Practical Relational Community Generation

Bo Long    Zhongfei Zhang
Department of Computer Science
State University of New York at Binghamton

Philip S. Yu
IBM T.J. Watson Research Center
Outline

- Introduction
- Homogeneous relational data clustering
  - Graph partitioning
  - Link pattern based clustering
- Heterogeneous relational data clustering
  - Co-clustering
  - K-partite graph clustering
- General relational data clustering
  - Feature based approaches
  - Collective clustering approaches
- Conclusions
Three Types of Relational Data

- Homogeneous relational data
  - Relations between the objects of the same type

- Heterogeneous relational data
  - Relations between the objects of the different type

- General relational data
  - Homogeneous relations
  - Heterogeneous relations
  - Attributes
Homogeneous Relational Data

Relations between objects of the same types
Heterogeneous Relational Data

Relations between objects of \textit{different} types
General Relational Data

Homogeneous relations, Heterogeneous relations, Attributes

Authors

Papers

Words

machine

supervise

SVM

partitioning

clustering

svd

relational
Relational Data Clustering

- Local cluster patterns for each type of nodes
- Global structure: the relations between clusters of different types of objects

Authors

Papers

Words

- machine
- learning
- supervise
- graph
- SVM
- clustering
- path
-svd
- relational
- protein
- gene
- cell
Challenges!

- Data objects are not identically distributed:
  - Heterogeneous data objects (papers, authors).

- Data objects are not independent
  - Data objects are related to each other.

No IID assumption 😞
Outline

- Introduction
- Homogeneous relational data clustering
  - Graph partitioning
  - Link pattern based clustering
- Heterogeneous relational data clustering
  - Co-clustering
  - K-partite graph clustering
- General relational data clustering
  - Feature based approaches
  - Collective clustering approaches
- Conclusions
Homogeneous Graphs
Graph Partitioning

- **Objective**
  - “Cut” a homogeneous graph into $k$ disjoint sub-graphs by finding the best edge cuts

- **Applications**
  - Circuit partitioning,
  - VLSI design,
  - Task scheduling,
  - Social network analysis
Edge Cut Objectives

\[ G(\mathcal{V}, \mathcal{E}, \mathcal{S}) \quad \text{link}(A, B) = \sum_{i \in A, j \in B} S_{ij} \]

**Ratio Cut**

\[ \sum_{j=1}^{k} \frac{\text{links}(V_j, V \setminus V_j)}{|V_j|} \]

**Normalized Cut**

\[ \sum_{j=1}^{k} \frac{\text{links}(V_j, V \setminus V_j)}{\text{degree}(V_j)} \]

**Kernighan-Lin Cut**

\[ \sum_{j=1}^{k} \frac{\text{links}(V_j, V \setminus V_j)}{\text{degree}(V_j)} \quad \text{s.t.} \quad |V_j| = |V|/k \quad \forall j = 1, \ldots, k. \]
Graph Partitioning Approaches

- **Spectral approaches**
  - Using eigenvectors of the affinity matrix, or a matrix derived from the affinity matrix.

- **Multi-level approaches**
  - Involving coarsening and un-coarsening graphs level by level
Spectral Graph Partitioning

Three basic stages:

1. Pre-processing
   - Construct a matrix representation of the dataset.

2. Decomposition
   - Compute eigenvalues and eigenvectors of the matrix.
   - Map each object to a lower-dimensional representation based on one or more eigenvectors.

3. Post-processing
   - Assign points to two or more clusters, based on the new representation.
Intuition: Why Eigenvectors?
Theory: Why Eigenvectors?

Normalized Cut

$$\min \sum_{j=1}^{k} \frac{\text{links} \ (V_j, V \setminus V_j)}{\deg \ ree (V_j)} \iff \max \sum_{j=1}^{k} \frac{\text{links} \ (V_j, V_j)}{\deg \ ree (V_j)}$$

$$\iff \max \sum_{j=1}^{k} \frac{c_j^T S c_j}{c_j^T D c_j} \iff \max \ \text{trace} \ (\tilde{C}^T S \tilde{C})$$

Edge cut minimization $\iff$ Trace maximization
Why Eigenvectors?

