy
labeled set size
CMN + VP
thresh + VP
VP
CMN
thresh


accuracy
labeled set size
CMN
thresh
VP
1NN


accuracy
labeled set size
CMN
thresh
VP
1NN


accuracy
labeled set size
CMN
thresh
VP
1NN