Multi-way Distributional Clustering via Pairwise Interactions

Ron Bekkerman  
Ran El-Yaniv  
Andrew McCallum

UMass  
Technion  
UMass

Email: ronb@cs.umass.edu
Contingency table clustering

- Not necessarily documents!
  - Images/features, genes/samples, movies/actors…
Contingency table clustering

Similarly we can cluster columns (words)

Two-way clustering

Or both!
Two-way clustering

a.k.a. Double clustering, Bi-clustering, Co-clustering, Coupled clustering etc.

- Main motivation:
  - Can overcome statistical sparseness

- Extensively studied:
  - Slonim & Tishby, SIGIR-2000
  - Getz, Levine & Domany PNAS-2000
  - El-Yaniv & Souroujon NIPS-2001
  - Dhillon, Mallela & Modha KDD-2003

- All showing impressive improvements
Multiple views

- Various views of data can be observed
- **Example**: email

  - messages
  - attachments
  - words
  - senders
  - recipients
  - subject lines
  - markup items
Multiple views

- Various views of data can be observed
- **Example**: email

- Statistical interaction of views
Multiple views

- Various views of data can be observed
- **Example:** email

- Statistical interaction of views
  - Not necessarily all interactions are relevant
Motivating question: can we extend two-way clustering to utilize multiple views?

Goal: construct $N$ “clean” clusterings of $N$ interdependent variables
Our contributions

- **Objective function** for fitting useful multi-way interaction model
- **Novel clustering algorithm** to maximize the objective
- **Striking empirical results**
Earlier attempts

- **Multivariate Information Bottleneck (mIB)**
  - Friedman et al. UAI-2001
  - Very general approach for dealing with several variables
  - **Objective**: Multi-Information

$$I(\tilde{X}_1;...;\tilde{X}_N) = \sum_{\tilde{x}_1,...,\tilde{x}_N} P(\tilde{X}_1,...,\tilde{X}_N) \log \frac{P(\tilde{X}_1,...,\tilde{X}_N)}{P(\tilde{X}_1)...P(\tilde{X}_N)}$$

- Not feasible for practical applications
Our approach

- Consider only pairwise interactions
- Pairwise interaction graph
  - Defines interactions between $N \geq 2$ variables
Our objective

- Let \((\tilde{X}, E)\) be pairwise interaction graph
- Extending Dhillon et al.:
- **Objective:** weighted sum of pairwise MI
  \[
  \max_{\tilde{X}_1, \ldots, \tilde{X}_N} \sum_{(\tilde{X}_i, \tilde{X}_j) \in E} w_{ij} \ I(\tilde{X}_i; \tilde{X}_j)
  \]
  - Subject to \(|\tilde{X}_i| = K_i, \ i = 1, \ldots, N\)
- No multi-dimensional probability tables
- Can be easily factorized
Objective factorization

- Consider triangle:

- Objective in this case:
  \[
  \max_{\tilde{X}, \tilde{Y}, \tilde{Z}} \ w_1 I(\tilde{X}; \tilde{Y}) + w_2 I(\tilde{Y}; \tilde{Z}) + w_3 I(\tilde{X}; \tilde{Z})
  \]

- ...is broken into 3 parts:
  \[
  \begin{align*}
  \max_{\tilde{X}} & \quad w_1 I(\tilde{X}; \tilde{Y}) + w_3 I(\tilde{X}; \tilde{Z}) \\
  \max_{\tilde{Y}} & \quad w_1 I(\tilde{X}; \tilde{Y}) + w_2 I(\tilde{Y}; \tilde{Z}) \\
  \max_{\tilde{Z}} & \quad w_2 I(\tilde{Y}; \tilde{Z}) + w_3 I(\tilde{X}; \tilde{Z})
  \end{align*}
  \]
Implementation

- We have tried various schemes:
  - Top-down
  - Bottom-up
  - Flat (K-means, sequential IB)
- Best results obtained with **hybrid**
  - Top-down for some variables
  - Bottom-up for other variables
  - Flat correction routine after each split/merge
Multi-way Distributional Clustering

- **Initialization**
  - If $i \in S^{up}$, put each $x_i$ in a singleton cluster
  - If $i \in S^{down}$, put all $x_i$ in one common cluster

- **Main loop**
  - If $i \in S^{up}$, merge every two closest clusters
  - If $i \in S^{down}$, split each cluster to two halves

- **Correction loop**
  - Pull each $x_i$ out of its cluster
  - Put it into $\tilde{x}_i$ s.t. the objective is maximized

MDC

Slonim et al.
SIGIR-2002
Example: 4-way MDC
Computational complexity

- General case
  - At each iteration of the main loop:
    - Pass over all $x_i$
    - Pass over all $\tilde{x}_i$
    - Pass over all $\tilde{x}_j$, $\forall j \neq i$

- If bottom-up system is only one
  $\implies o(|X_i|^3)$

- 2-way case
  $\mathcal{X} \xrightarrow{\mathcal{Y}}$
  - At each iteration $|\tilde{x}_i|$ is doubled
  - While $|\tilde{y}_i|$ is halved

$O(|X_i|^2)$
Experimental setup

- 2-way MDC
  - *Documents* and *Words*

- 3-way MDC
  - *Documents*, *Words* and *Authors*

- 4-way MDC
  - *Documents*, *Words*, *Authors* and *documents’ Titles*

- **Documents**: bottom-up, the rest: top-down
Evaluation methodology

- Clustering evaluation
  - Is generally unintuitive
  - Is an entire ML research field