Ratio Cut

\[
\min \sum_{j=1}^{k} \frac{\text{links}(V_j, V \setminus V_j)}{|V_j|} \iff \min \sum_{j=1}^{k} \frac{c_j^T (D - S)c_j}{c_j^T c_j} \iff \\
\min \text{trace}(\tilde{C}^T (D - S)\tilde{C}) \iff \max \text{trace}(\tilde{C}^T (I - D + S)\tilde{C})
\]

Ratio cut minimization \iff Trace maximization
Why Eigenvectors?

Relax the weighted indicator matrix to an arbitrary orthonormal matrix

\[
\begin{align*}
\max_{\tilde{\mathbf{C}}^T\tilde{\mathbf{C}}=\mathbf{I}} \quad \text{trace}(\tilde{\mathbf{C}}^T \mathbf{A} \tilde{\mathbf{C}})
\end{align*}
\]

Ky-Fan Theorem: the optimal solution is given by the eigenvectors corresponding to the k largest eigenvalues.
Post-processing Eigenvectors

Eigenvectors do not directly provide cluster membership

K-means is used to cluster eigenvectors to provide the cluster membership
Multilevel Graph Partitioning

Three basic stages:

1. Coarsening
   - Merge the nodes and links level by level to construct a sequence of smaller graphs
   - Approaches: maximal matching, random matching, etc.

2. Initial partitioning
   - Initial partitioning on the small, coarsest graph

3. Uncarsening
   - Project back to the original graph level by level
   - Refine the partitioning at each level.
Example

How to coarsen a graph using a maximal matching

\[ G = (N, E) \]

- \( E_m \) is shown in red
- Edge weights shown in blue
- Node weights are all one

\[ G_c = (N_c, E_c) \]

- \( N_c \) is shown in red
- Edge weights shown in blue
- Node weights shown in black

Courtesy of Kathy Yelick
Example (cont.)

Converting a coarse partition to a fine partition

Partition shown in green
Multilevel Partitioning Algorithms

- Multilevel Kernighan/Lin
  - G. Karypis and Vipin Kumar, 1998 (METIS)
  - ParMETIS - parallel version

- Multilevel Spectral Bisection
  - S. Barnard and H. Simon, 1993
  - Bruce Hendrickson and Robert Leland, 1994 (Chaco)
Outline

- Introduction
- Homogeneous relational data clustering
  - Graph partitioning
  - Link pattern based clustering
- Heterogeneous relational data clustering
  - Co-clustering
  - K-partite graph clustering
- General relational data clustering
  - Feature based approaches
  - Collective clustering approaches
- Conclusions
A link-pattern based cluster is a group of nodes which have the similar link patterns.
Applications

- Social network analysis
  - A Web community of “music fans” Web pages could link to the same set of music Web pages but sparsely linked to each other

- Bioinformatics
  - A group of proteins could interact with the same set of proteins but not interact with each other.
Statistical Block Modeling

- Basic Formulation

\[ P(S \mid Z, \theta) = \prod_{i,j} P(s_{ij} \mid z_i, z_j, \theta) \]
Hoff et al. 2002

\[ P(S \mid Z, X, \theta) = \prod_{i,j} P(s_{ij} \mid z_i, z_j, x_{ij}, \theta) \]
Kemp et al. 2004

\[
P(S \mid Z, \theta) = \prod_{i,j} P(s_{ij} \mid z_i, z_j, \theta)
\]

Generated by Chinese restaurant processes

\[
p(z_i = a \mid z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\alpha} & n_a > 0 \\ \frac{\alpha}{i-1+\alpha} & a \text{ is a new class} \end{cases}
\]
\[ P(S \mid Z, \theta, \alpha) = \prod_{i,j} P(s_{ij} \mid z_i, z_j, \theta, \alpha) \]
Block Patterns of Affinity Matrix

