Learning Global Probabilistic Models for Analyzing Large Structured and Semi-Structured Data

Chao Wang
Department of Computer Science and Engineering
The Ohio State University
wachao@cse.ohio-state.edu
Advisor: Prof. Srinivasan Parthasarathy

Copyright 2007, Data Mining Research Lab at the Ohio State University
Outline

• Introduction
• Background
• Our Previous Work
• Ongoing Work
• Concluding Remarks
Introduction

• Structured and semi-structured data
  – Transactional data. E.g., market basket data
  – Real graph data. E.g., co-authorship network, protein-protein interaction network
  – XML data

• Data modeling
  – Modeling interactions among domain entities
  – Our focus: Using local patterns to learn global probabilistic models

• Applications
  – Business intelligence
    • Recommender system/collaborative filtering
  – Graph analysis
    • Link prediction
    • Anomaly detection. E.g., anomalous link detection
  – Database processing
    • Selectivity estimation for query optimization
  – Many more …
Background

• Probabilistic graphical models
  – Undirected graphical models (Markov random field)
  – Directed graphical models (Bayesian network)

• Local patterns
  – Frequent itemsets
  – Frequent structural (sequence/tree/graph) patterns

• Using frequent itemsets to construct an MRF (First proposed by Pavlov et al. in 2000 for solving selectivity estimation problem)
  – View each $k$-itemset and its support as a constraint on the underlying data distribution
  – For a set of itemsets, a maximum entropy (ME) distribution satisfying all these constraints is selected as the estimated data distribution
    • This ME distribution specifies an MRF
Previous Work (Part 1) – Learning Approximate MRFs on Large Transactional Data

Mining frequent itemsets

<table>
<thead>
<tr>
<th>TID</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Exact MRF (model structure)

decompose

$k$-MinCut

augment

Interaction importance & treewidth-based scheme

Approximate MRF

derive

greedy inference
Experimental Results on Selectivity Estimation

Microsoft Web Anonymous Dataset
minSupp=20, |FI|=9901, tw=28

Varying tw (k = 25):

![Graphs showing estimation accuracy and response time](image)

- **Online time**: Increasing with query size and tw.
- **Offline time**: Increasing with tw, but relatively stable for different query sizes.

![Histogram showing average relative error](image)

- **Estimation accuracy**: Dependent on query size and tw, with lower error for smaller tw and query size.
There exist redundancy in a large collection of frequent itemsets

- Select only non-redundant patterns to learn probabilistic models
- Eliminate redundancy using MRFs
  - Relate to itemset summarization: To provide a more concise representation of a large collection of itemsets

1: Start from itemsets of size $k = 1$

2: Use $k$-itemsets to construct an MRF $M$ (learning)

3: Use $M$ to estimate the supports of $(k+1)$-itemsets (inference)
   - If estimation is accurate enough ($<\text{error\_threshold}$), do nothing
   - Else, augment $M$ using the corresponding patterns

4: Repeat in a level-wise fashion
Itemset Summarization Results

- Chess dataset (minSup=2000, |FI|=166581, |CFI|=68967, |NDFI|=1276|)

**Summarization quality (restoration error)**

**Summarization size**

**Summarization time**
Ongoing Work

• Improving model learning
  – Exploiting sampling methods (e.g., importance sampling) to learn truly large models
  – Combined with convex optimization techniques (e.g., CG, BFGS, L-BFGS, etc)

• Modeling data with incremental updates (e.g., Evolving real graphs) – preliminary results are promising
  – Incremental modeling is a special case
  – Link prediction
  – Novel pattern discovery
Concluding Remarks

- My thesis work is an *inter-disciplinary* effort that closely relates to various data mining/machine learning techniques, including:
  - Frequent pattern mining / Pattern post-processing
  - Statistical modeling / Probabilistic inference
  - Incremental data mining / Mining stream data
  - Graph mining / Social network analysis
  - Data characterization