II. Cluster Analysis

- Cluster Analysis Basics
- Hierarchical Cluster Analysis
- Iterative Cluster Analysis
- Density-Based Cluster Analysis
- Cluster Evaluation
- Constrained Cluster Analysis
Constrained Cluster Analysis
Person Resolution Task
Constrained Cluster Analysis
Person Resolution Task
Constrained Cluster Analysis
Person Resolution Task

target name
*Michael Jordan*

other names
Constrained Cluster Analysis
Person Resolution Task

The basket ball player.

The statistician.
Multi-document resolution task:

Names, Target names: \( N = \{n_1, \ldots, n_l\}, \quad T \subset N \)

Referents: \( R = \{r_1, \ldots, r_m\}, \quad \tau : R \to T, \quad |R| \gg |T| \)

Documents: \( D = \{d_1, \ldots, d_n\}, \quad \nu : D \to \mathcal{P}(N), \quad |\nu(d_i) \cap T| = 1 \)

A solution: \( \gamma : D \to R, \quad \text{s.t.} \quad \tau(\gamma(d_i)) \in \nu(d_i) \)
Constrained Cluster Analysis
Person Resolution Task

The basket ball player. The statistician.

Facts about the Spock data mining challenge:

Target names: \( |T| = 44 \)
Referents: \( |R| = 1101 \)
Documents: \( |D_{\text{train}}| = 27,000 \) (labeled \( \approx 2.3\)GB)
\( |D_{\text{test}}| = 75,000 \) (unlabeled \( \approx 7.8\)GB)
Constrained Cluster Analysis
Person Resolution Task

- up to 105 referents for a single target name
- about 25 referents on average per target name
- about 23 documents on average per referent

Facts about the Spock data mining challenge:

- Target names: $|T| = 44$
- Referents: $|R| = 1101$
- Documents:
  - $|D_{\text{train}}| = 27,000$ (labeled $\approx 2.3$GB)
  - $|D_{\text{test}}| = 75,000$ (unlabeled $\approx 7.8$GB)
Constrained Cluster Analysis
Applied to Multi-Document Resolution

1. Model similarities ➔ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➔ logistic regression

3. Find equivalence classes (unsupervised) ➔ cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization
1. Model similarities → new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) → logistic regression

3. Find equivalence classes (unsupervised) → cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization
Constrained Cluster Analysis
Applied to Multi-Document Resolution

1. Model similarities ➔ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➔ logistic regression

3. Find equivalence classes (unsupervised) ➔ cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization
Constrained Cluster Analysis
Applied to Multi-Document Resolution

1. Model similarities ➔ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➔ logistic regression

3. Find equivalence classes (unsupervised) ➔ cluster analysis:
   - adaptive graph thinning
   - multiple, density-based cluster analysis
   - clustering selection by expected density maximization
Constrained Cluster Analysis
Applied to Multi-Document Resolution

1. Model similarities \(\Rightarrow\) new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) \(\Rightarrow\) logistic regression

3. Find equivalence classes (unsupervised) \(\Rightarrow\) cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization
1. Model similarities ➔ new and established retrieval models:
   - global and context-based vector space models
   - explicit semantic analysis
   - ontology alignment

2. Learn class memberships (supervised) ➔ logistic regression

3. Find equivalence classes (unsupervised) ➔ cluster analysis:
   (a) adaptive graph thinning
   (b) multiple, density-based cluster analysis
   (c) clustering selection by expected density maximization
Constrained Cluster Analysis

Idealized Class Membership Distribution over Similarities

Similarity distributions for document pairs from different referents and same referent.