Strongly intra-connected cluster

Weakly intra-connected cluster
Symmetric Convex Coding

Long et al. ICML’07

\[
\min D(S, CBC^T)
\]
A Special Case

1. Using Euclidean distance function
2. \( B \) is restricted to be diagonal, i.e., \( rI \)

\[
\min \| S, C(rI)C^T \|^2 \iff \max \text{trace}(C^TSC)
\]

Edge cut objectives
SCC under Euclidean Distance

\[
\min_{C \in \mathbb{R}^{n \times k}_+} \|A - CBC^T\|^2 + \alpha \|C1 - 1\|^2.
\]

\[
C = \tilde{C} \odot \left( \frac{A\tilde{C}B + \frac{\alpha}{2}}{\tilde{C}B\tilde{C}^T\tilde{C}B + \frac{\alpha}{2}\tilde{C}E} \right)^{\frac{1}{4}}
\]

\[
B = \tilde{B} \odot \frac{C^TAC}{C^TC\tilde{B}C^TC}
\]

ICDM 2007
$C_{jh} = \tilde{C}_{jh} \left( \frac{\sum_i A_{ij} [\tilde{C}B]_{i1} + \alpha}{\sum_i [\tilde{C}B]_{i1} + \alpha \tilde{1}_{j}} \right)^{\frac{1}{2}}$

$B_{gh} = \tilde{B}_{gh} \frac{\sum_{ij} A_{ij} C_{ig} C_{jh}}{\sum_{ij} C_{ig} C_{jh}}$
Outline

- Introduction
- Homogeneous relational data clustering
  - Graph partitioning
  - Link pattern based clustering
- Heterogeneous relational data clustering
  - Co-clustering
  - K-partite graph clustering
- General relational data clustering
  - Feature based approaches
  - Collective clustering approaches
- Conclusions
Applications

Bi-partite heterogeneous relational data

- Web: documents - words.
- Collaborative filtering: users - movies.
- Bioinformatics: gene expressions - experimental conditions.
- Market basket data: customers - products
- .........
### Summary of Co-clustering Alg.

<table>
<thead>
<tr>
<th>Method</th>
<th>Publish</th>
<th>Cluster Model</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheng &amp; Church</td>
<td>ISMB 2000</td>
<td>Background + row effect + column effect</td>
<td>Minimize mean squared residue of biclusters</td>
</tr>
<tr>
<td>Getz et al. (CTWC)</td>
<td>PNAS 2000</td>
<td>Depending on plugin clustering algorithm</td>
<td>Depending on plugin clustering algorithm</td>
</tr>
<tr>
<td>Lazzeroni &amp; Owen (Plaid Models)</td>
<td>Bioinformatics 2000</td>
<td>Background + row effect + column effect</td>
<td>Minimize modeling error</td>
</tr>
<tr>
<td>Ben-Dor et al. (OPSM)</td>
<td>RECOMB 2002</td>
<td>All genes have the same order of expression values</td>
<td>Minimize the p-values of biclusters</td>
</tr>
<tr>
<td>Tanay et al. (SAMBA)</td>
<td>Bioinformatics 2002</td>
<td>Maximum bounded bipartite subgraph</td>
<td>Minimize the p-values of biclusters</td>
</tr>
<tr>
<td>Yang et al. (FLOC)</td>
<td>BIBE 2003</td>
<td>Background + row effect + column effect</td>
<td>Minimize mean squared residue of biclusters</td>
</tr>
<tr>
<td>Kluger et al. (Spectral)</td>
<td>Genome Res. 2003</td>
<td>Background × row effect × column effect</td>
<td>Finding checkerboard structures</td>
</tr>
</tbody>
</table>
## Summary of Co-clustering Alg.

