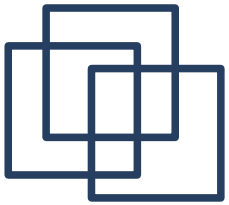
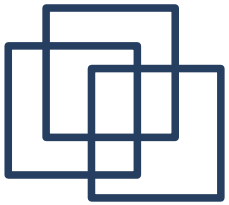


Cop k-means

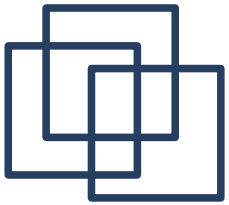
Danilo Vasconcellos Vargas



If we had background knowledge...

- Sometimes there is information about the problem

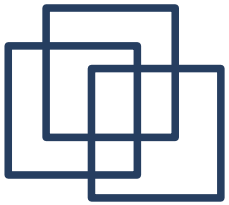
- Constraints
- Also called Semi-supervised clustering



Constraints (Cop-kmeans)

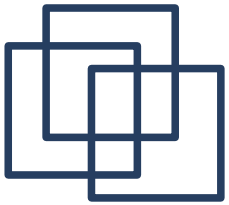
- Must-link
- Cannot-link

- **Constraints are never broken!**



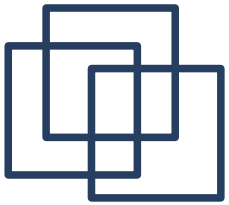
k-means

- Initialize k cluster centers
- Assign Phase
  - objects are assigned to closest cluster center
- Update Cluster Centers
  - update the cluster centers to the mean of constituent objects



Cop k-means

- Initialize k cluster centers
- Assign Phase
  - objects are assigned to closest cluster center **without violating constraints**
- Update Cluster Centers
  - update the cluster centers to the mean of constituent objects

