ALADDIN Workshop on Graph Partitioning in Vision and Machine Learning

Jan 9-11, 2003

Welcome!

[Organizers: Avrim Blum, Jon Kleinberg, John Lafferty, Jianbo Shi, Eva Tardos, Ramin Zabih]
Graph partitioning

Coming up recently in

- **Vision** (image segmentation, image cleaning,...)
- **Machine Learning** (learning from labeled & unlabeled data, clustering).

Central problem in **Algorithms** (max-flow min-cut, balanced separators)
Goals of the Workshop

• Exchange of ideas among people in 3 areas.
• Understanding of similarities and differences in problems, objectives.
• Formulate good questions.

This is supposed to be informal!
Thanks to our sponsor

ALADDIN Center

• NSF-supported center on ALgorithms, ADaptation, Dissemination, and INtegration
• Support work/interaction between algorithms and application areas.

More announcements at the break
Graph partitioning for Machine Learning from Labeled and Unlabeled data

Avrim Blum, CMU
Combining Labeled and Unlabeled Data

- Hot topic in recent years. Many applications have lots of unlabeled data, but labeled data is rare or expensive. E.g.,
  - Web page, document classification
  - OCR, Image classification
  - Text extraction

Can we use the unlabeled data to help?

[lots of relevant references omitted here]
How can unlabeled data help?

- Unlabeled data + assumptions reduce space of “reasonable” distinctions.
  - E.g., OCR data might cluster. We hope each digit is one or more clusters.
  - Assumptions about world add a “self-consistency” component to optimization.

- In the presence of other kinds of info, can provide ability to bootstrap (co-training).
  - e.g., video, word-sense disambiguation.
Unlabeled data + assumptions! reasonableness criteria

• Suppose we are looking for a linear separator. We believe should exist one with large separation. SVM.
Unlabeled data + assumptions !
reasonableness criteria

• Suppose we are looking for linear separator. We believe should exist one with large separation. Transductive SVM [Joachims]:
Suppose we are looking for linear separator. We believe should exist one with large separation. Transductive SVM [Joachims]:

Unlabeled data + assumptions! reasonableness criteria
Unlabeled data + assumptions + reasonableness criteria

• Suppose we believe that in general, similar examples have the same label.
  - Suggests NearestNeighbor or locally-weighted voting alg for standard problem.
  - Why not extend to objective function over unlabeled data too?
Unlabeled data + assumptions!
reasonable criteria

• Suppose we believe that in general, similar examples have the same label.
  - Given set of labeled and unlabeled data, classify unlabeled data to minimize penalty = #pairs of similar examples with different labels.
The good, the bad, and the ugly...
The good

Suggests natural alg approach along lines of [GPS,BVZ,SVZ,KT]:

1. Define graph with edges between similar examples (perhaps weighted).
The good

Suggests natural alg approach along lines of \[GPS,BVZ,SVZ,KT\]:

2. Solve for labeling that minimizes weight of bad edges.
The good

Much of ML is just 2-class problems, so (2) becomes just a minimum cut.

S.Chawla will discuss some exptl results and design issues.
The good

Another view: if we created graph by connecting each to nearest neighbor, this is the labeling s.t. NN would have smallest leave-one-out error. [see also Joachims’ talk]
The bad

- Is this really what we want to do?
- Assumptions swamp our evidence?
The bad

• Is this really what we want to do? Assumptions swamp our evidence?
The ugly

1. Who defined “similar” anyway?
2. Given a distance metric, how should we construct graph?
3. Given graph, several possible objectives.

Will skip 1 but see Murphy, Dietterich, Lafferty talks.
2: given \( d \), how to create \( G \)?

- weird issue: for just labeled data, kNN (\( k=1,3,5 \)) makes more sense than fixed radius because of unevenness of distribution. (I.e., for each test point you want to grow radius until hit \( k \) labeled points).
- But for unlabeled data, fixed \( k \) has problems.
Given $d$, how to create $G$?

- Say we connect each example to nearest nbr.
Given $d$, how to create $G$?