<table>
<thead>
<tr>
<th>Method</th>
<th>Allow overlap?</th>
<th>Bicluster Discovery</th>
<th>Complexity</th>
<th>Testing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheng &amp; Church</td>
<td>Yes (rare in reality)</td>
<td>One at a time</td>
<td>O(MN) or O(MlogN)</td>
<td>Yeast (2884×17), lymphoma (4026×96)</td>
</tr>
<tr>
<td>Getz et al. (CTWC)</td>
<td>Yes</td>
<td>One set at a time</td>
<td>Exponential</td>
<td>Leukemia (1753×72), colon cancer (2000×62)</td>
</tr>
<tr>
<td>Lazzeroni &amp; Owen</td>
<td>Yes</td>
<td>One at a time</td>
<td>Polynomial</td>
<td>Food (961×6), forex (276×18), yeast (2467×79)</td>
</tr>
<tr>
<td>Lazzeroni &amp; Owen</td>
<td>Yes</td>
<td>One at a time</td>
<td>Polynomial</td>
<td></td>
</tr>
<tr>
<td>Ben-Dor et al.</td>
<td>Yes</td>
<td>All at the same time</td>
<td>O(NM^3l)</td>
<td>Breast tumor (3226×22)</td>
</tr>
<tr>
<td>Ben-Dor et al.</td>
<td>Yes</td>
<td>All at the same time</td>
<td>O(NM^3l)</td>
<td></td>
</tr>
<tr>
<td>Tanay et al.</td>
<td>Yes</td>
<td>All at the same time</td>
<td>O((N2^{d+1})^{log_{r+1}}r^{rd})</td>
<td>Lymphoma (4026×96), yeast (6200×515)</td>
</tr>
<tr>
<td>Yang et al.</td>
<td>Yes</td>
<td>All at the same time</td>
<td>O((N+M)^2kp)</td>
<td>Yeast (2884×17)</td>
</tr>
<tr>
<td>Kluger et al.</td>
<td>No</td>
<td>All at the same time</td>
<td>Polynomial</td>
<td>Lymphoma (1 rel., 1 abs.), leukemia, breast cell line, CNS</td>
</tr>
</tbody>
</table>
A bicluster is represented as the submatrix $A$ of the whole expression matrix (the involved rows and columns need not be contiguous in the original matrix).

Each entry $A_{ij}$ in the bicluster is the superposition (summation) of:

1. The background level
2. The row (gene) effect
3. The column (condition) effect

A dataset contains a number of biclusters, which are not necessarily disjoint.
Example:

<table>
<thead>
<tr>
<th>Back.: 5</th>
<th>Col 0: 1</th>
<th>Col 1: 3</th>
<th>Col 2: 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 0: 2</td>
<td>8</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Row 1: 4</td>
<td>10</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>Row 2: 1</td>
<td>7</td>
<td>9</td>
<td>8</td>
</tr>
</tbody>
</table>

Correlation between any two columns = correlation between any two rows = 1.

\[ a_{ij} = a_{iJ} + a_{Ij} - a_{IJ} \]

where \( a_{iJ} = \text{mean of row } i \), \( a_{Ij} = \text{mean of column } j \), \( a_{IJ} = \text{mean of } A \).

Biological meaning: the genes have the same (amount of) response to the conditions.
Information-Theoretic Co-clustering
(Dhillon et al, 2003)

- View “co-occurrence” matrix as a joint probability distribution over row & column random variables

- seek a “hard-clustering” of both rows and columns such that “information” in the compressed matrix is maximized.
Lemma: “Loss in mutual information” equals

\[ I(X,Y) - I(\hat{X},\hat{Y}) = KL(p(x,y) \| q(x,y)) \]

\[ = H(\hat{X},\hat{Y}) + H(X \| \hat{X}) + H(Y \| \hat{Y}) - H(X,Y) \]

