The power of negative thinking: Exploiting label disagreement in the min-cut classification framework

with an application to sentiment classification of speeches in legislative debates

Mohit Bansal, IIT Kanpur

Claire Cardie, Cornell University

Lillian Lee, Cornell University

Minimum (or "min") cut framework: elegant technique for using **relationships between items** to aid classification.

Example from Thomas, Pang and Lee '06 (TPL):

Items x_i: speeches uttered during legislative debates.

Labels c_i: item x_i is either "pro" or "con" the debate topic.

Suppose x_i is hard to classify, but...

...from textual clues, x_i's speaker seems to *agree* with x_j's,

...and we can determine c_j

$\Rightarrow$ we should strongly consider setting $c_i := c_j$

[Also: Pang&Lee '04, Agarwal&Bhattacharyya '05, Barzilay&Lapata '05, Greene '07]

Min-cut classification assumes individual preference scores and association preference scores:

Large $\mathrm{Ind}(x, \text{pro}) \Rightarrow$ "x probably belongs in class 'pro' "

Large $\mathrm{Assoc}(x_i, x_j) \Rightarrow$ "probably $c_i = c_j$"

A graphical representation for an example with just two x_is:

$$
\begin{array}{cccccc}
 & \leftarrow [\mathrm{Ind}(x_1, \text{pro}) = .7]- & \boxed{x_1} & & -[\mathrm{Ind}(x_1, \text{con}) = .3]\rightarrow & \\
\boxed{\text{pro}} & & & \Updownarrow [\mathrm{Assoc}(x_1, x_2) = -2] & & \boxed{\text{con}} \\
 & \leftarrow [\mathrm{Ind}(x_2, \text{pro}) = .6]- & \boxed{x_2} & & -[\mathrm{Ind}(x_2, \text{con}) = .4]\rightarrow &
\end{array}
$$

Note: these scores can come from different sources and so here we see that **they can encode conflicting preferences**: x_1 and x_2 both individually prefer “pro”, but also want to be in different classes (negative assoc).

(Binary) min-cut classification = finding c_is that minimize:

$$\sum_i \mathrm{Ind}(x_i, \overline{c_i}) + \alpha \times \sum_{i,j:\, i<j,\, c_i = \overline{c_j}} \mathrm{Assoc}(x_i, x_j),$$

where $\overline{c_i}$ is the "opposite" class from c_i.

Or, in English: minimizing the total "pining" of the x_is for the class they were not assigned to, where that "pining" is due to either individual *or* associational preferences.

If the scores are non-negative, this partitioning problem can be solved efficiently and exactly via max flow!

But if the assocs can be negative ...
(negative individual preferences easily handled)

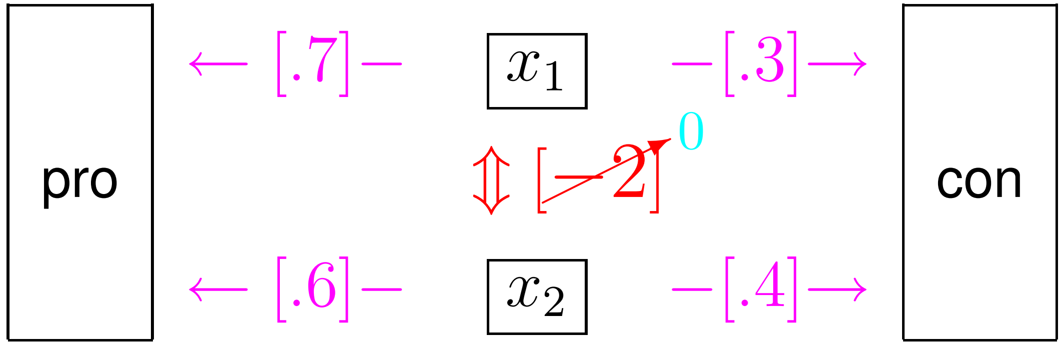
- Important example: SVM-based scoring utilizing signed distance from the separating hyperplane

... then, performance guarantees disappear. (Unless P=NP.)

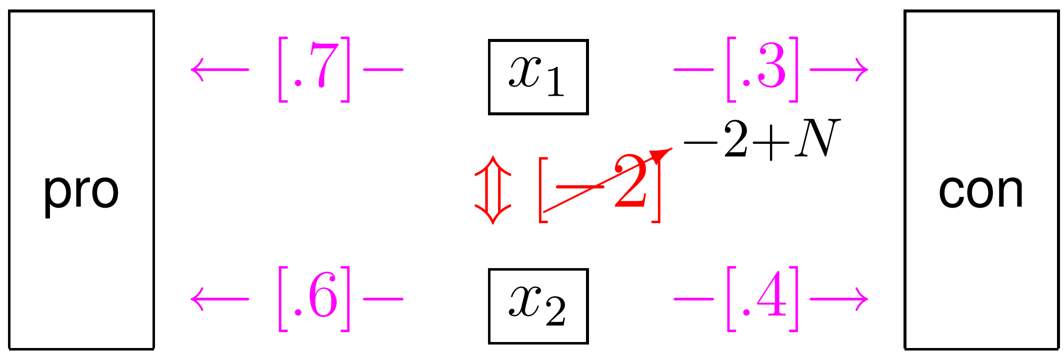
Heuristic-based approach: Change the scores so there aren't any negative ones, making max flow again applicable.