- Say we connect each example to nearest nbr.
Given $d$, how to create $G$?

- Say we connect each example to nearest nbr.
Given \( d \), how to create \( G \)?

- Say we connect each example to nearest nbr.

- \( w(u,v) = f(d(u,v)) \) at least has property that graph gets more connected...
Given $d$, how to create $G$?

- [BC]: use unweighted graph. Edge between any pair of distance $< \delta$.

| Dataset | $|L| \times |U|$ | Number of features | MinCut-3 | MinCut-$\delta_{up}$ | MinCut-$\delta_0$ | MinCut-$\delta_{1/2}$ | ID3 | 3-NN |
|---------|-----------------|-------------------|---------|-------------------|-----------------|-----------------------|-----|-----|
| Mush   | 20+1000         | 22                | 82.1    | 97.7              | 97.7            | 97.0                  | 93.3 | 91.1 |
| Mush*  | 20+1000         | 22                | 71.2    | 88.7              | 56.9            | 87.0                  | 80.8 | 83.3 |
| TAK    | 10+100          | 5                 | 86.0    | 99.0              | 96.0            | 97.9                  | 86.0 | 80.0 |
| TAK*   | 10+100          | 5                 | 76.0    | 96.0              | 86.0            | 94.0                  | 76.0 | 62.0 |
| VOTING | 15+300          | 16                | 89.1    | 91.3              | 66.1            | 83.3                  | 86.1 | 89.6 |
| MusK   | 10+200          | 166               | 73.0    | 92.5              | 91.0            | 92.5                  | 83.5 | 87.0 |
| PUMA   | 50+718          | 8                 | 63.8    | 92.3              | 48.8            | 72.3                  | 70.0 | 68.1 |
| IonO   | 50+300          | 34                | 71.0    | 81.6              | 78.0            | 77.6                  | 86.6 | 69.6 |
| Bupa   | 15+300          | 6                 | 53.3    | 59.3              | 48.0            | 41.7                  | 55.7 | 52.7 |
| MI     | 124+132         | 6                 | 70.0    | 61.1              | 61.4            | 61.4                  | 99.4 | 81.1 |
| MII    | 160+132         | 6                 | 68.6    | 67.2              | 57.2            | 67.2                  | 67.9 | 63.6 |
| MIII*  | 122+132         | 6                 | 79.1    | 90.6              | 61.8            | 80.6                  | 94.4 | 83.6 |

- Is there a “correct” way? GW-moats?
Given $G$, several natural objectives

Say $f(v)$= fractional label of $v$.

• Mincut: minimize $\sum_{(u,v) \in E} |f(u) - f(v)|$

• $[GZ]$: minimize $\sum_{(u,v) \in E} (f(u) - f(v))^2$

nice random walk / electrical networks interp.
Given $G$, several natural objectives

- (When) is one better than the other?

- Optimize other fns too?
Given $G$, several natural objectives

• (When) is one better than the other?

• Optimize other fns too?
Given $G$, several natural objectives

- If we view $G$ as MRF, then mincut is finding most likely configuration.
  
  Cut of size $k$ has prob $\propto e^{-k/T}$

- Instead, ask for Bayes-optimal prediction on each individual example?
Given $G$, several natural objectives

- If we view $G$ as MRF, then mincut is finding most likely configuration.

  Cut of size $k$ has prob $\propto e^{-k/T}$

- Instead, ask for Bayes-optimal prediction on each individual example?

- Nice open problem: efficiently sample from this distrib? (extend [JS]?)
Given $G$, several natural objectives

• If we view $G$ as MRF, then mincut is finding most likely configuration.

  Cut of size $k$ has prob $\propto e^{-k/T}$

• Instead, ask for Bayes-optimal prediction on each individual example?

• Hack: Repeatedly add noise to edges and solve.
More questions

• Tradeoff between *assumptions* over unlabeled data, and *evidence* from labeled data? Esp if non-uniform.

• Hypergraphs? find labeling to minimize number of points that are different from majority vote over k nearest nbrs? See [VK,RZ].
More questions

• ... (we’ll see over the next 2.5 days)