Logistic regression task:

- sample size: 400 000
- classes imbalance: non-target class : target class ≈ 25:1
- items are drawn uniformly distributed wrt. non-targets and targets
- items are uniformly distributed over the groups of target names
Constrained Cluster Analysis
Membership Distribution under $tf\cdot idf$ Vector Space Model

Model details:

- corpus size: 25,000 documents
- dictionary size: 1,2 Mio terms
- stopwords number: 850
- stopword volume: 36%
Constrained Cluster Analysis
Membership Distribution under Context-Based Vector Space Model

Model details:
- corpus size: 25,000 documents
- dictionary size: 1.2 Mio terms
- stopwords number: 850
- stopword volume: 36%
Constrained Cluster Analysis
Membership Distribution under Ontology Alignment Model

Model details:

- DMOZ open directory project
- > 5 million documents
- 12 top-level categories
- 31 second level categories
- ML: hierarchical Bayes
- training set: 100 000 pages

Top

Arts Business Computers Games World

Virtual Reality Algorithms AI

Classifier
Constrained Cluster Analysis
In-Depth: Multi-Class Hierarchical Classification

Flat (big-bang) classification

- simple realization
- loss of discriminative power with increasing number of categories

Hierarchical (top-down) classification

- specialized classifiers (divide and conquer)
- misclassification at higher levels can never become repaired
Constrained Cluster Analysis
In-Depth: Multi-Class Hierarchical Classification

State of the art of effectiveness analyses:

1. independence assumption between categories
2. neglection of both hierarchical structure and degree of misclassification

Improvements:

- Consider similarity $\varphi(C_i, C_j)$ between correct and wrong category.
- Consider graph distance $d(C_i, C_j)$ between correct and wrong category.
Constrained Cluster Analysis
In-Depth: Multi-Class Hierarchical Classification

Improvements continued:

- Multi-label (multi path) classification
- Multi-classifier (ensemble) classification

- Traverse more than one path and return all labels
- Employ probabilistic classifiers with a threshold: split a path or not
- Classification result is a majority decision
- Employ different classifier (different types or differently parameterized)
### Constrained Cluster Analysis

**Membership Distribution under Optimized Retrieval Model Combination**

<table>
<thead>
<tr>
<th>Retrieval Model</th>
<th>$F_{1/3}$ -Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>tfidf vector space</td>
<td>0.39</td>
</tr>
<tr>
<td>context-based vector space</td>
<td>0.32</td>
</tr>
<tr>
<td>ESA Wikipedia persons</td>
<td>0.30</td>
</tr>
<tr>
<td>phrase structure grammar</td>
<td>0.17</td>
</tr>
<tr>
<td>ontology alignment</td>
<td>0.15</td>
</tr>
<tr>
<td>optimized combination</td>
<td><strong>0.42</strong></td>
</tr>
</tbody>
</table>
Constrained Cluster Analysis

Membership Distribution under Optimized Retrieval Model Combination

<table>
<thead>
<tr>
<th>Retrieval Model</th>
<th>$F_{1/3}$ -Measure</th>
<th>Referent 1</th>
<th>Referent 2</th>
<th>Referent m</th>
</tr>
</thead>
<tbody>
<tr>
<td>tfidf vector space</td>
<td>0.39</td>
<td>⊙</td>
<td>⊙</td>
<td></td>
</tr>
<tr>
<td>context-based vector space</td>
<td>0.32</td>
<td>⊙</td>
<td>⊙</td>
<td></td>
</tr>
<tr>
<td>ESA Wikipedia persons</td>
<td>0.30</td>
<td>⊙</td>
<td>⊙</td>
<td>⊙</td>
</tr>
<tr>
<td>phrase structure grammar</td>
<td>0.17</td>
<td>⊙</td>
<td>⊙</td>
<td>...</td>
</tr>
<tr>
<td>ontology alignment</td>
<td>0.15</td>
<td></td>
<td></td>
<td>⊙</td>
</tr>
<tr>
<td>optimized combination</td>
<td>0.42</td>
<td></td>
<td></td>
<td>⊙ ⊙</td>
</tr>
</tbody>
</table>
Constrained Cluster Analysis
Membership Distribution under Optimized Retrieval Model Combination

**Retrieval Model** | $F_{1/3}$ -Measure
---|---
tfidf vector space | 0.39
context-based vector space | 0.32
ESA Wikipedia persons | 0.30
phrase structure grammar | 0.17
ontology alignment | 0.15
optimized combination | **0.42**