- We use the “accuracy” measure
  - Following Slonim et al. and Dhillon et al.
  - Ground truth:
  - Our results:

\[
Acc = \frac{1}{|X|} \sum_{c} \gamma_c
\]

Size of dominant class in cluster \(c\)
Datasets

- Three CALO email datasets:
  - acheyer: 664 messages, 38 folders
  - mgervasio: 777 messages, 15 folders
  - mgondek: 297 messages, 14 folders

- Two Enron email datasets:
  - kitchen-l: 4015 messages, 47 folders
  - sanders-r: 1188 messages, 30 folders

- The 20 Newsgroups: 19997 messages
## Results

<table>
<thead>
<tr>
<th></th>
<th>Slonim et al.</th>
<th>Dhillon et al.</th>
<th>2-way MDC</th>
<th>3-way MDC</th>
<th>4-way MDC</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>acht</strong></td>
<td>44.7±0.6</td>
<td>47.0±0.2</td>
<td>48.1±0.7</td>
<td>50.5±0.4</td>
<td>52.1±0.8</td>
<td>65.8±2.9</td>
</tr>
<tr>
<td><strong>mgervasio</strong></td>
<td>40.2±2.3</td>
<td>36.6±1.6</td>
<td>44.9±1.2</td>
<td>48.6±0.8</td>
<td>54.2±0.6</td>
<td>77.6±1.0</td>
</tr>
<tr>
<td><strong>mgondek</strong></td>
<td>62.1±1.4</td>
<td>69.5±1.6</td>
<td>77.1±1.4</td>
<td>80.8±1.2</td>
<td>81.6±1.0</td>
<td>92.6±0.8</td>
</tr>
<tr>
<td><strong>kitchen-l</strong></td>
<td>33.2±0.5</td>
<td>33.0±0.3</td>
<td>41.9±0.7</td>
<td>38.5±0.2</td>
<td></td>
<td>73.1±1.2</td>
</tr>
<tr>
<td><strong>sanders-r</strong></td>
<td>64.8±0.4</td>
<td>59.3±1.2</td>
<td>67.7±0.3</td>
<td>67.1±0.8</td>
<td></td>
<td>87.6±1.0</td>
</tr>
<tr>
<td><strong>20NG</strong></td>
<td>61.0±0.7</td>
<td>57.7±0.2</td>
<td>71.8±0.7</td>
<td></td>
<td></td>
<td>91.3±0.3</td>
</tr>
</tbody>
</table>
Improvement over the baseline

<table>
<thead>
<tr>
<th></th>
<th>Slonim et al.</th>
<th>Dhillon et al.</th>
<th>2-way MDC</th>
<th>3-way MDC</th>
<th>4-way MDC</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>acheyer</td>
<td></td>
<td>47.0±0.2</td>
<td>+1.1</td>
<td>+3.5</td>
<td>+5.1</td>
<td></td>
</tr>
<tr>
<td>mgervasio</td>
<td>40.2±2.3</td>
<td></td>
<td>+2.7</td>
<td>+8.4</td>
<td>+14.0</td>
<td></td>
</tr>
<tr>
<td>mgondek</td>
<td></td>
<td>69.5±1.6</td>
<td>+7.6</td>
<td>+11.3</td>
<td>+12.1</td>
<td></td>
</tr>
<tr>
<td>kitchen-l</td>
<td>33.2±0.5</td>
<td></td>
<td>+8.7</td>
<td>+5.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sanders-r</td>
<td>64.8±0.4</td>
<td></td>
<td>+2.9</td>
<td>+2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20NG</td>
<td>61.0±0.7</td>
<td></td>
<td>+10.8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
More results

- 4-way: full graph $\geq$ original graph $\geq$
  - Some interactions are unnecessary
- 2-way: agglomerative MDC $\approx$ original MDC
  - But it is dramatically less efficient
    - Would run 300 times longer on 20NG
- Documents can be clustered top-down
  - And words bottom-up
  - Eventually close results
Even more results

- Scheduling matters
  - More splits usually improve the results
  - MDC runs 7X slower with split/merge ratio = 2

- Social network analysis
  - Goal: to cluster people in a social network
  - Tested on Melinda Gervasio’s email
    - She created 4 groups of correspondents
    - 62.3% accuracy with 4 clusters, 76.6% precision with 8 clusters
Even more results

- Scheduling matters
  - More splits usually improve the results
  - MDC runs 7X slower with split/merge ratio = 2

- Social network analysis
  - Goal: to cluster people in a social network
  - Tested on Melinda Gervasio’s email
    - She created 4 groups of correspondents
    - 62.3% accuracy with 4 clusters, 76.6% precision with 8 clusters
Discussion

- Improvement over Slonim et al.
  - Which is a 1-way clustering algorithm
  - Shows that **multi-modality** helps
- Improvement over Dhillon et al.
  - Which is a 2-way clustering algorithm
  - Shows that **hierarchical setup** helps
- MDC is an efficient method
  - Which allows exploring complex models
    - 3-way, 4-way etc.
Conclusion

- Unsupervised model without generative assumptions
- Exploit multiple views of your data
- Efficient algorithm
- Impressive empirical results 😊
Future work

- Inference of optimal schedule
- Inference on “optimal” number of clusters?
- Extend to semi-supervised setup
Future work

- Inference of optimal schedule
- Inference on “optimal” number of clusters?
- Extend to semi-supervised setup

MDC 0.1 can be downloaded from http://www.cs.umass.edu/~ronb/mdc.html

Thanks