Pattern-Based Classification of Demographic Sequences

Dmitry I. Ignatov\textsuperscript{1}, Danil Gizdatullin\textsuperscript{1}, Ekaterina Mitrofanova\textsuperscript{1}, Anna Muratova\textsuperscript{1}, Jaume Baixeries\textsuperscript{2}

\textsuperscript{1}National Research University Higher School of Economics, Moscow

\textsuperscript{2}Universitat Politècnica de Catalunya, Barcelona

Intelligent Data Processing 2016, Barcelona
Possible life events

- First job (job)
- The highest education degree is obtained (education)
- Leaving parents’ home (separation)
- First partner (partner)
- First marriage (marriage)
- First child birth (children)
- Break-up (parting)
- ... (divorce)
Data and problem statement
[Ignatov et al., 2015],[Blockeel et al., 2001]

Generation and Gender Survey (GGS): three waves panel data for 11 generations of Russian citizens starting from 30s

Binary classification
1545 men
3312 women

Examples of sequential patterns
- ⟨{education, separation}, {work}, {marriage}, {children}⟩(m)
- ⟨{work}, {marriage}, {children} {education}⟩(f)
- ⟨{partner}, {marriage, separation}, {children}⟩(f)
Basic definitions

$s = \langle s_1, \ldots, s_k \rangle$ is the **subsequence** of $s' = \langle s'_1, \ldots, s'_k \rangle$ ($s \preceq s'$) if $k \leq k'$ and there exist $1 \leq r_1 < r_2 < \ldots < r_k \leq k'$ such $s_j = s'_{r_j}$ for all $1 \leq j \leq k$.

**support**($s$, $D$) is the **support** of a sequence $s$ in $D$, i.e. the number of sequences in $D$ such that $s$ is their subsequence.

$$support(s, D) = |\{s'| s' \in D, s \preceq s'\}|$$

$s$ is a **frequent closed sequence** (sequential pattern) if there is no $s'$ such that $s \prec s'$ and

$$support(s, D) = support(s', D)$$
Let $D$ be a set of sequences:

**Table: Dataset $D$.**

<table>
<thead>
<tr>
<th></th>
<th>${a, b, c}{a, b}{b}$</th>
<th>${a}{a, c}{a}$</th>
<th>${a, b}{b, c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s_3$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- $I = \{a, b, c\}$ is the set of all items (atomic events)
- $\langle\{a, b\}\{b\}\rangle$ belongs to $s_1$ and $s_3$ but it is missing in $s_2$
- $support_D(\langle\{a, b\}\{b\}\rangle) = 2$
- $\{\langle\{a\}\rangle, \langle\{c\}\rangle, \langle\{a\}\{c\}\rangle, \langle\{a, b\}\{b\}\rangle, \langle\{a, c\}\{a\}\rangle\}$ is the set of closed sequences.
CAEP: Classification by Aggregating Emerging Patterns
G. Dong et al., 1999

Growth Rate

\[
growth\_rate_{D' \rightarrow D''}(X) = \begin{cases} 
\frac{\text{supp}_{D''}(X)}{\text{supp}_{D'}(X)} & \text{if supp}_{D'}(X) \neq 0 \\
0 & \text{if supp}_{D''}(X) = \text{supp}(X) = 0 \\
\infty & \text{if supp}_{D''}(X) \neq 0 \text{ and supp}_{D'}(X) = 0 
\end{cases}
\]

Class score

\[
score(s, C) = \sum_{e \subseteq s, e \in E(c)} \frac{growth\_rate_{C}(e)}{growth\_rate_{C}(e) + 1} \cdot \text{supp}_{C}(e)
\]
CAEP: Classification by Aggregating Emerging Patterns

Score normalization

\[
\text{normal score}(s, C) = \frac{\text{score}(s, C)}{\text{median} \left\{ \text{growth rate}_C(e_i) \right\}}
\]

Classification rule

\[
\text{class}(s) = \begin{cases} 
C_1, & \text{if } \text{normal score}(s, C_1) > \text{normal score}(s, C_2) \\
C_2, & \text{if } \text{normal score}(s, C_1) < \text{normal score}(s, C_2) \\
\text{undetermined}, & \text{if } \text{normal score}(s, C_1) = \text{normal score}(s, C_2)
\end{cases}
\]
Gapless prefix-based sequential patterns

- $s = \langle s_1, ..., s_k \rangle$ is a **gapless prefix-based subsequence** of $s' = \langle s'_1, ..., s'_k \rangle$ ($s^* = s'$) if $k \leq k'$ and $\forall i \in k': s_i = s'_i$.

