{ Mining, Sets, of, Patterns }

A tutorial at ECMLPKDD2010
September 20, 2010, Barcelona, Spain

by
B. Bringmann, S. Nijssen, N. Tatti, J. Vreeken, A. Zimmermann
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>00:00</td>
<td><em>Introduction</em> &lt;br&gt;Siegfried Nijssen</td>
</tr>
<tr>
<td>00:45</td>
<td><em>Unsupervised, explorative pattern set mining</em> &lt;br&gt;Jilles Vreeken</td>
</tr>
<tr>
<td>01:30</td>
<td>Break</td>
</tr>
<tr>
<td>02:00</td>
<td><em>Supervised pattern set mining</em> &lt;br&gt;Björn Bringmann</td>
</tr>
<tr>
<td>02:45</td>
<td>End</td>
</tr>
</tbody>
</table>
Practical information

- Even though we did our best to achieve otherwise:

WARNING

This TUTORIAL is neither complete nor unbiased

REFERENCES are not necessarily authoritative or complete

- More information (including references):
Part I
Introduction
Overview part I

- Patterns
- Definitions
- Motivations
- Dimensions
- Algorithms
- Pattern sets
Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms
Overview part I

- Patterns
- Definitions
- Motivations
- Dimensions
- Algorithms
What is a pattern?

Recurring structure

<table>
<thead>
<tr>
<th>Data</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Data" /></td>
<td><img src="image2" alt="Pattern" /></td>
</tr>
<tr>
<td><img src="image3" alt="Data" /></td>
<td><img src="image4" alt="Pattern" /></td>
</tr>
<tr>
<td><img src="image5" alt="Data" /></td>
<td><img src="image6" alt="Pattern" /></td>
</tr>
</tbody>
</table>
What is a pattern?

Data

Pattern

\[ y = x - 1 \]
What is a pattern?

Data

Pattern

\[ y = x - 1 \]
What is a pattern?

In this tutorial we are looking for

- Recurring structures ...
- ... in enumerable, discrete domains

Hence we do not consider a regression model to be a pattern...
Overview part I

Patterns

Definitions
Motivations
Dimensions
Algorithms
Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms
What is a pattern?

Example 1: Frequent Itemset in Market Basket Data
What is a pattern?

Example 1: Frequent Itemset in Market Basket Data

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
</tbody>
</table>

support( )=3
What is a pattern?

Example 2: Co-cluster in Gene Expression Data

Lyssiotis et al.
What is a pattern?

Example 3: Conjunctive Formula in UCI Data

- Petal length $\geq 2.0$
- and Petal width $\leq 0.5$
What is a pattern?

Example 4: Frequent Subgraph in Molecules
What is a pattern?

Recurring structure in enumerable, discrete domain

Enumerable, discrete domains:
- itemsets, graphs, sequences, trees, ...

Recurrence as determined by constraints:
- support constraint, size constraint, area constraint, ...
The problem: too many patterns
Too many patterns...

Solution 1: constraint-based mining

Solution 2: pattern set mining
Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms
Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms
Solution 1:

pattern constraints

Constraint on each pattern individually based on

- background knowledge
- condensed representations
- class labels
Constraints: background knowledge

- Support constraints
- Syntactical constraints
- Statistical constraints
  - difference with expectation
  - taxonomies

= diapers
Constraints:
condensed representations

If we pass a pattern through the data, we obtain another pattern
Constraints: condensed representations

- Closed patterns
  Pasquier et al.

- Free/generator patterns
  Pasquier et al.

- Maximal frequent patterns
  Bayardo

- Non-derivable patterns
  Calders et al.
Constraints: class labels

Mutagenic

Clean

VS
Constraints: class labels

GIVEN database $D$, target $c$, threshold $t$, class of patterns

FIND all patterns $p$ with $f(p,D,c) > t$
Constraints: class labels
Constraints: class labels

Many different names for this setting

<table>
<thead>
<tr>
<th>Pattern name</th>
<th>Typical measure</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emerging pattern</td>
<td>Growth rate</td>
<td>Dong et al.</td>
</tr>
<tr>
<td>Contrast set</td>
<td>Difference in rel support</td>
<td>Bay et al.</td>
</tr>
<tr>
<td>Correlated pattern</td>
<td>Chi2</td>
<td>Morishita et al.</td>
</tr>
<tr>
<td>Subgroup</td>
<td>Weighted relative accuracy</td>
<td>Kloesgen et al.</td>
</tr>
<tr>
<td>Discriminative pattern</td>
<td>Information gain</td>
<td>Cheng et al.</td>
</tr>
<tr>
<td>Class association rule</td>
<td>Confidence</td>
<td>Liu et al.</td>
</tr>
</tbody>
</table>

Novak, Webb and Lavrac
Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms
Overview part I

Patterns

- Definitions
- Motivations
- Dimensions
- Algorithms
How to find patterns?