- \( p \) is the input distribution
- \( q \) is an approximation to \( p \)

\[ q(x,y) = p(\hat{x},\hat{y})p(x \| \hat{x})p(y \| \hat{y}), x \in \hat{x}, y \in \hat{y} \]

- Can be shown that \( q(x,y) \) is a maximum entropy approximation subject to cluster constraints.
**Example**

\[
p(x | \hat{x}) \quad p(\hat{x}, \hat{y}) \quad p(y | \hat{y}) \quad q(x, y)
\]

\[
\begin{bmatrix}
.5 & 0 & 0 \\
.5 & 0 & 0 \\
0 & .5 & 0 \\
0 & .5 & 0 \\
0 & 0 & .5 \\
0 & 0 & .5 \\
\end{bmatrix}
\begin{bmatrix}
.3 & 0 \\
0 & .3 \\
.2 & .2 \\
\end{bmatrix}
\begin{bmatrix}
.36 & .36 & .28 & 0 & 0 & 0 \\
0 & 0 & 0 & .28 & .36 & .36 \\
\end{bmatrix}
\begin{bmatrix}
.05 & .05 & .05 & 0 & 0 & 0 \\
0 & 0 & 0 & .05 & .05 & .05 \\
0 & 0 & 0 & .05 & .05 & .05 \\
.04 & .04 & 0 & .04 & .04 & .04 \\
.04 & .04 & .04 & 0 & .04 & .04 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
.054 & .054 & .042 & 0 & 0 & 0 \\
.054 & .054 & .042 & 0 & 0 & 0 \\
0 & 0 & 0 & .042 & .054 & .054 \\
0 & 0 & 0 & .042 & .054 & .054 \\
.036 & .036 & .028 & .028 & .036 & .036 \\
.036 & .036 & .028 & .028 & .036 & .036 \\
.036 & .036 & .028 & .028 & .036 & .036 \\
\end{bmatrix}
\]

\[
p(x, y)
\]

ICDM 2007
Outline

- Introduction
- Homogeneous relational data clustering
  - Graph partitioning
  - Link pattern based clustering
- Heterogeneous relational data clustering
  - Co-clustering
  - K-partite graph clustering
- General relational data clustering
  - Feature based approaches
  - Collective clustering approaches
- Conclusions
K-partite Heterogeneous Relations

Tri-partite relations

Penta-partite relations
Consistent Co-Partitioning

Gao et al.  2005
Consistent Co-Partitioning (cont.)

\[ q = (x, y)^T \quad \text{and} \quad p = (y, z)^T \]

\[
\begin{cases}
\min \beta \frac{q^T L^{(1)} q}{q^T D^{(1)} q} + (1 - \beta) \frac{p^T L^{(2)} p}{p^T D^{(2)} p}, \\
\text{subject to} \quad q^T D^{(1)} e = 0, \quad q \neq 0 \\
\quad \quad \quad \quad \quad p^T D^{(2)} e = 0, \quad p \neq 0 \\
\quad \quad \quad \quad \quad 0 < \beta < 1
\end{cases}
\]
Consistent Co-Partitioning (cont.)

How about different number of clusters?
Relation Summary Network (RSN)

Long et al. 2006
RSN: Example

(a) (b)
Distance between \( G \) and \( G^s \)

\[
D(G, G^s) = \sum_{i,j} \sum_{v_i \sim v_j, v_{ih} \in V_i, v_{jkl} \in V_j} D(e(v_{ih}, v_{jkl}), e^s(s_{ip}, s_{jq})),
\]

where \( 1 \leq i, j \leq m, 1 \leq h \leq |V_i|, 1 \leq l \leq |V_j|, 1 \leq p \leq |S_i|, \) and \( 1 \leq q \leq |S_j| \).