In 2006, TPL simply zeroed negative assocs. The positive assocs alone boosted performance above strong baselines.

TPL:



Scale all up: Instead of ignoring disagreement, add big enough positive N to all assocs.

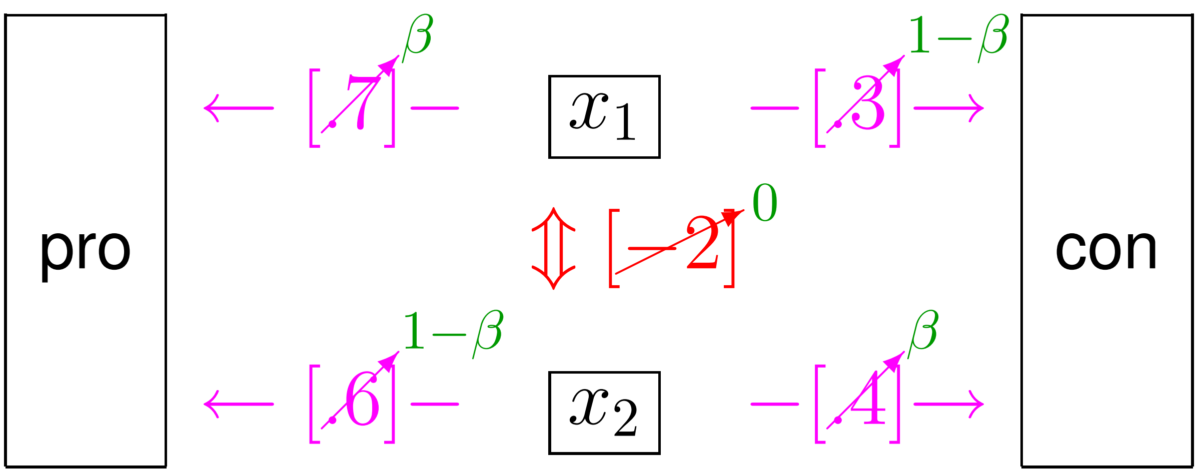


But can we more directly encourage c_1 to be opposite to c_2?

SetTo: To decide which of x_1 and x_2 is "pro",

- First, support the preference expressed by the largest individual score — here, $\mathrm{Ind}(x_1, \text{pro}) = .7$ — by setting it to large β.
- Then, support $c_2 = \overline{c_1}$ by setting $\mathrm{Ind}(x_2, \text{pro})$ to $1 - \beta$.

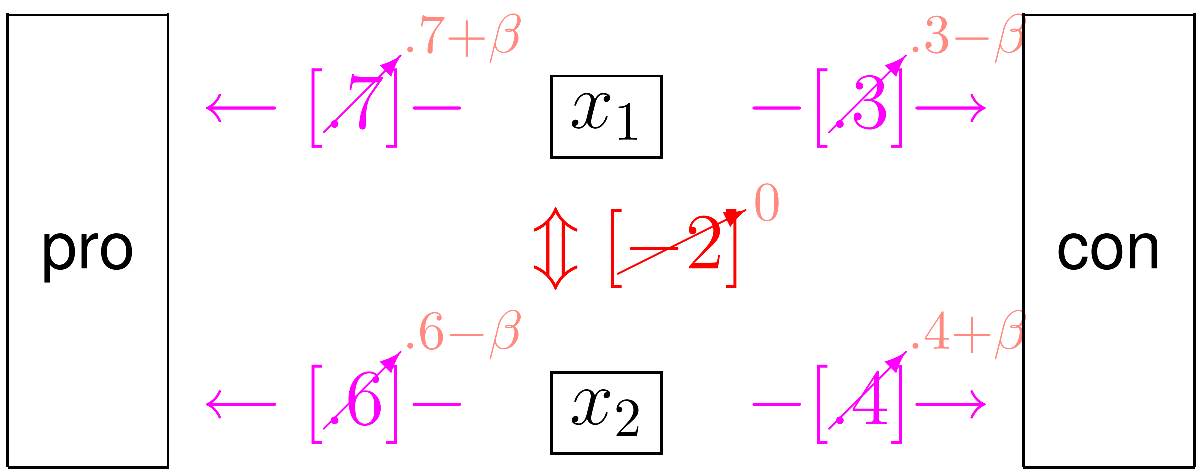
Etc.



IncBy: Similar to SetTo, but preserves some initial individual information.