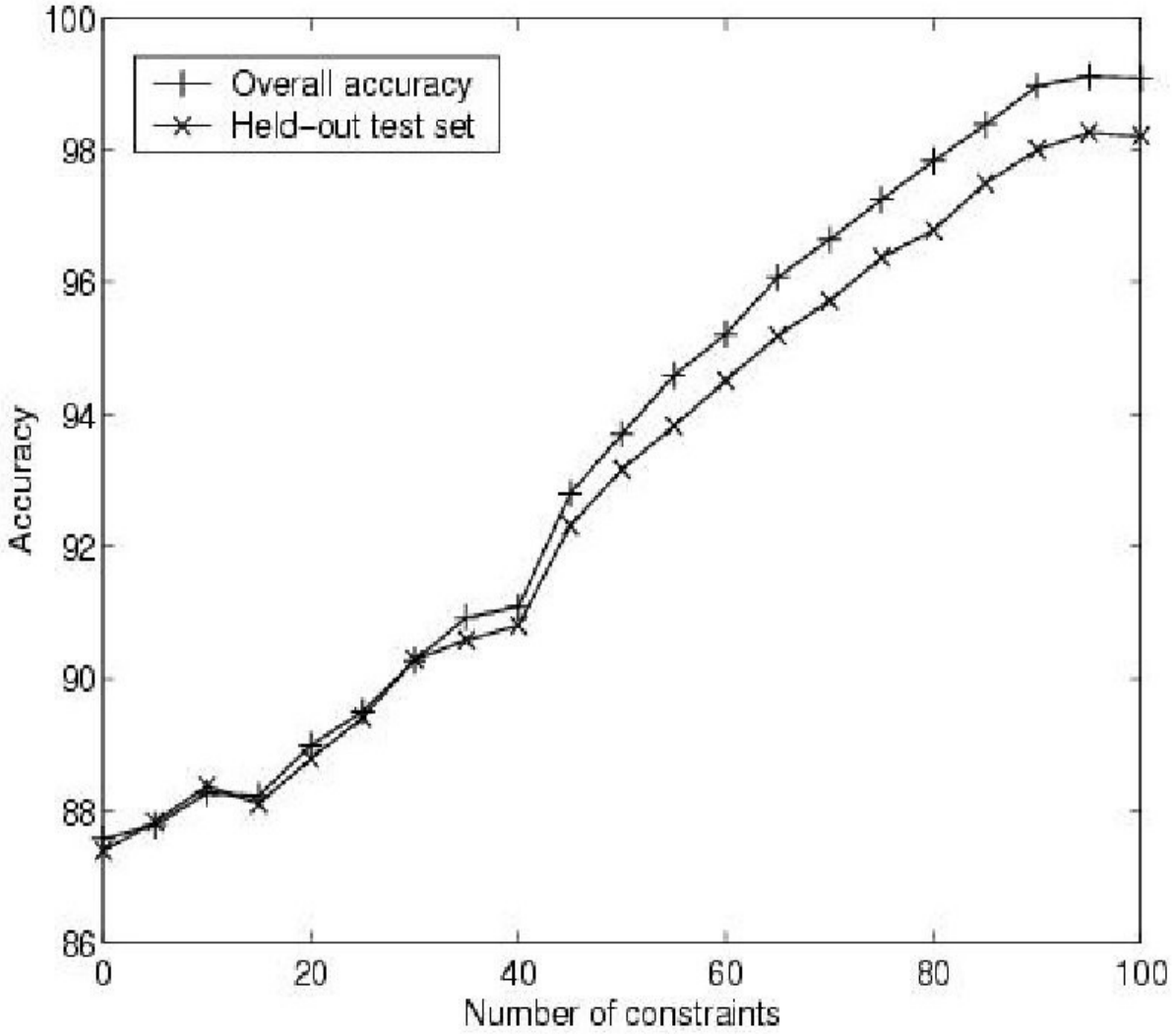


Handling Constraints

For all objects try to assign it to closest k

- 1. No constraint broken:
  - Assign object *o* to cluster *k*.
- 2. Broken → is there a next closest cluster ?
  - Yes → Back to 1.
  - No → 3.
- 3. fail

