

Clustering Using Integer Programming

Katie Kuksenok

Michael Brooks

What are we doing?

- Investigating IP formulations of the clustering problem
- Developing a web application for clustering
- Applying clustering to several datasets
- Visualizing resulting clusters

The Clustering Problem

- Input:
 - N entities
 - Each entity has a value for each of M features
- Goal:
 - Partition the entities “optimally”
- Issues with clustering:
 - Definition of “optimal”
 - Different feature types don’t work well together

The Clustering Problem

- 😊 Exposes high-level properties of data
- 😊 Important for exploratory data analysis in data mining
- 😞 Finding the optimal clustering is NP-Complete

Grotschel and Wakabayashi IP

- Recall the Grotschel-Wakabayashi formulation (clustering wild cats):
- Variables: $\mathbf{x}_{ij} = 1$ if \mathbf{i} and \mathbf{j} are in the same cluster
- Data: $\mathbf{w}_{ij} = \text{disagreements}(\mathbf{i}, \mathbf{j}) - \text{agreements}(\mathbf{i}, \mathbf{j})$
- Minimize: SUM $\{ \mathbf{w}_{ij} \mathbf{x}_{ij} : \text{over all } \mathbf{i}, \mathbf{j} \}$

Formulation: Constraints

- Constraints for reflexivity, symmetry, transitivity reduce to the following constraints:

$$\forall 1 \leq i < j < k \leq n$$

$$x_{ij} + x_{jk} - x_{ik} \leq 1$$

$$x_{ij} - x_{jk} + x_{ik} \leq 1$$

$$-x_{ij} + x_{jk} + x_{ik} \leq 1$$

Exploratory Clustering Application

- Currently in development
- User provides data to website
- Web server processes data
 - Produces OPL code
 - Runs the model in LPSolve (another MIP solver)
- Results are returned to the user, with visuals

Preliminary Exploration

- With working parts of application:
- can cluster and visualize datasets

- verified Grotschel/Wakabayashi clustering results
- used clustering to explore senate voting records

Data Sets (Cats)

- Cluster 30 varieties of wild cats (from paper)
- Cats have 14 features
- Reproduced results from paper

- Model has 12180 constraints and 435 variables
- Runs in less than 1 second
- No branching

	Cluster 1 (2)	Cluster 2 (6)	Cluster 3 (1)	Cluster 4 (21)
<i>Aspect of the pelt:</i>	without spots, uniformly colored with stripes	with spots	with spots	without spots, uniformly colored with spots
<i>Fur:</i>	short-haired	short-haired long-haired	short-haired	short-haired long-haired
<i>Ears:</i>	rounded	rounded	rounded	rounded
<i>Height (H) up to shoulder:</i>	$H > 70 \text{ cm}$	$H > 70 \text{ cm}$ $50 \text{ cm} < H < 70 \text{ cm}$	$H > 70 \text{ cm}$	$H \leq 50 \text{ cm}$
<i>Weight (W):</i>	$W > 80 \text{ kg}$	$W > 80 \text{ kg}$ $10 \text{ kg} < W \leq 80 \text{ kg}$	$10 \text{ kg} < W \leq 80 \text{ kg}$	$W \leq 10 \text{ kg}$
<i>Length (L) of body:</i>	$L > 150 \text{ cm}$	$80 \text{ cm} < L \leq 150 \text{ cm}$	$80 \text{ cm} < L \leq 150 \text{ cm}$	$L \leq 80 \text{ cm}$
<i>Length of tail compared with length of body:</i>	median	long short	long	short median
<i>(Teeth) canines:</i>	very developed	very developed	little developed	little developed
<i>(Larynx) Lingual bone:</i>	present	present absent	absent	absent
<i>Retractile claws:</i>	yes	yes	no	yes
<i>Predatory behavior:</i>	diurnal nocturnal	diurnal and nocturnal	diurnal	diurnal and nocturnal nocturnal
<i>Type of Prey:</i>	big prey	big or small prey	big or small prey	small prey
<i>Climbs trees:</i>	no	yes	no	yes
<i>Chases after or lies in wait for the prey:</i>	chase wait	wait	chase	wait

Data Sets (Senate Voting Records)

- 100 senators
- Senate records for 8 passed bills from this year
- Model has 485100 constraints and 4950 variables
- Runs in 1 minute
- No branching

	Cluster 1 (74)	Cluster 2 (20)	Cluster 3 (2)	Cluster 4 (2)	Cluster 5 (1)	Cluster 6 (1)
<i>Party:</i>	D, a	R	D	R	R	D
<i>H.R. 2847:</i>	Yea	Nay	Yea Not Voting	Nay	Yea	Yea
<i>H.R. 3326:</i>	Yea	Nay Yea	Yea	Yea	Yea	Yea
<i>H.R. 2996:</i>	Yea	Nay	Yea Not Voting	Yea	Yea	Not Voting
<i>H.R. 3288:</i>	Yea	Nay	Yea	Nay	Yea	Not Voting
<i>S.1023:</i>	Yea	Yea Nay	Yea	Nay	Nay	Not Voting
<i>H.R. 3435:</i>	Yea	Nay	Not Voting	Nay	Nay	Not Voting
<i>H.R. 2997:</i>	Yea	Nay	Not Voting	Yea	Nay	Not Voting
<i>H.R. 3357:</i>	Yea	Nay Yea	Not Voting	Yea	Nay	Not Voting

Data Sets (Senate Voting By State)

- Grouped senate voting by state
- If senators for a state disagreed, 'Maybe'
- Model has 58800 constraints and 1225 variables
- Runs in 1.5 seconds
- No branching

	Cluster 1 (35)	Cluster 2 (8)	Cluster 3 (5)	Cluster 4 (2)
<i>Party:</i>	D _{RR}	RD _R	R	R
<i>H.R. 2847:</i>	Yea _{Maybe}	Maybe	Nay	Nay
<i>H.R. 3326:</i>	Yea _{Maybe}	Yea	Nay _{Maybe}	Yea
<i>H.R. 2996:</i>	Yea _{Maybe}	Maybe	Nay	Yea Maybe
<i>H.R. 3288:</i>	Yea _{Maybe}	Maybe	Nay	Nay
<i>S.1023:</i>	Yea _{Maybe}	Yea _{Maybe}	Nay _{Maybe}	Nay Maybe
<i>H.R. 3435:</i>	Yea _{Maybe}	Nay _{Maybe}	Nay	Nay
<i>H.R. 2997:</i>	Yea _{Maybe}	Yea _{Maybe}	Nay _{Maybe}	Yea Nay
<i>H.R. 3357:</i>	Maybe _{Yea} Yea	Maybe _{Yea}	Maybe _{Nay} Nay	Yea

What now?

- The Application
 - We'd like to put it all together
- Branch and Bound
 - No real-life datasets so far have required branching
 - We have constructed datasets that did
 - Can we determine under what conditions branching will be required?

This is a guepard (cheetah in French?):



...any other questions?

Approaches to Clustering

- What makes a good clustering algorithm?

- Many algorithms
- Runtime and space efficiency (when N is large)
 - Ability to quickly incorporate new data
- IP
- Ability to deal with many-dimensional data (when M is large)
 - Optimality (closeness to)