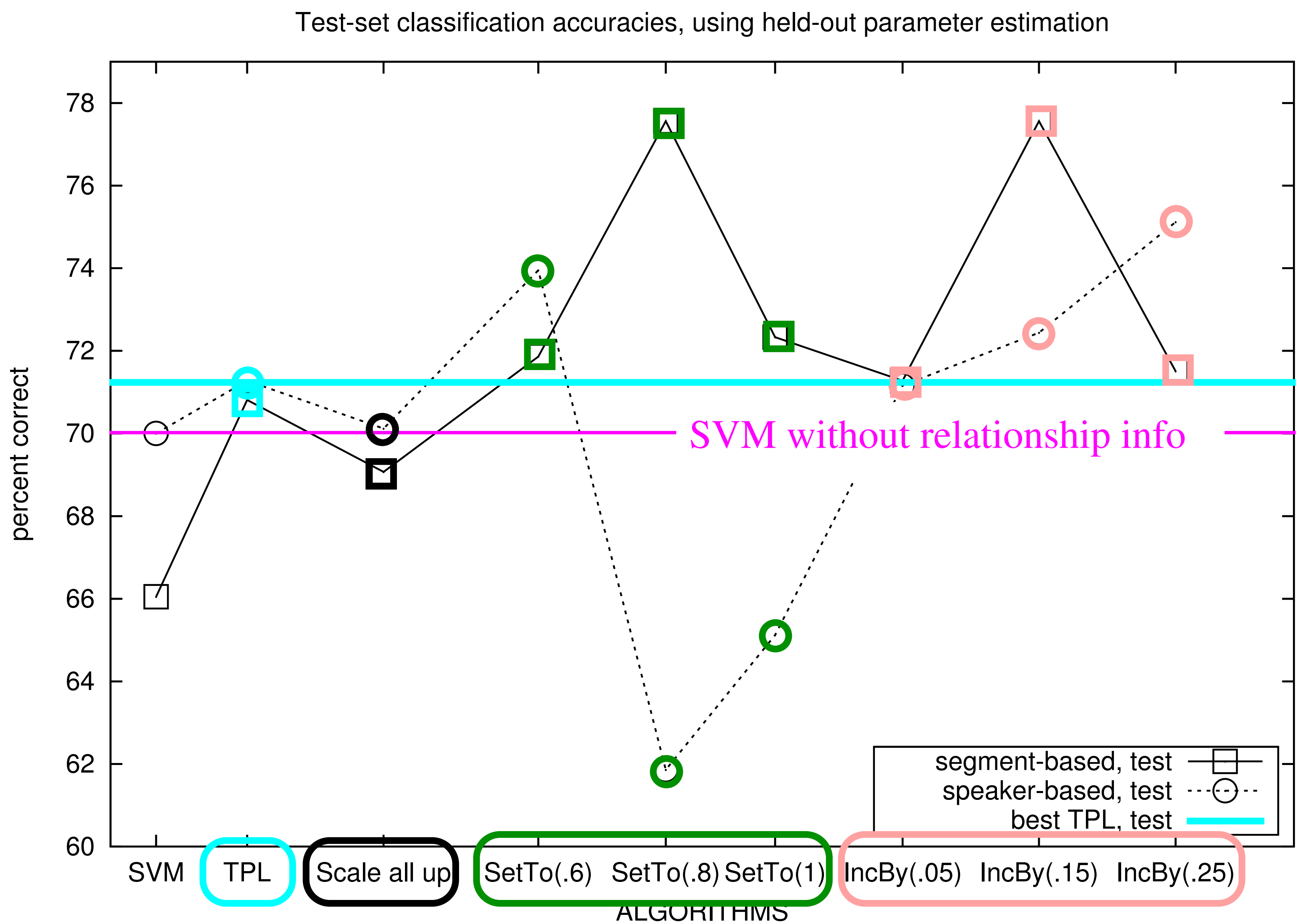
- **Increment** the largest individual preference — here, $\mathrm{Ind}(x_1, \textsf{pro}) = .7$ — rather than completely re-write it.
- Then, support $c_2 = \overline{c_1}$ by **decrementing** $\mathrm{Ind}(x_2, \textsf{pro})$.

Etc.



Evaluation: we used TPL 2006's experimental set-up and data, including train/dev/test splits and individual and association scores, available at www.cs.cornell.edu/home/llee/data/convote.html

Technical issue: the TPL 2006 gold-standard labels are the same for every speech uttered by the same speaker. They evaluated two different ways to enforce this constraint, called "segment-based" and "speaker-based", so we do, too.


Test-set classification accuracies, using held-out parameter estimation
percent correct
78
76
74
72
70
68
66
64
62
60
SVM without relationship info
segment-based, test
speaker-based, test
best TPL, test
SVM
TPL
Scale all up
SetTo(.6)
SetTo(.8)
SetTo(1)
IncBy(.05)
IncBy(.15)
IncBy(.25)
ALGORITHMS

Analysis of results:

- Simply scaling up assocs (***Scale all up***) is worse than ignoring disagreement (*TPL*).
- ***SetTo* and *IncBy* incorporate disagreement in a more sophisticated fashion that can yield better results than previous work.**

To do: explain the non-regularities in results and perform more analysis; consider more sophisticated heuristics; consider as alternatives approximation algorithms, more sophisticated formalisms, ... [your suggestions here!]