Exam III Syllabus

Clustering

- K-means, EM
- Hierarchical
- Density based
  - DBSCAN
  - DENCLUE
  - Density Estimation
  - Density gradient/attractors

- Spectral/Graph clustering
  - Graph Cuts — Laplacian, modularity
  - MCL (Markov)

Evaluation

- External
- Internal
- Relative

Frequent Pattern Mining

- Stemsets & Association Rule
- Sequence & substrings

- Graphs
  - graph isomorphism (duplicate elimination)
  - subgraph isomorphism (frequency counting) \( \text{NP-complete} \)
- Maximal & closed patterns
Candidate generation without duplicates (Prefix based extension)

\[ A, B, C, D, E \]

\[ AC, AE, BC, CE \]

Subtree checking

Graphs

\[ G_1 \]

\[ G_2 \]

Isomorphic

\[ \overline{G}_1 \]

\[ \overline{G}_2 \]
Sets

A  B  C

Sequence

A → B → C  B → A → C

Order?

Trees: branching

XML documents

Phase transition

graph:

Cycle

Social network "substructure"
biological network

graph isomorphism
Frequent subgraphs

\[ \text{DB:} \]

\[ G_1: \]

\[ G_2: \]

\[ \text{MinSup} = 2 \]

\[ \text{edge\_growth} \]

\[ \Sigma = \{ a, b \} \]

\[ \begin{array}{c}
\text{2 Canonical} \\
(a\ b) < (b\ c)
\end{array} \]

\[ \begin{array}{c}
\text{1 not Canonical} \\
\text{not frequent}
\end{array} \]

\[ \text{Aut}(G) = \text{automorphism group of } G = \{ \text{all graphs isomorphic to } G \} \]

\[ \text{Canonical} = \min \{ \text{Aut}(G) \} \]

\[ G \rightarrow \text{Am I canonical?} \]
Transomism:

ode &
label
preserving
mapping

Forward extensions
(new vertex is added)

Backward extension
(only a new edge is added)
Given a frequent graph $G$, how to extend.

**Rightmost Path Extensions**

- Rightmost child (RMC)
- Path from RMC to the root is the RMP

$\text{RMP} = \{0, 1, 3\}$

1. **Forward extensions** allowed only from RMP nodes; ranked from bottom to top.

2. **Backward extensions**: from RMC to some node on RMP; ranked based on how close to the root the edge is.
3) all back edges come before forward edges

Impose a total order.

\[ G \]

\[ \text{DFS}(G) = (0, 1, a, a) \]
\[ (0, 2, a, b) \]

Is this canonical?
No.

Looking for a smaller code.

\[ G' \]

\[ G_1 \]

\[ G_2 \]

\[ G_3 \]

\[ G_4 \]

\[ G_5 \]
$\text{Span} \ (DB = \{ G_1, G_2, \ldots, G_n \})$

$F_i \ : \ \text{find edge tree in lex order}$

\[ \text{ac} \quad \text{bb} \quad \text{cc} \quad \text{aa} \quad \text{ac} \quad \text{bc} \]

for each edge graph $g_i \in F_i$

$\text{extend} \ (g_i)$

$\text{extend} \ (g_i, DB)$

If $\text{canonical} \ (g_i)$

then $\text{compute support of } g_i \ \text{in } DB$

if $\text{sup} (g_i) \geq \text{minsup}$

print $g_i$

Application:

local context $G(x)$

Large social network (labeled)

classify the node ($Y, N$)

$S \setminus \{v_1, v_2, \ldots, v_n\}$
\[ D_B = \{ G_1, G_2, \ldots, G_N \} \]

Find all frequent subgraphs (similar local context) and narrow the closed ones.

\[ C = \{ \text{closed frequent subgraphs} \} = \{ g_1, g_2, \ldots, g_m \} \]

\[ G(v_1) \]
\[ G(v_2) \]
\[ G(v_n) \]

\[ K(v_i, v_j) = \text{ # a common subgraph} \]

\[ \mathbf{K} = N \times N \]