Generating Diverse Realistic Data Sets (for episode mining)

Workshop “Practical Theories of Data Mining” @ ICDM 2012

Albrecht Zimmermann, KU Leuven
My Motivation

• Involved in industry cooperation

• Time-stamped event data

\(\langle (E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots \rangle\)

• Approach: episode mining
  • e.g. sliding window, minimal occurrence

• Off-the-shelf miner
My Motivation

• Involved in industry cooperation

• Time-stamped event data

\[\{(E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots\}\]

• Approach: episode mining
  • e.g. sliding window, minimal occurrence
  
• Off-the-shelf miner
My Motivation

- Involved in industry cooperation
- Time-stamped event data

\[ \langle (E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots \rangle \]

- Approach: episode mining
  - e.g. sliding window, minimal occurrence
- Off-the-shelf miner

Monday 10 December 12
My Motivation

• Involved in industry cooperation

• Time-stamped event data
  \( \langle (E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots \rangle \)

• Approach: episode mining
  • e.g. sliding window, minimal occurrence

• Off-the-shelf miner
My Motivation

• Involved in industry cooperation
• Time-stamped event data

$$\langle (E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots \rangle$$

• Approach: episode mining
  • e.g. sliding window, minimal occurrence
• Off-the-shelf miner
My Motivation

• Involved in industry cooperation
• Time-stamped event data
  \((E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots\)
• Approach: episode mining
  • e.g. sliding window, minimal occurrence
• Off-the-shelf miner
My Motivation

• Involved in industry cooperation

• Time-stamped event data

\[ ((E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots) \]

• Approach: episode mining
  • e.g. sliding window, minimal occurrence

• Off-the-shelf miner
My Motivation

- Involved in industry cooperation
- Time-stamped event data
  \[
  \langle (E, 1), (A, 12), (B, 15), (C, 25), (D, 26), (A, 36), (B, 38), \ldots \rangle
  \]
- Approach: episode mining
  - e.g. sliding window, minimal occurrence
- Off-the-shelf miner
My Motivation

• Involved in industry cooperation
• Time-stamped event data:
  \[\{(E, 1), (A, 12), (B, 15), (A, 36), (B, 38), \ldots\}\]
• Approach: episode mining
  • e.g. sliding window, minimal occurrence
• Off-the-shelf miner

NO idea what to do with patterns!
Going to the literature

- Guidance which approach to use - none
- Significance measures - (almost) none
- Guidance where in the output relevant patterns are - (almost) none
- Guarantees that patterns are found at all - (almost) none

15 years of research
Why’s that?

- Few temporal (real-life) data sets
- Locked by NDAs
- Real-life data sets have no ground truth!
- Post-hoc evaluation by domain experts
- Opposed to a priori class labels
Why’s that?

- Few temporal (real-life) data sets
- Locked by NDAs
- Real-life data sets have no ground truth!
- Post-hoc evaluation by domain experts
- Opposed to a priori class labels
Why’s that?

- Few temporal (real-life) data sets
- Locked by NDAs
- Real-life data sets have no ground truth!
- Post-hoc evaluation by domain experts
- Opposed to a priori class labels
Straight-Up Solution

- Generate diverse artificial data w/known patterns
- Building on Laxman’s generator
- Extensively evaluate different techniques/measures
- Develop guidelines when methods expected to work
Straight-Up Solution

• Generate diverse artificial data w/known patterns
• Building on Laxman’s generator
• Extensively evaluate different techniques/measures
• Develop guidelines when methods expected to work

(Related episodes and HMMs)
Comparative Data Mining

A detour to knowledge discovery

1. Get hands on real life data
2. Generate artificial data w/same characteristics
3. Mine patterns on artificial & real life data
4. Use relationship known & mined patterns on artificial data to select patterns from real data
Laxman’s generator

- n sequential patterns
- length N
- alphabet size M
- length of data sequence
- noise probability p
- uniform distributions for noise/time stamps
Laxman’s generator

- n sequential patterns
- length N
- alphabet size M
- length of data sequence
- noise probability p
- uniform distributions for noise/time stamps

- n=2, \( p \in [0.2,0.5] \)
- fixed M
- no sharing/repetition of elements
- interleaved episodes
- embedded concurrently
What’s “realistic”?

- Time information matters
- Events might not be logged
- There might be several patterns
  - Differently likely
- Patterns might interleave/share events/repeat events
- Patterns might occur successively
- Not only uniform distributions
What’s “realistic”? 

- Time information matters
- Events might not be logged
- There might be several patterns
  - Differently likely
  - Patterns might interleave/share events/repeat events
- Patterns might occur successively
- Not only uniform distributions

This is anecdotal
What’s “realistic”?

- Time information matters
- Events might not be logged
- There might be several patterns
  - Differently likely
  - Patterns might interleave/share events/repeat events
- Patterns might occur successively
- Not only uniform distributions

This is anecdotal

Episodes probably time-constrained
Adding parameters

- Failure (to log) probability
- Maximal delays explicit
- Enforcement in episode
- Switches for sharing/repetition/interleaving/concurrency/weights
- Poisson distribution for noise
- (Mixture of) normal distribution(s) for delays
Different kinds of data

HMM-generated data

Number of Occurrences vs. Event

Monday 10 December 12
Different kinds of data

HMM-generated data

n=2, p=0.3
interleaved uniform noise
Different kinds of data

Data as in Tatti, Cule '11

Event types

Occurrence counts (log-scaled)

Monday 10 December 12
Different kinds of data

Data as in Tatti, Cule '11

Large M
p=0.38
uniform noise
Different kinds of data

![Bar chart showing occurrence counts for different event types in real-life data.](image-url)
Different kinds of data

Real-life data

Occurrence Counts vs. Event types
Can I rebuild my data?

![Bar Chart]

**Real life-like data 01**

- **Number of Occurrences**
- **Event types**

Monday 10 December 12
Can I rebuild my data?

Real life-like data 01

n=2, N=4, p=0.7 uniform noise
Can I rebuild my data?

![Histogram of event types](image)

Real life-like data 02

- Number of Occurrences
- Event types
Can I rebuild my data?

- Real life-like data 02
- n=3, p=0.3
- different weights
- uniform noise

Monday 10 December 12
Can I rebuild my data?

Real life-like data 03

Event types

Number of Occurrences

Monday 10 December 12
Can I rebuild my data?

Real life-like data 03

n=3, p=0.3
different weights
Poisson noise
Harder for time

HMM-generated data

Number of Occurrences

Length of delay

Monday 10 December 12
Harder for time
Harder for time

![Real life data 01](image_url)
Harder for time
Experimental results

- Time constraint seems more important than matching semantic
- Best case: pattern within top-10
- Several patterns: very hard
- Real life data: patterns swamped by other stuff
Beyond episode mining

- Comparative data mining: general framework
- Currently working on itemset mining
- Extending to supervised settings:
  - Data harder to generate
  - Augment theoretical/UCI guarantees