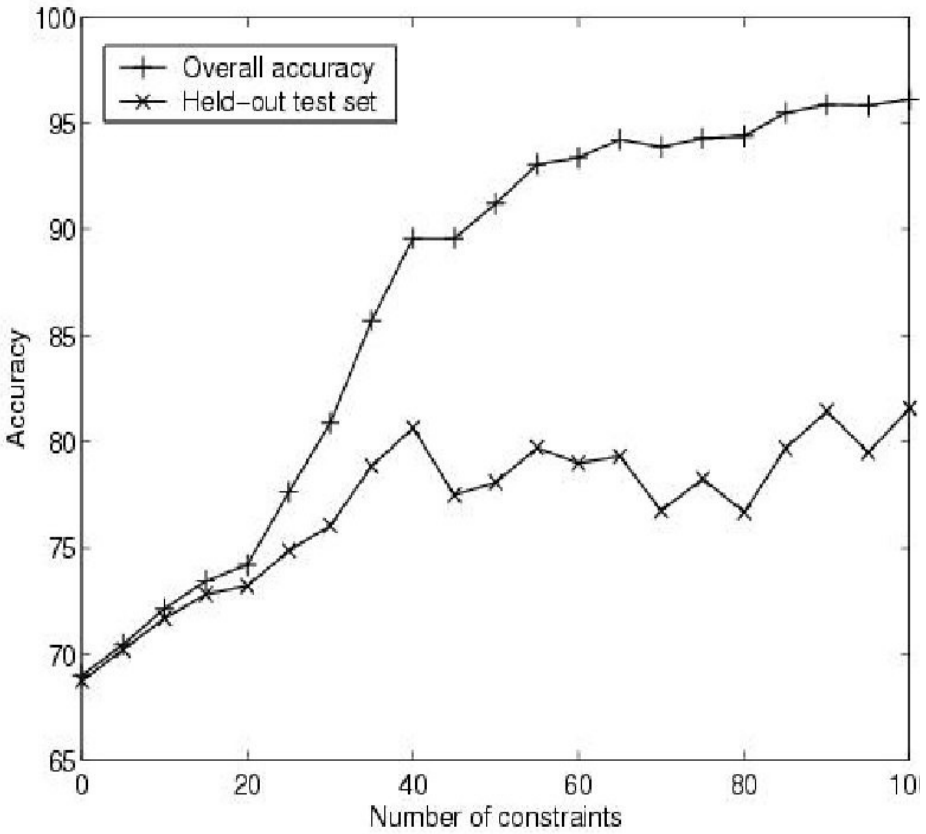
Experimental Results

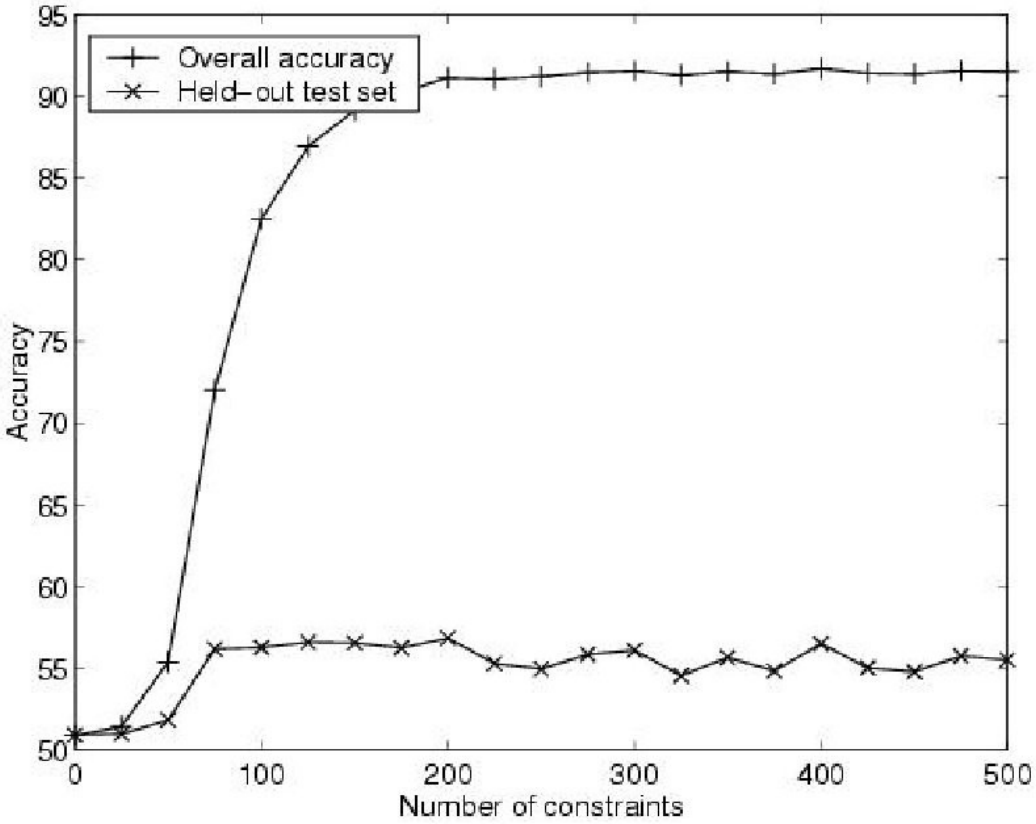


- 48% can be determined solely by constraints.
- 100 trials
- Held-out test is not directly affected by constraints

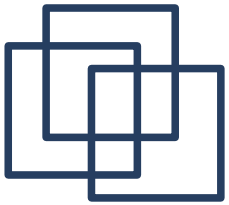
(Wagstaff et al, 2001)

More Results





(Wagstaff et al, 2001)



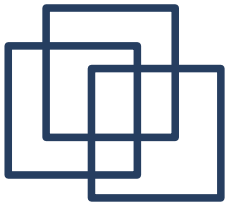
Comments on Results

- Are constraints worthwhile?
  - Depends on the dataset
  - Constraints can be generalized to the full dataset?
- Sensitivity to assignment order
  - Studied and solved by (Hong and Kwong, 2009) using an ensemble algorithm.

Comments on Results

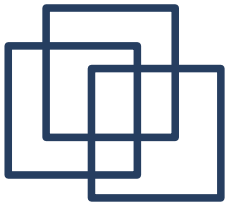
- The set of constraints can vary and so their impact on the accuracy (Wagstaff, Basu and Davidson 2006)

Data set	Accuracy		
	Min	Mean	Max
Glass	67.6	69.9	72.3
Iris	82.2	88.4	93.4
Ionosphere	58.2	60.1	62.3
Wine	68.0	71.3	74.3



Semi Supervised Clustering

- Similarity-adapting methods
  - Example: modifying the Euclidian Distance
- Search-based methods
  - Example: Cop-kmeans



Extensions

- Soft Constraints

Bibliography

Yi Hong and Sam Kwong “Learning Assignment Order of Instances for the constrained k-means clustering algorithm” IEEE Transactions on Systems, Man, and Cybernetics, Vol 39, No 2. April, 2009.

Wagstaff, Kiri L., Basu, Sugato, Davidson, Ian “When is constrained clustering beneficial, and why?” National Conference on Aritficial Intelligence, Boston, Massachusetts 2006.

Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl “Constrained K-means Clustering with Background Knowledge” ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, 2001.