- **Support of gapless prefix-based sequences**

  Let $T$ be a set of sequences.

  $$support(s, T) = \frac{|\{s'| s' \in T, s^* = s'\}|}{|T|}$$
Let $0 < \text{minSup} \leq 1$ be a minimal support parameter and $D$ is a set of sequences then searching for prefix-based gapless sequential patterns is the task of enumeration of all prefix-based gapless sequences $s$ such that $\text{support}(s, D) \geq \text{minSup}$. Every sequence $s$ with $\text{support}(s, D) \geq \text{minSup}$ is called a prefix-based gapless sequential pattern.

Prefix-based gapless sequential pattern (PGSP) $p$ is called closed if there is no PGSP $d$ of greater of equal support such that $d = p^*$. 
Example

Table: $D$ is a set of sequences.

<table>
<thead>
<tr>
<th>$s_1$</th>
<th>${a}{b}{d}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_2$</td>
<td>${a}{b}{c}$</td>
</tr>
<tr>
<td>$s_3$</td>
<td>${a, b}{b, c}$</td>
</tr>
</tbody>
</table>

$s = \langle\{a\}\{b\}\rangle$

- $I = \{a, b, c\}$ is the set of all items (atomic events)
- $s_1 = s^*; s_2 = s^*$
- $s_3 \neq s^*$
- $Supp_D(s) = \frac{2}{3}$
- $\langle\{a\}\{b\}\rangle$ is closed, $\langle\{a\}\rangle$ is not closed.
(S, (D, ⊓), δ) is a pattern structure

S is a set of objects, D is a set of their possible descriptions

δ(g) is the description of g from S

Galois connection is given by ◊ operator as follows:

\[ A^{◊} := \bigcap_{g \in A} δ(g) \text{ for } A \subseteq S \]

\[ d^{◊} := \{ s \in S | d \sqsubseteq δ(g) \} \text{ for } d \in D \]

For two sequences ⊓ may result in their largest common prefix subsequence
A pair \((A, d)\) is called a **pattern concept** of a pattern structure \((S, (D, \sqcap), \delta)\) if

1. \(A \subseteq S\)
2. \(d \in D\)
3. \(A^\diamond = d\)
4. \(d^\diamond = A\)
Pattern Structures

Example

\[ s_1 : \langle a, b, c \rangle \]
\[ s_2 : \langle a, b, c \rangle \]
\[ s_3 : \langle a, b, d \rangle \]

Tree

```
0
   \text{a(3)}
   \text{b(3)}
 /   \   \n\text{c(2)} \text{d(1)}
```

Pattern concepts (PCs)

\[
(\{s_1, s_2, s_3\}, \langle a, b \rangle); (\{s_1, s_2\}, \langle a, b, c \rangle)
\]
\[
(\{s_1\}, \langle a, b, c \rangle) \text{ is not a PC}; (\{s_3\}, \langle a, b, d \rangle)
\]
Pattern-based JSM-hypotheses


Positive, negative and undetermined pattern structures

\[ K_\oplus = (S_\oplus, (D, \sqcap), \delta_\oplus) \]
\[ K_\ominus = (S_\ominus, (D, \sqcap), \delta_\ominus) \]

There is a pattern structure of undetermined examples:

\[ K_\tau = (S_\tau, (D, \sqcap), \delta_\tau) \]

Hypothesis

A **hypothesis** is a pattern intent that belongs to examples from a fixed class only.

A pattern intent \( h \) is a positive hypothesis (dually for negative hypotheses) if

\[ \forall s \in S_\ominus (s \in S_\oplus) : h \nsubseteq s_\ominus (h \nsubseteq s_\oplus) \]
Hypotheses generation: An example

Sequential classification rules

\[ s_1 : \langle a, b, c \rangle - class \ 0 \]
\[ s_2 : \langle a, b, c \rangle - class \ 0 \]
\[ s_3 : \langle a, b, d \rangle - class \ 1 \]

Prefix-tree

```
0
   \|-- a(2; 1)
     \   \|-- b(2; 1)
          \   \|-- c(2; 0)
             \   \   \|-- d(0; 1)
```

Hypotheses

\[ \langle \{a\}, \{b\}, \{c\} \rangle \text{ is a hypothesis of class } 0 \]
\[ \langle \{a\}, \{b\}, \{d\} \rangle \text{ is a hypothesis of class } 1 \]
Classification via hypotheses

\[
\text{class}(g_\tau) = \begin{cases} 
\text{positive} & \text{if } \exists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \not\exists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) \\
\text{negative} & \text{if } \not\exists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \exists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) \\
\text{undetermined} & \text{if } \exists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \exists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) \\
\text{undetermined} & \text{if } \not\exists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \not\exists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) 
\end{cases}
\]
Emerging patterns based on pattern structures

