Frequent Pattern Based Graph Classification

Peixiang Zhao
February 26th, 2009

Acknowledgement: Xifeng Yan etc. (KDD’08 tutorial)
Wei fan etc. (KDD’08)
Graph Classification

- **Structure-based Approach**
  - Local structures in a graph, e.g., neighbors surrounding a vertex, paths with fixed length

- **Kernel-based Approach**
  - Random walk (Gärtner ’02, Kashima et al. ’02, ICML’03, Mahé et al. ICML’04)
  - Optimal local assignment (Fröhlich et al. ICML’05)

- **Pattern-based Approach**
  - Subgraph patterns from domain knowledge or from graph mining
    - Model based Search Tree (Fan et al. KDD’08)
    - Boosting (Kudo et al. NIPS’04)
    - LAR-LASSO (Tsuda, ICML’07)
Pattern-Based Graph Classification

**Basic Idea**
- Transform each graph in the dataset into a feature vector

\[ G \rightarrow x = \{x_1, x_2, \ldots, x_n\} \]

where \(x_i\) is the frequency of the \(i\)-pattern in \(G\). Each vector is associated with a class label. *Classify these vectors in a vector space*

**Features**
- Sequential patterns *(De Raedt and Kramer IJCAI’01)*
- *(Closed)* Frequent subgraphs *(Deshpande et al, ICDM’03; Liu et al. SDM’05)*
- Coherent frequent subgraphs *(Huan et al. RECOMB’04)*
- Acyclic Subgraphs *(Wale and Karypis, technical report ’06)*
Discriminative Frequent Patterns as Features

- Frequent pattern is a good candidate for discriminative features (Hong etc. ICDE’07, 08)
- Conventional Procedure
  1. Mine frequent patterns (> minsup)
  2. Select most discriminative patterns
  3. Represent data in the feature space using such patterns
  4. Build classification models
Problems on the Conventional Procedure

• Feature Mining Step
  – # frequent patterns can be too large for effective feature selection
  – If the frequency of discriminative features is below $\textit{minsup}$, those features won’t even be considered

• Feature Evaluation Step
  – The discriminative power of patterns is evaluated against the complete dataset, but not on subset that other patterns do not predict well
  – The correlation among multiple features are not directly evaluated
MODEL-BASED SEARCH TREE – (1)

• Basic Idea
  – Partition the data in a top-down manner and construct the tree using the best feature at each step according to some criterion (e.g., information gain)
  – Partition the data set into two subsets, one containing this feature and the other does not
MODEL-BASED SEARCH TREE – (2)

• A top-down divide-and-conquer approach toward a combination of feature mining and search tree construction
  – As each node, a frequent-pattern algorithm is invoked only on the examples that the node is responsible of
  – The search and tree construction terminates when either
    • every example in the node belongs to the same class
    • the number of examples is less than a given threshold
 MODEL-BASED SEARCH TREE – (3)
Properties

• The algorithm could get over the barrier of explosive growth of frequent patterns and successfully identify discriminative patterns with very small support

• Bound on Number of Returned Features
  – $O(n)$: the number of examples $n$ for the model based search tree

• Non-overfitting

• Optimality under Exhaustive Search
Experimental Studies

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Proposed Method (≥ 10%)</th>
<th>SVM Bchmk 5% + fs</th>
<th>C4.5 Bchmk 5% + fs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M^b_T$</td>
<td>$DT$</td>
<td>$M^b_T$</td>
</tr>
<tr>
<td>NCI1</td>
<td>0.685</td>
<td>0.74</td>
<td>0.583</td>
</tr>
<tr>
<td>NCI33</td>
<td>0.743</td>
<td>0.745</td>
<td>0.512</td>
</tr>
<tr>
<td>NCI41</td>
<td>0.765</td>
<td>0.763</td>
<td>0.679</td>
</tr>
<tr>
<td>NCI47</td>
<td>0.708</td>
<td>0.727</td>
<td>0.501</td>
</tr>
<tr>
<td>NCI81</td>
<td>0.696</td>
<td>0.723</td>
<td>0.541</td>
</tr>
<tr>
<td>NCI83</td>
<td>0.734</td>
<td>0.722</td>
<td>0.633</td>
</tr>
<tr>
<td>NCI109</td>
<td>0.699</td>
<td>0.746</td>
<td>0.508</td>
</tr>
<tr>
<td>NCI123</td>
<td>0.667</td>
<td>0.679</td>
<td>0.517</td>
</tr>
<tr>
<td>NCI145</td>
<td>0.747</td>
<td>0.752</td>
<td>0.55</td>
</tr>
<tr>
<td>H1</td>
<td>0.675</td>
<td>0.667</td>
<td>0.632</td>
</tr>
<tr>
<td>H2</td>
<td>0.707</td>
<td>0.695</td>
<td>0.519</td>
</tr>
</tbody>
</table>

7 Wins, 4 losses
Conclusion

• Model-based Search Tree
  – Integrated feature mining and construction
  – Can mine extremely small support discriminative patterns
  – Not limited to one type of frequent pattern: plug-play

• More details
  – Wei Fan, Kun Zhang, Hong Cheng, Jing Gao, Jiawei Han, Philip Yu, and Olivier Verscheure, "Direct Mining of Discriminative and Essential Frequent Patterns via Model-based Search Tree", 2008 ACMKDD International Conference on Knowledge Discovery and Data Mining (KDD'08), Las Vegas, NE, USA.
  – The software package for Linux environment can be founded http://www1.cs.columbia.edu/~wfan/SOFTWARE/MBT_Itemset_Binary.tar.gz
Thank you