In principle two ways:

• Greedy / heuristic
  + Fast
  - Overlooks solutions

• Complete search
  + Finds everything
  - Slower
How to find patterns?

Complete search under constraints often feasible

**GIVEN** database $D$, constraint $\varphi$ on $D$, class of patterns $C$

**FIND** all patterns $p$ in class $C$ satisfying $\varphi$

Key Observation:
(Anti-)monotonicity
Lots of solutions... what’s their problem?
Overview part I

Patterns

Motivations
Definition
Dimensions
Algorithms
Overview part I

- Motivations
- Definition
- Dimensions
- Algorithms

Pattern sets
The problem - complex pattern relationships

Unsupervised descriptive task
The problem - complex pattern relationships

Unsupervised descriptive task
The problem - complex pattern relationships

Unsupervised descriptive task
The problem - complex pattern relationships

Unsupervised descriptive task
The problem - complex pattern relationships

Unsupervised descriptive task
The problem - complex pattern relationships

Unsupervised descriptive task
The problem - complex pattern relationships

Unsupervised descriptive task
The problem - complex pattern relationships

Supervised predictive task
The problem - complex pattern relationships

Supervised predictive task

All patterns mined
The problem - complex pattern relationships

Supervised predictive task

<table>
<thead>
<tr>
<th>P1</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Pattern 1" /></td>
<td><img src="image2.png" alt="Pattern 2" /></td>
</tr>
</tbody>
</table>

All patterns mined
The problem - complex pattern relationships

Supervised predictive task

<table>
<thead>
<tr>
<th>P1</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Pattern 1" /></td>
<td><img src="image2.png" alt="Pattern 2" /></td>
</tr>
</tbody>
</table>

All patterns mined

32
Overview part I

- Motivations
- Definitions
- Dimensions
- Algorithms
Overview part I

- Motivations
- Definitions
- Dimensions
- Algorithms

Pattern sets
Pattern set mining

GIVEN a data mining task
FIND an interrelated set of patterns useful for this task
Overview part I

Motivations
Definition
Dimensions
Algorithms

Pattern sets
### Patterns vs Pattern sets

<table>
<thead>
<tr>
<th></th>
<th>Unsupervised</th>
<th>Supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pattern mining</strong></td>
<td><strong>No target relationships</strong></td>
<td><strong>Relevant</strong> to target relationships</td>
</tr>
<tr>
<td><strong>Pattern set mining</strong></td>
<td><strong>No target relationships</strong></td>
<td><strong>Relevant</strong> to target relationships</td>
</tr>
</tbody>
</table>

- **Part II**: Pattern mining
- **Part III**: Pattern set mining
Task dimensions

<table>
<thead>
<tr>
<th>descriptive</th>
<th>unsupervised</th>
<th>supervised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Association Analysis</td>
<td>Subgroup discovery</td>
</tr>
<tr>
<td></td>
<td>(Co-)Clustering</td>
<td>Exceptional model mining</td>
</tr>
<tr>
<td></td>
<td>Probabilistic models</td>
<td></td>
</tr>
<tr>
<td>predictive</td>
<td>Predictive clustering</td>
<td>Classification</td>
</tr>
<tr>
<td></td>
<td>Regression</td>
<td>Regression</td>
</tr>
</tbody>
</table>

part II

part III
Task dimensions

- Supervised vs unsupervised
- Predictive vs descriptive
Task dimensions

- Supervised vs unsupervised
- Predictive vs descriptive
- (Semi-)Structured data vs Binary data
Task dimensions

- Supervised vs unsupervised
- Predictive vs descriptive
- (Semi-)Structured data vs Binary data
Task dimensions

- Supervised vs unsupervised
- Predictive vs descriptive
- (Semi-)Structured data vs Binary data
Task dimensions

- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained
Task dimensions

- Supervised vs unsupervised
- Predictive vs descriptive
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained
Task dimensions

- **Supervised vs unsupervised**
- **Predictive vs descriptive**
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained
Task dimensions

- Supervised vs unsupervised
- Predictive vs descriptive
- (Semi-)Structured data vs Binary data
- Constrained vs Unconstrained
- Interpretable model vs Black box
Overview part I

- Motivations
- Definitions
- Dimensions
- Algorithms

Pattern sets
Overview part I

- Motivations
- Definitions
- Dimensions
- Algorithms

Pattern sets
How to find pattern sets?
How to find pattern sets?
How to find pattern sets?