\[(1)\]
Matrix Representation of K-partite Graphs

\[ A^{(12)} \]

\[ B^{(12)} \]

\[ C^{(1)} \]

\[ C^{(2)} \]
RSN Objective Function

\[
\begin{align*}
\text{arg min}_{G^s} \ & \ D(G, G^s) \ \Leftrightarrow \\
\text{arg min} \ & \ \sum_{C^i, B^{(ij)}} D(A^{(ij)}, C^{(i)} B^{(ij)} (C^{(j)})^T) \\
\text{subject to} \ & \ 1 \leq i < j \leq m
\end{align*}
\]
RSN with Bregman Divergences

Table 1: A list of Bregman divergences and the corresponding convex functions.
Main Theorem

Theorem 1. Assume that $D$ in Definition 2 is a Bregman Divergence $D_{\phi}$. If $\{C^{(i)}\}_{1 \leq i \leq m}$ and $\{B^{(ij)}\}_{1 \leq i < j \leq m}$ are the optimal solution to the minimization in Definition 2, then

$$(C^{(i)})^T (C^{(i)} B^{(ij)} (C^{(j)})^T - A^{(ij)}) C^{(j)} = 0$$

for $1 \leq i < j \leq m$. 

\[ (C^{(i)})^T (C^{(i)} B^{(ij)} (C^{(j)})^T - A^{(ij)}) C^{(j)} = 0 \]
Iterative Algorithm: Update C \(^{(i)}\)

\[ C^{(i)}_{h,l^*} = 1 \text{ for } l^* = \arg\min_l L_l \]

\[ e^*(v_{ih}, s_{il^*}) = 1 \text{ for } l^* = \arg\min_l D_{\phi}(G, G^i_l). \]
Iterative Algorithm: Update B

\[ B^{(ij)} = ((C^{(i)})^T C^{(i)})^{-1} (C^{(i)})^T A^{(ij)} C^{(j)} ((C^{(j)})^T C^{(j)})^{-1} \]

\[ e^s(s_{ip}, s_{jq}) = \frac{1}{|U| \times |Z|} \sum_{v_{ih} \in U, v_{jl} \in Z} e(v_{ih}, v_{jl}), \]
Experiments

- Four RSN-BD algorithms:
  - RSN-ED: Euclidean Distance—normal dist.
  - RSN-LL: Logistic Loss—Bernoulli dist.
  - RSN-GI: Generalized I-divergence—Poisson dist.
  - RSN-IS: Itakura-Saito—Exponential dist.

- Comparing algorithms:
  - K-Means with ED, LL, GI and IS.
  - Bipartite Spectral Graph Partitioning (BSGP)
  - Consistent Bipartite Graph Co-partitioning (CBGC)
Experiments on Bi-partite Graphs

NMI scores on real bipartite graphs

- RSN-ED
- KM-ED
- RSN-LL
- KM-LL
- RSN-GI
- KM-GI
- RSN-IS
- KM-IS
- BSGP

BP-NG1
BP-NG2
BP-NG3
Experiments on Tri-partite Graphs

NMI scores on tri-partite graphs

Normalized Mutual Information

TP-e  TP-TM 1  TP-TM 2
Outline

- Introduction
- Homogeneous relational data clustering
  - Graph partitioning
  - Link pattern based clustering
- Heterogeneous relational data clustering
  - Co-clustering
  - K-partite graph clustering
- General relational data clustering
  - Feature based approaches
  - Collective clustering approaches
- Conclusions
General Relational Data

Homogeneous relations, Heterogeneous relations, Attributes
Relational Database

**Attributes**

- **Professor**
  - Teaching-Ability

- **Course**
  - Difficulty

**Relations**

- **Classes**
- **Student**
  - Intelligence

- **Registration**
  - Grade
  - Satisfaction

- **Teach**
- **Take**
- **In**
Feature Based Approaches

- **RDBC** (Kirsten et al. 98)
  - RIBL (Relational Instance-Based Learning)
  - Relational distance based clustering (agglomerative)

- **First-order K-Means clustering** (Kirsten et al. 00)
  - Distance-based K-Means clustering

- **CrossClus** (Yin et al. 2005)
  - Choose cross-relational features in a relational data set with user guidance
**RIBL (Relational Instance-Based Learning)**

- To measure distance between objects $O_1$ and $O_2$, neighbor objects of $O_1$ and $O_2$ are also considered.