Referents:
- Referent 1
- Referent 2
- Referent m
Constrained Cluster Analysis
Membership Distribution under Optimized Retrieval Model Combination

<table>
<thead>
<tr>
<th>Retrieval Model</th>
<th>$F_{1/3}$ Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>tfidf vector space</td>
<td>0.39</td>
</tr>
<tr>
<td>context-based vector space</td>
<td>0.32</td>
</tr>
<tr>
<td>ESA Wikipedia persons</td>
<td>0.30</td>
</tr>
<tr>
<td>phrase structure grammar</td>
<td>0.17</td>
</tr>
<tr>
<td>ontology alignment</td>
<td>0.15</td>
</tr>
<tr>
<td>optimized combination</td>
<td><strong>0.42</strong></td>
</tr>
</tbody>
</table>

![Graph showing membership distribution](image)

- Different referents: Orange bars
- Same referent: Purple bars

*DM:II-296 Cluster Analysis © STEIN 2007-2020*
Constrained Cluster Analysis
Membership Distribution under Optimized Retrieval Model Combination

In the example:
- precision = 0.4
- recall = 0.43
- $F_{1/3} = 0.41$

(if false negatives are uniformly distributed)
Consideration of imbalance:
Constrained Cluster Analysis
In-Depth: Analysis of Classifier Effectiveness

- class imbalance factor (CIF) of 25
  ⇒ precision in interval [0.725; 1] for edges between same referents: ≈ 0.17

How can $F_{1/3} = 0.42$ be achieved via cluster analysis?

Consideration of imbalance:

![Graph showing interval with different referents and same referents]
Constrained Cluster Analysis
In-Depth: Analysis of Classifier Effectiveness

Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow |TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)
$\Rightarrow \frac{|TP|}{|C|}$ degree of true positives per node (here: 11)
$\Rightarrow |TP|(\frac{1}{\text{precision}} - 1)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, $FP^*$, connect to same cluster

$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)
$\Rightarrow$ edge tie factor (ETF) specifies the excess of true positives until tie (here: 3…5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}, \quad \text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$
Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness

Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow \frac{|TP|}{|C|}$ true 1-similarities per cluster (here: 130 @ threshold 0.725)

$\Rightarrow \frac{|TP|}{|C|}$ degree of true positives per node (here: 11)

$\Rightarrow |TP| \left(\frac{1}{\text{precision}} - 1\right)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, $FP^*$, connect to same cluster

$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

$\Rightarrow$ edge tie factor ($ETF$) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}, \quad \text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$
Constrained Cluster Analysis
In-Depth: Analysis of Classifier Effectiveness

Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow |TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)

$\Rightarrow \frac{|TP|}{|C|}$ degree of true positives per node (here: 11)

$\Rightarrow |TP|(\frac{1}{\text{precision}} - 1)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, $FP^*$, connect to same cluster

$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

$\Rightarrow$ edge tie factor ($ETF$) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)}$$

effective precision $= precision \cdot \frac{CIF}{ETF}$
Constrained Cluster Analysis
In-Depth: Analysis of Classifier Effectiveness

Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow |TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)

$\Rightarrow \frac{|TP|}{|C|}$ degree of true positives per node (here: 11)

$\Rightarrow |TP|\left(\frac{1}{\text{precision}} - 1\right)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, $FP^*$, connect to same cluster

$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

$\Rightarrow$ edge tie factor ($ETF$) specifies the excess of true positives until tie (here: 3 ... 5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*)},$$

effective precision $= \text{precision} \cdot \frac{CIF}{ETF}$
Constrained Cluster Analysis

In-Depth: Analysis of Classifier Effectiveness

Assumption: uniform distribution of referents over documents (here: 25 clusters with $|C| = 23$)

$\Rightarrow |TP|$ true 1-similarities per cluster (here: 130 @ threshold 0.725)

$\Rightarrow \frac{|TP|}{|C|}$ degree of true positives per node (here: 11)

$\Rightarrow |TP| \left(1 - \frac{1}{\text{precision}}\right)$ false 1-similarities per cluster (here: 760)