Pattern set constraint

Model constraint

Optimisation Criteria

Mining Constraint

PS

Pattern Selection

Pattern

Pattern set constraint

PS

Model Induction

PS

Pattern Mining

DB

How to find pattern sets?
How to find pattern sets?

Pattern set constraint

Model constraint

DB -> Pattern Mining -> PS -> Pattern Selection -> PS -> Model Induction

Mining Constraint

Optimisation Criteria
How to find pattern sets?
How to find pattern sets?

Pattern set constraints

Model constraints

DB -> Pattern Mining -> PS -> Pattern Selection -> PS -> Model Induction -> M

Optimisation Criteria

Mining Constraint
How to find pattern sets?

Pattern set constraint

Model Independent Iterative Mining

Model constraint

Mining Constraint

Optimisation Criteria
How to find pattern sets?

1. Model Independent Iterative Mining
2. Model Independent Post Processing

Pattern set constraint

DB → Pattern Mining → PS → Pattern Selection → PS → Model Induction → M

Model constraint

Mining Constraint
Optimisation Criteria
How to find pattern sets?

Pattern set constraint

Model constraint

1. Model Independent Iterative Mining
2. Model Independent Post Processing
3. Model Dependent Iterative Mining

DB → Pattern Mining → PS → Pattern Selection → PS → Model Induction → M

Mining Constraint

Optimisation Criteria
How to find pattern sets?

Pattern set constraint

Model Independent
Iterative Mining

Pattern Mining

Model Independent
Post Processing

Pattern Selection

PS

PS

Model Induction

Model constraint

Miner

Optimisation

Criteria

Model Dependent
Iterative Mining

Pattern Dependent
Post Processing

DB
Pattern set = Feature set

<table>
<thead>
<tr>
<th></th>
<th><img src="image1" alt="Pattern" /></th>
<th><img src="image2" alt="Pattern" /></th>
<th><img src="image3" alt="Pattern" /></th>
<th><img src="image4" alt="Pattern" /></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5" alt="Pattern" /></td>
<td><img src="image6" alt="Pattern" /></td>
<td><img src="image7" alt="Pattern" /></td>
<td><img src="image8" alt="Pattern" /></td>
<td><img src="image9" alt="Pattern" /></td>
</tr>
<tr>
<td><img src="image10" alt="Pattern" /></td>
<td><img src="image11" alt="Pattern" /></td>
<td><img src="image12" alt="Pattern" /></td>
<td><img src="image13" alt="Pattern" /></td>
<td><img src="image14" alt="Pattern" /></td>
</tr>
</tbody>
</table>
Pattern set = Feature set

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Pattern1" /></td>
<td><img src="image2" alt="Pattern2" /></td>
<td><img src="image3" alt="Pattern3" /></td>
<td><img src="image4" alt="Pattern4" /></td>
<td><img src="image5" alt="Pattern5" /></td>
</tr>
<tr>
<td><img src="image6" alt="Feature1" /></td>
<td><img src="image7" alt="Feature2" /></td>
<td><img src="image8" alt="Feature3" /></td>
<td><img src="image9" alt="Feature4" /></td>
<td><img src="image10" alt="Feature5" /></td>
</tr>
<tr>
<td><img src="image11" alt="Feature6" /></td>
<td><img src="image12" alt="Feature7" /></td>
<td><img src="image13" alt="Feature8" /></td>
<td><img src="image14" alt="Feature9" /></td>
<td><img src="image15" alt="Feature10" /></td>
</tr>
<tr>
<td><img src="image16" alt="Feature11" /></td>
<td><img src="image17" alt="Feature12" /></td>
<td><img src="image18" alt="Feature13" /></td>
<td><img src="image19" alt="Feature14" /></td>
<td><img src="image20" alt="Feature15" /></td>
</tr>
<tr>
<td><img src="image21" alt="Feature16" /></td>
<td><img src="image22" alt="Feature17" /></td>
<td><img src="image23" alt="Feature18" /></td>
<td><img src="image24" alt="Feature19" /></td>
<td><img src="image25" alt="Feature20" /></td>
</tr>
</tbody>
</table>
Feature vs pattern selection

We know more about patterns

- constraints used
- generality relationships
Overview

Unsupervised Pattern set mining
Part II

Supervised Pattern set mining
Part III

How to score pattern sets
How to find pattern sets
End of Part I