Combinatorial Markov Random Fields and their Applications to Information Organization

Ron Bekkerman

Center for Intelligent Information Retrieval
Research overview

- Classification of documents represented as bags of their word clusters
  [Bekkerman, El-Yaniv, Tishby, and Winter, SIGIR 2001, JMLR 2003]
- Clustering words and “useful” bigrams
  [Bekkerman and Allan, UMass TR IR-408 2004]
Research overview

- Focused Web crawling for social network analysis
  
  [Culotta, Bekkerman, and McCallum, CEAS 2004]

- Disambiguation of people who have the same name
  
  [Bekkerman and McCallum, WWW 2005]

- Link analysis for clustering Web search results
  
  [Bekkerman, Zilberstein, and Allan, IJCAI 2007]

---

- Web mining
- Information retrieval
- Link analysis
- Person name disambiguation
- Focused crawl
- Multi-way distributional clustering algorithm
- Image clustering
- Clustering by genre
- Formal model
- Interactive clustering
Research overview

Supervised learning

- Feature clustering for text classification

Unsupervised learning

- One-class clustering
- Clustering by genre
- Clustering by genre

Web mining

- Formal model
- Person name disambiguation

Information retrieval

- Link analysis
- Focused crawl

Multi-modal clustering

- Multi-way distributional clustering algorithm
- Image clustering
- Formal model

Semi-supervised clustering

- Interactive clustering
Thesis contributions

- Proposed a new framework for ML modeling
  - Combinatorial Markov Random Fields (Comrafs)
- Constructed particular Comraf models for multi-modal learning problems:
  - (Unsupervised) clustering
  - Semi-supervised clustering
  - Transfer learning
  - Interactive clustering
  - One-class clustering
Thesis contributions (continued)

- Applied constructed models to real-world tasks:
  - Email clustering
  - Social network analysis
  - Web appearance disambiguation
  - Clustering scientific papers
  - Clustering documents by genre
  - Clustering documents by authors’ sentiment
  - Re-ranking Web information retrieval results
  - Detecting the topic of the week in a newswire stream
  - Clustering images with captions
Trinity of a Comraf model

- **Comraf graph**
  - Nodes are *combinatorial* random variables

- **Objective function factored over the Comraf graph**
  - We use information-theoretic objectives

- **Optimization procedure for the objective function**
  - We use an iterative method for traversing the graph
  - And local search at each node
Comrafs for multi-modal clustering

- Simultaneously constructing $N$ clusterings of $N$ data modalities
- Combinatorial random variables are defined over all possible clusterings of a modality
Objective function for multi-modal clustering

- Best clusterings maximize the objective:

\[ I(\tilde{A};\tilde{B}) + I(\tilde{B};\tilde{C}) + I(\tilde{B};\tilde{G}) + I(\tilde{A};\tilde{F}) + I(\tilde{A};\tilde{G}) + I(\tilde{F};\tilde{G}) + I(\tilde{C};\tilde{G}) + I(\tilde{G};\tilde{E}) + I(\tilde{F};\tilde{E}) + I(\tilde{G};\tilde{D}) + I(\tilde{C};\tilde{D}) + I(\tilde{D};\tilde{E}) \]

- A potential function is defined on every edge
- Potentials are Mutual Information (MI) between interacting clusterings
Optimization procedure for multi-modal clustering

- **Iterative Conditional Modes (ICM)**
  - Fix current values of all variables but one
  - Optimize this variable wrt its neighbors (i.e. its Markov blanket)
  - Fix its new value and move to another variable
  - Round-robin over the variables
Optimization procedure (continued)

- **Clique-wise optimization (CWO)**
  - Choose one clique, ignore all the rest
  - Optimize this clique
  - Fix its nodes’ values and move to another clique
  - Round-robin over the cliques
Local search at each node

- For each variable
- Start with some solution
  - Say, (0,0,0)
  - All data points are in cluster $C_0$
- Traverse the lattice
  - While maximizing the objective
Evaluation methodology

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

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

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

- We fix number of clusters = number of categories
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: 19,997 messages
Melinda Gervasio’s data (SRI)