**Relational data**
- `member(person1 ; 45 ; male; 20 ; gold)`
- `member(person2 ; 30 ; female; 10 ; platinum)`
- `car(person1 ; wagon; 200 ; volkswagen)`
- `car(person1 ; sedan; 220 ; mercedesbenz)`
- `car(person2 ; roadster; 240 ; audi)`
- `car(person2 ; coupe; 260 ; bmw)`
- `house(person1 ; murgle; 1987 ; 560)`
- `house(person1 ; montecarlo; 1990 ; 210)`
- `house(person2 ; murgle; 1999 ; 430)`
- `district(montecarlo; famous; large; monaco)`
- `district(murgle; famous; small ; slovenia)`

**Neighbor data of level 2**

```
member(person1, 45, male, 20, gold)

<table>
<thead>
<tr>
<th>car(person1, wagon, 200, volkswagen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>car(person1, sedan, 220, mercedesbenz)</td>
</tr>
<tr>
<td>car(person2, roadster, 240, audi)</td>
</tr>
<tr>
<td>car(person2, coupe, 260, bmw)</td>
</tr>
<tr>
<td>house(person1, murgle, 1987, 560)</td>
</tr>
<tr>
<td>house(person1, montecarlo, 1990, 210)</td>
</tr>
<tr>
<td>district(murfyle, famous, small, slovenia)</td>
</tr>
<tr>
<td>district(montecarlo, famous, large, monaco)</td>
</tr>
</tbody>
</table>
```
RDBC (Kirsten et al. 98)

- Use distance measure of RIBL
- Agglomerative clustering approach
  - Every object is used as a cluster at beginning
  - Keep merging clusters that are most similar
CrossClus (Yin et al. 2005)
CrossClus (Yin et al. 2005) (Cont.)

- A multi-relational feature is defined by:
  - A join path. E.g., *Student* → *Register* → *OpenCourse* → *Course*
  - An attribute. E.g., *Course. area*
  - An aggregation operator. E.g., sum or average

- Use user’s guidance and a heuristic method to select features

- Use a *k*-medoids-based algorithm (CLARANS) for clustering
Feature Based Approaches

Summary

- Derive new features for each type of objects based on the related objects.
- Apply traditional clustering approaches to clustering each type of objects individually.
- Aggregation causes link information loss.
- Usually do not consider homogeneous relations.
Outline

- Introduction
- Homogeneous relational data clustering
  - Graph partitioning
  - Link pattern based clustering
- Heterogeneous relational data clustering
  - Co-clustering
  - K-partite graph clustering
- General relational data clustering
  - Feature based approaches
  - Collective clustering approaches
- Conclusions
Collective Clustering Approaches

- A probabilistic model to formulate a whole relational data set
- Simultaneously cluster different types of objects.
- Probabilistic Relational Model (PRM) (Getoor et al. 2001)
- Mixed Membership Relational Clustering (MMRC) (Long et al. 2007)
Probabilistic Relational Models

Professor
Teaching-Ability

Student
Intelligence

Course
Difficulty

Reg
Grade
Satisfaction

Professor

Student

http://robotics.stanford.edu/~koller/acl-talk-web_files/frame.htm

ICDM 2007
73
PRM for Clustering

Learn model using EM

Actor

Director

Movie

Gender

C

Genres

C

Rating

Year

#Votes

MPAA Rating

[Taskar, Segal, K., 2001]
PRM, Summary

- Aggregation

  ![Diagram of Actor and Movies with features f1, f2, f3]

- Feature Based
  - Cross relation features
  - Do not explicitly use relation (link) information
MMRC (Long et al. 2007)