Density-based cluster analysis: effective false positives, $FP^*$, connect to same cluster

$\Rightarrow$ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

$\Rightarrow$ edge tie factor ($ETF$) specifies the excess of true positives until tie (here: 3...5)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*)}, \quad \text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$
Density-based cluster analysis: effective false positives, $FP^*$, connect to same cluster

⇒ analyze $P(|FP^*| > k \mid D, R_{iid})$ (here: $E(|FP^*|) = 2.7$)

⇒ edge tie factor ($ETF$) specifies the excess of true positives until tie (here: $3\ldots5$)

$$ETF = \frac{|TP|}{|C| \cdot E(|FP^*|)},$$

$$\text{effective precision} = \text{precision} \cdot \frac{CIF}{ETF}$$
Constrained Cluster Analysis
In-Depth: Analysis of Classifier Effectiveness

Assumption: uniform distribution of referents over documents (here: 25 clusters with \(|C| = 23\))

\[ |TP| \text{ true 1-similarities per cluster (here: 130 \text{ @ threshold 0.725})} \]

\[ |TP| |C| \text{ degree of true positives per node (here: 11)} \]

\[ |TP| \text{ (1 - precision)} \text{ false 1-similarities per cluster (here: 760)} \]

Density-based cluster analysis: effective false positives, \(FP^*\), connect to same cluster

\[ \frac{E(|FP^*)}{|D,R|_{iid}} \text{ (here: 2.7)} \]

\[ \text{edge tie factor (ETF)} = \frac{|TP| |C| \cdot E(|FP^*)}{\text{effective precision} = \text{precision} \cdot \text{CIF}} \]

Determine optimum similarity threshold for class-membership function:

\[ \theta^* = \arg \max_{\theta \in [0;1]} \left\{ \frac{1 + \alpha \cdot \text{ETF}}{\text{precision}_{\theta} \cdot \text{CIF}} \right\} \]

\(\theta^*\) considers co-variate shift, introduces model formation bias and sample selection bias.
Constrained Cluster Analysis
Model Selection: Our Risk Minimization Strategy

<table>
<thead>
<tr>
<th>Retrieval Model</th>
<th>$F_{1/3}$ -Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>tfidf vector space</td>
<td>0.39</td>
</tr>
<tr>
<td>context-based vector space</td>
<td>0.32</td>
</tr>
<tr>
<td>ESA Wikipedia persons</td>
<td>0.30</td>
</tr>
<tr>
<td>phrase structure grammar</td>
<td>0.17</td>
</tr>
<tr>
<td>ontology alignment</td>
<td>0.15</td>
</tr>
<tr>
<td>optimized combination</td>
<td><strong>0.42</strong></td>
</tr>
<tr>
<td>Ensemble cluster analysis</td>
<td><strong>0.40</strong></td>
</tr>
</tbody>
</table>

Ensemble cluster analysis: higher bias, better generalization.

1) Do we speculate on a better fit for $D_{test}$?

2) Do we expect a significant covariate shift, more noise, etc. in $D_{test}$?
Constrained Cluster Analysis

Recap

1. Multi-document resolution can be tackled with constrained cluster analysis.

2. Constraints are derived from labeled examples.

3. Class membership function ties constraints to multiple retrieval models.

4. Advanced density-based clustering technology is key.
Constrained Cluster Analysis

References

- Disambiguating Web Appearances of People in a Social Network.
  [R. Bekkerman, A. McCallum. WWW 2005]

- A Bayesian Model for Supervised Clustering with the Dirichlet Process Prior.

  [E. Gabrilovich, S. Markovitch. IJCAI 2007]

- Unsupervised Discrimination of Person Names in Web Contexts.
  [T. Pedersen, A. Kulkarni. CICLing 2007]

- On Information Need and Categorizing Search.

- Weighted Experts: A Solution for the Spock Data Mining Challenge.
  [B. Stein, S. Meyer zu Eissen. I-KNOW 2008]

- GRAPE: A System for Disambiguating and Tagging People Names in Web Search.