[rnob@vinci6 mgervasio]$ l total 68
drwx------   2 ronb ciir 4096 Feb 15 2004 email-external/
drwx------   2 ronb ciir 4096 Feb 15 2004 email-internal/
drwx------   2 ronb ciir 4096 Feb 15 2004 learning/
drwx------   2 ronb ciir 4096 Feb 15 2004 lsi/
drwx------   2 ronb ciir 4096 Feb 15 2004 lsi-comm/
drwx------   2 ronb ciir 4096 Feb 15 2004 lsi-dev/
drwx------   2 ronb ciir 4096 Feb 15 2004 lsi-plan/
drwx------   2 ronb ciir 4096 Feb 15 2004 lsi-weekly/
drwx------   2 ronb ciir 4096 Feb 15 2004 metalearning/
drwx------   2 ronb ciir 4096 Feb 15 2004 ra/
drwx------   2 ronb ciir 4096 Feb 15 2004 research/
drwx------   2 ronb ciir 4096 Feb 15 2004 scenarios/
drwx------   2 ronb ciir 4096 Feb 15 2004 yltest/
Richard Sanders’ data (Enron)

drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 agency_com/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 annex_5/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 bastos/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 duke/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 ecogas/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 ees_neg_ctc/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 gleason_sound/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 heof_intrust/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 india/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 infineum/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 international/
drwxr-xr-x   2 ronb ciir 8192 Feb 15 2004 iso_pricecaps/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 kafus/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 metals/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 misc/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 monetization_el_paso/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 nsm/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 pacific_valour/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 pacific_virgo/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 pca/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 personal_addresses/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 private_folders_beeson/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 project_stanley/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 px/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 radack/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 recruiting/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 sempra/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 senator_dunn_inv/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 tenaska/
drwxr-xr-x   2 ronb ciir 4096 Feb 15 2004 tva/
## Clustering results

<table>
<thead>
<tr>
<th></th>
<th>Agglomer. IB</th>
<th>IT Co-clustering</th>
<th>Sequent. IB</th>
<th>LDA</th>
<th>2-way Comraf</th>
<th>SVM (superv.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>acheyer</td>
<td>36.4</td>
<td>46.1±0.3</td>
<td>47.0±0.5</td>
<td>44.3±0.4</td>
<td>47.8±0.4</td>
<td>65.8±2.9</td>
</tr>
<tr>
<td>mgervasio</td>
<td>30.9</td>
<td>34.2±0.5</td>
<td>35.1±0.6</td>
<td>38.5±0.4</td>
<td>42.4±0.4</td>
<td>77.6±1.0</td>
</tr>
<tr>
<td>mgondek</td>
<td>43.3</td>
<td>63.4±1.1</td>
<td>68.2±1.2</td>
<td>68.0±0.8</td>
<td>75.9±0.6</td>
<td>92.6±0.8</td>
</tr>
<tr>
<td>kitchen-l</td>
<td>31.0</td>
<td>31.8±0.2</td>
<td>34.6±0.5</td>
<td>36.7±0.3</td>
<td>42.4±0.6</td>
<td>73.1±1.2</td>
</tr>
<tr>
<td>sanders-r</td>
<td>48.8</td>
<td>60.2±0.4</td>
<td>63.1±0.6</td>
<td>63.8±0.4</td>
<td>67.4±0.3</td>
<td>87.6±1.0</td>
</tr>
<tr>
<td>20NG</td>
<td>26.5</td>
<td>57.7±0.2</td>
<td>61.0±0.7</td>
<td>56.7±0.6</td>
<td>69.5±0.7</td>
<td>91.3±0.3</td>
</tr>
</tbody>
</table>
## Clustering results (continued)

<table>
<thead>
<tr>
<th></th>
<th>2-way Comraf</th>
<th>3-way Comraf</th>
<th>4-way Comraf</th>
</tr>
</thead>
<tbody>
<tr>
<td>acheyer</td>
<td>47.8±0.4</td>
<td>49.1±0.4</td>
<td>50.2±0.6</td>
</tr>
<tr>
<td>mgervasio</td>
<td>42.4±0.4</td>
<td>52.4±0.7</td>
<td>54.1±0.5</td>
</tr>
<tr>
<td>mgondek</td>
<td>75.9±0.6</td>
<td>80.1±0.7</td>
<td>80.9±0.5</td>
</tr>
<tr>
<td>kitchen-l</td>
<td>42.4±0.6</td>
<td>40.2±0.3</td>
<td></td>
</tr>
<tr>
<td>sanders-r</td>
<td>67.4±0.3</td>
<td>69.0±0.4</td>
<td></td>
</tr>
<tr>
<td>20NG</td>
<td>69.5±0.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Clustering results (3-way clustering)

