ICDM workshop on:
Practical Theories for Exploratory Data Mining

Tijl De Bie
Kleanthis Nikolaos Kontonasios
Eirini Spyropoulou
Topic of the workshop: Exploratory data mining

What it is not:
- Data mining with a clear objective or utility function, motivated by clear goals
- E.g. prediction problems, reinforcement learning,...

What it is:
- Facilitating the search for interesting *nuggets of information* in data (here referred to as *patterns*)
- Giving serendipity a chance
- Exploratory data analysis (EDA), but allowing for greater complexity

Very much an art... Can we make it a science?
Components of Exploratory Data Mining

- Pattern syntax
  - What kinds of things are we looking for? What *form* should it have?
- Interestingness measures
  - Can we quantify what is more interesting?
- Efficient search
  - Assuming we can quantify it, how do we find it?
- Visualization and presentation
  - Are there principles for optimally presenting patterns of this syntax to the users?

(Evaluation,...)
Pattern* syntax

Examples

- Cluster, or clustering (yes – but what exactly…?)
- Tile or tiling of a binary database; an itemset being frequent
- Association between attributes and a target label (subgroup discovery)
- Association between entities in a graph or relational database (plenty of ways to do this)
- …

Need a language to describe pattern syntaxes that is both

- powerful and
- easy to use for non-experts

* ‘Pattern’ = ‘data mining result’ / ‘information nugget’
Interestingness of a pattern

**Objective** interestingness – most common approach
- Clustering (k-means objective, eigenvalue in spectral clustering, etc)
- Itemset mining (frequency and many, often more useful, alternatives)
- Subgroup discovery (e.g. Weighted Relative Accuracy, etc)
- ...

**Subjective** interestingness – takes user’s prior knowledge into account
- Approach based on statistical testing – how significant is the pattern given a reasonable background model?
- Information theoretic approaches – how informative is the pattern considering the prior beliefs of the data miner?
Effective search paradigms

- Pattern syntax can be anything depending on user needs and data type
- Interestingness depends on context
- We need search frameworks that are automated, flexible, and ‘rarely much less efficient’

Examples:
- Inductive databases
- Declarative data mining (using constraint programming / convex optimization)
- Exhaustive (e.g. levelwise) search (when applicable)
- Pattern sampling (when applicable)
- …
Presentation to the user and interaction

- Present patterns in *intuitive* ways
- Allowing for feedback and *interactivity*
- Facilitating the *navigation* of the pattern space to help find the unsuspected
- A *generic* solution is needed

Examples:
- Visual analytics
- Visually controllable data mining
(Partial) data mining frameworks

- Inductive databases
- Declarative data mining
- Search paradigms such as levelwise search, pattern sampling, etc
- Visual analytics and visually controllable data mining
- Minimum description length (MDL) principle (data mining as model selection?)
- Information theory more generally for formalizing (subjective) interestingness
- Statistical testing and pattern significance
- Probabilistic modeling and statistical relational learning
- …
Workshop schedule

08:30 - 09:00  Introduction
09:00 - 10:00  Keynote talk: Kathleen Marchal
Network-based data integration for computational systems biology

Coffee break

10:30 - 11:30  Keynote talk: Luc De Raedt
From inductive querying to declarative modeling for data mining

11:30 - 11:50  Wilhelmiina Hamalainen
Thorough analysis of log data with dependency rules: Practical solutions and theoretical challenges

11:50 - 12:10  Christiane Kamdem Kengne et al.
Enhancing the Analysis of Large Multimedia Applications Execution Traces with FrameMiner

12:10 - 12:30  Michael Nett et al.
Generalized Expansion Dimension

Lunch break
Workshop schedule

Lunch break

14:00 - 15:00  Keynote talk: Kai Puolamaki
               The use of randomization and statistical significance in data mining
15:00 - 15:20  Albrecht Zimmermann
               Generating Diverse Realistic Data Sets for Episode Mining
15:20 - 15:40  Shailesh Kumar et al.
               Logical Itemset Mining

Coffee break

16:00 - 17:00  Keynote talk: Pieter Adriaans
               Data mining, looking backward, looking forward
17:00 - 18:00  Pieter Adriaans, Luc De Raedt, Kathleen Marchal, Kai Puolamaki, Jilles Vreeken
               Panel discussion
Some discussion points

- The user should have full control over the pattern syntax
  - How do we achieve this?
- The user cannot be forgotten in determining which patterns are interesting
  - How do we achieve this?
- The user cannot be burdened with the design or choice of the search strategy
  - Does this mean search will not be efficient?
- Iterative exploration important
  - Which strategies allow for this?
- Intuitive presentation
  - To what extent is visualization still an art rather than a science?
Thanks

Jose Balcazar
Mario Boley
Jean-Francois Boulicaut
Arno Knobbe
Jefrey Lijffijt
Emmanuel Mueller
Siegfried Nijssen
Kai Puolamaki
Thomas Seidl
Arno Siebes
Padhraic Smyth
Nikolaj Tatti
Tim Van den Bulcke
Jilles Vreeken
Geoff Webb

Pieter Adriaans
Luc De Raedt
Kathleen Marchal
Kai Puolamaki
Jilles Vreeken
Siegfried Nijssen
Florian Mansmann