Growth Rate

\[
GrowthRate(g, K_\oplus, K_\ominus) = \frac{Sup_{K_\oplus}(g)}{Sup_{K_\ominus}(g)}
\]

Emerging patterns

A pattern is called **emerging pattern** if its growth rate is greater than or equal to \( \Theta_{\text{min}} \)

\[
GrowthRate(g, K_\oplus, K_\ominus) > \Theta_{\text{min}}
\]
s is a new object

\[
\begin{align*}
\text{normal score}_\oplus(s) &= \frac{\sum_{p \in P_\oplus} \text{GrowthRate}(p, K_\oplus, K_\ominus)}{\text{median(GrowthRate}(P_\oplus))} : p \sqsubseteq s \\
\text{normal score}_\ominus(s) &= \frac{\sum_{p \in P_\ominus} \text{GrowthRate}(p, K_\ominus, K_\oplus)}{\text{median(GrowthRate}(P_\ominus))} : p \sqsubseteq s
\end{align*}
\]

Classification via emerging patterns

\[
\text{class}(s) = \begin{cases} 
\text{positive} \text{ if } \text{normal score}_\oplus(s) > \text{score}_\ominus(s) \\
\text{negative} \text{ if } \text{normal score}_\oplus(s) < \text{score}_\ominus(s) \\
\text{undetermined} \text{ if } \text{normal score}_\oplus(s) = \text{normal score}_\ominus(s)
\end{cases}
\]
Classification algorithm for gapless prefix-based sequential patterns

1. Build the prefix tree for the input sequences.
2. For each tree node calculate its Growth Rate.
3. For every new sequence traverse the tree and compute the Score for each class.
4. Compare the Score value for different classes and classify the new sequence.
Execution example

Input sequences

class 0 : \{⟨\{a\}\{b\}\{c\}⟩, ⟨\{b\}\{a\}\{c\}⟩, ⟨\{b\}\{a\}\{c\}⟩, ⟨\{b\}\{c\}⟩\}

class 1 : \{⟨\{a\}\{c\}\{b\}⟩, ⟨\{b\}\{c\}\{a\}⟩, ⟨\{b\}\{c\}\{a\}⟩\}

Prefix tree

```
0
  / \    /
 a(1; 1) b(2; 2)
  \   \   \   \   \   \   \   \
 b(1; 0) c(0; 1) a(2; 0) c(1; 2) c(1; 0) b(0; 1) c(2; 0) a(0; 2)
```
Counting Growth Rate

\[
\begin{align*}
0 & \quad a(0.75; 1.33) \quad b(0.75; 1.33) \\
& \quad \quad \quad b(\infty; 0) \quad c(0; \infty) \quad a(\infty; 0) \quad c(0.38; 2.67) \\
& \quad \quad \quad \quad \quad c(\infty; 0) \quad b(0; \infty) \quad c(\infty; 0) \quad a(0; \infty)
\end{align*}
\]

New sequence

\[
\langle \{b\}; \{c\}; \{a\} \rangle - ???
\]

\[
\begin{align*}
\text{Score}_0 &= 0 \\
\text{Score}_1 &= 2.67 + \infty = \infty
\end{align*}
\]
Comparison of closed and non-closed patterns

Figure: TPR vs FPR for closed and non-closed patterns
Experiments and results

Figure: TPR-FPR for classification via gapless prefix-based patterns
Interesting patterns (women)

\[
\langle \{\text{work, separation}\}, \{\text{marriage}\}, \{\text{children}\}, \{\text{education}\} \rangle, [\text{inf}, 0.006] \]

\[
\langle \{\text{separation, partner}\}, \{\text{marriage}\} \rangle, [\text{inf}, 0.006] \]

\[
\langle \{\text{work, separation}\}, \{\text{marriage}\}, \{\text{children}\} \rangle, [\text{inf}, 0.008] \]

\[
\langle \{\text{work, separation}\}, \{\text{marriage}\} \rangle, [\text{inf}, 0.009] \]
Interesting patterns (men)

\[
\langle \{education\}, \{marriage\}, \{work\}, \{children\}, \{separation\} \rangle, [10.6, 0.006]
\]

\[
\langle \{education\}, \{marriage\}, \{work\}, \{children\} \rangle, [12.7, 0.007]
\]

\[
\langle \{educ\}, \{work\}, \{part\}, \{mar\}, \{sep\}, \{ch\} \rangle, [10.6, 0.006]
\]
We have studied several pattern mining techniques for demographic sequences including pattern-based classification in particular.

We have fitted existing approaches for sequence mining of a special type (gapless and prefix-based ones).

The results for different demographic groups (classes) have been obtained and interpreted.

In particular, a classifier based on emerging sequences and pattern structures has been proposed.
Conclusion

Thank you!

Questions?