- Capable of using all kinds of information
  - Attributes
  - Homogeneous relations
  - Heterogeneous relations

- A good generalization of all types of relational clustering
  - Graph partitioning, link pattern based clustering
  - Bi-partite and k-partite relational data clustering
Matrix representation of relational data

\[ \{ \{ F^{(i)} \}_{i=1}^{m}, \{ S^{(i)} \}_{i=1}^{m}, \{ R^{(ij)} \}_{i,j=1}^{m} \} \]

Type 1: Papers
Type 2: Authors
Type 3: Words
Parameters and Variables

Membership matrices

Parameters
\[ \Omega = \left\{ \{ \Lambda^{(j)} \}_{j=1}^m, \{ \Theta^{(j)} \}_{j=1}^m, \{ \Gamma^{(j)} \}_{j=1}^m, \{ \Upsilon^{(ij)} \}_{i,j=1}^m \right\} \]

Variables
\[ \Psi = \left\{ \{ C^{(j)} \}_{j=1}^m, \{ F^{(j)} \}_{j=1}^m, \{ S^{(j)} \}_{j=1}^m, \{ R^{(ij)} \}_{i,j=1}^m \right\} \]

Latent cluster indicator

\[ \Lambda^{(1)} = \begin{bmatrix} 0.7 & 0.3 \\ 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.1 & 0.9 \end{bmatrix} \]

\[ C^{(1)} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \]
Generative Process

1. For each object $x_p^{(j)}$
   - Sample $C_p^{(j)} \sim \text{Multinomial}(\Lambda_p^{(j)}, 1)$.
2. For each object $x_p^{(j)}$
   - Sample $F_p^{(j)} \sim P_r(F_p^{(j)} | \Theta^{(j)} C_p^{(j)})$.
3. For each pair of objects $x_p^{(j)}$ and $x_q^{(j)}$
   - Sample $S_{pq}^{(j)} \sim P_r(S_{pq}^{(j)} | (C_p^{(j)})^T \Gamma^{(j)} C_q^{(j)})$.
4. For each pair of objects $x_p^{(i)}$ and $x_q^{(j)}$
   - Sample $R_{pq}^{(ij)} \sim P_r(R_{pq}^{(ij)} | (C_p^{(i)})^T Y^{(ij)} C_q^{(j)})$.

$\Gamma^{(13)} = \begin{bmatrix} 0.8 & w_1 \\ 0.1 & 0.7 \end{bmatrix}$
Likelihood Function

\[ Pr(\Psi|\Omega) = \prod_{j=1}^{m} Pr(C^{(j)}|A^{(j)}) \prod_{j=1}^{m} Pr(F^{(j)}|\Theta^{(j)}C^{(j)}) \]
\[ \quad \prod_{j=1}^{m} Pr(S^{(j)}|(C^{(j)})^T\Gamma^{(j)}C^{(j)}) \]
\[ \prod_{i=1}^{m} \prod_{j=1}^{m} Pr(R^{(ij)}|(C^{(i)})^T\Theta^{(ij)}C^{(j)}) \]

1. MMRC with exponential family distributions
2. Monte Carlo EM algorithm
3. Hard version and soft version algorithms
Experiments

Homogeneous data

Bi-partite data

Tri-partite data
Conclusions

- Relational data clustering has wide applications
- Homogeneous relational data clustering
  - Graph partitioning
  - Social network analysis
- Heterogeneous relational data clustering
  - Bi-partite and k-partite graph clustering
- General relational data clustering
  - Find hidden patterns from any relational database
References (1)


E. Erosheva, S. Fienberg, and J. LaFerty. Mixed membership models of scientific publications. In NAS.


References (3, incomplete)

- X. Yin, J. Han, and P. Yu. Cross-relational clustering with user's guidance. In *KDD 2005*.
- Bo Long, Zhongfei Zhang, Philip S. Yu A probabilistic framework for relational clustering, *KDD’07*.