<table>
<thead>
<tr>
<th></th>
<th>ICM, tree graph</th>
<th>CWO, tree graph</th>
<th>ICM, full graph</th>
<th>CWO, full graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>acheyer</td>
<td>49.1±0.4</td>
<td>47.2±0.3</td>
<td>48.9±0.4</td>
<td>46.1±0.2</td>
</tr>
<tr>
<td>mgervasio</td>
<td>52.4±0.7</td>
<td>48.4±0.5</td>
<td>51.3±0.8</td>
<td>51.1±0.4</td>
</tr>
<tr>
<td>mgondek</td>
<td>80.1±0.7</td>
<td>76.1±1.2</td>
<td>79.1±0.4</td>
<td>72.2±1.1</td>
</tr>
<tr>
<td>kitchen-l</td>
<td>40.2±0.3</td>
<td>39.5±0.5</td>
<td>42.2±0.4</td>
<td></td>
</tr>
<tr>
<td>sanders-r</td>
<td>69.0±0.4</td>
<td>63.9±0.2</td>
<td>68.4±0.5</td>
<td>68.8±0.2</td>
</tr>
</tbody>
</table>
Choosing the best Comraf graph (on \textit{mgervasio})
Clustering scientific papers

- Papers / title words
- Papers / citations
- Papers / title words / citations
- Papers / title words / citations

<table>
<thead>
<tr>
<th>Papers / title words</th>
<th>Papers / citations</th>
<th>Papers / title words / citations</th>
<th>Papers / title words / citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>38.8±0.5</td>
<td>40.7±0.7</td>
<td>55.0±0.7</td>
<td>61.4±0.6</td>
</tr>
</tbody>
</table>
Multi-modal clustering of images

- Image collections are multi-modal:
  - *Images*
  - Their *colors*
  - *Regions*: rectangular segments of images
  - *Blobs*: clusters of image regions
  - *Texture*: Gabor features
  - *Words* in image captions

[Bekkerman and Jeon, CVPR 2007]
An important observation of image clustering

- Many modalities are dense enough
  - Such as *colors*: no need to cluster them
  - Even *caption words* may not be clustered

- We end up with one *target* node $G^c$
  - And *observed* nodes

- Observed nodes do not interact with each other
Comraf* models

- Comraf models of an asterisk topology
  - With observed nodes around the target node
- A general Comraf can be translated into a sequence of Comraf*

1. 

2. 

A general Comraf can be translated into a sequence of Comraf*.
Particular models for image clustering

A general Comraf model: images / words / colors / regions / texture

2-step Comraf* model: regions are clustered first, then images
Datasets

- **Corel**
  - A benchmark dataset for image processing
  - A subset of 4500 images, 50 categories

- **Israel Images**
  - Collected especially for this project
  - 1823 images, 11 categories
Clustering accuracy on Corel

- 46.6 ± 0.5%
- 55.3 ± 0.5%
- 60.1 ± 0.3%
- 61.2 ± 0.4%

k-means: 22%
Clustering accuracy on Israel Images

- 44.2 ± 1.0%
- 54.2 ± 0.9%
- 68.6 ± 1.0%
- 69.0 ± 0.6%

k-means: 22%
An efficient multi-modal clustering algorithm

- Complexity of MDC is $O(n^2 \log n)$
  - Not useful for large datasets
- **Rooted MDC** is its efficient implementation:
  - Choose $x\%$ of data uniformly at random
  - Cluster it using MDC
  - Each data point from the rest of $(100-x)\%$ data is assigned into one of the clusters such that the MDC’s objective is maximized:
    \[
    \arg \max_{\tilde{x}_1^c, \tilde{x}_2^c} I(\tilde{X}_1; \tilde{X}_2)
    \]
- Rooted MDC shows promising results on RCV1
  - 800,000 documents

[Bekkerman and Allan, in preparation]
Conclusion

- Comraf is a flexible framework for multi-modal learning, consisting of
  - A graphical representation
  - An information-theoretic objective
  - A combinatorial optimization method

- Modifying Comraf graphs leads to new models for
  - Semi-supervised clustering and transfer learning
  - Image clustering
  - Text clustering by genre, etc.

- Modifying objective → one-class clustering
- Modifying algorithm → interactive clustering