Pattern-Based Classification of Demographic Sequences

Dmitry I. Ignatov[1], Danil Gizdatullin[1], Ekaterina Mitrofanova[1], Anna Muratova[1], Jaume Baixeries[2]

[1]National Research University Higher School of Economics, Moscow

[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

- $\langle\{education, separation\}, \{work\}, \{marriage\}, \{children\}\rangle(m)$
- $\langle\{work\}, \{marriage\}, \{children\}\{education\}\rangle(f)$
- $\langle\{partner\}, \{marriage, separation\}, \{children\}\rangle(f)$

Basic definitions

Texbooks of Han et al., Zaki & Meira, Aggarwal et al., etc

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

$$support(s, D) = support(s', D)$$

Example

Let D be a set of sequences:

Table: Dataset D.

s_1	$\{a, b, c\}\{a, b\}\{b\}$
s_2	$\{a\}\{a, c\}\{a\}$
s_3	$\{a, b\}\{b, c\}$

- $I = \{a, b, c\}$ 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

$$growth\_rate_{D' \to D''}(X) = \begin{cases} \dfrac{supp_{D''}(X)}{supp_{D'}(X)} \text{ if } supp_{D'}(X) \neq 0 \\ 0 \text{ if } supp_{D''}(X) = supp(X) = 0 \\ \infty \text{ if } supp_{D''}(X) \neq 0 \text{ and } supp_{D'}(X) = 0 \end{cases}$$

Class score

$$score(s, C) = \sum_{e \subseteq s, e \in E(c)} \frac{growth\_rate_C(e)}{growth\_rate_C(e) + 1} \cdot supp_c(e)$$

CAEP: Classification by Aggregating Emerging Patterns

Score normalization

$$normal\_score(s, C) = \frac{score(s, C)}{median(\{growth\_rate_C(e_i)\})}$$

Classification rule

$$class(s) = \begin{cases} C_1, \text{ if } normal\_score(s, C_1) > normal\_score(s, C_2) \\ C_2, \text{ if } normal\_score(s, C_1) < normal\_score(s, C_2) \\ undetermined \text{ if } normal\_score(s, C_1) = normal\_score(s, C_2) \end{cases}$$

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.

$$support(s, T) = \frac{|\{s' | s' \in T, s* = s'\}|}{|T|}$$

Gapless sequential patterns

- Let $0 < minSup \leq 1$ 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 $support(s, D) \geq minSup$. Every sequence s with $support(s, D) \geq minSup$ 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 $d = p*$.

Gapless sequential patterns

Example

Table: D is a set of sequences.

s_1	$\{a\}\{b\}\{d\}$
s_2	$\{a\}\{b\}\{c\}$
s_3	$\{a,b\}\{b,c\}$

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

- $I = \{a, b, c\}$ 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.

Pattern Structures

Ganter & Kuznetsov, 2001

- $(S, (D, \sqcap), \delta)$ is a pattern structure
- S is a set of objects, D is a set of their their possible descriptions
- $\delta(g)$ is the description of g from S
- Galois connection is given by $\diamond$ operator as follows:

$$A^{\diamond} := \prod_{g \in A} \delta(g) \text{ for } A \subseteq S$$

$$d^{\diamond} := \{s \in S | d \sqsubseteq \delta(g)\} \text{ for } d \in D$$

- For two sequences $\sqcap$ may result in their largest common prefix subsequence

Pattern concepts

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

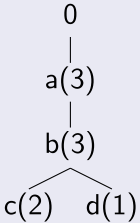
Example

$$s_1 : \langle a, b, c\rangle$$

$$s_2 : \langle a, b, c\rangle$$

$$s_3 : \langle a, b, d\rangle$$

Tree



Pattern concepts (PCs)

$$(\{s_1, s_2, s_3\}, \langle a, b\rangle); (\{s_1, s_2\}, \langle a, b, c\rangle)$$

$(\{s_1\}, \langle a, b, c\rangle)$ is not a PC; $(\{s_3\}, \langle a, b, d\rangle)$

Pattern-based JSM-hypotheses

[Finn, 1981], [Kuznetsov, 1993], [Ganter et al, 2004]

Positive, negative and undetermined pattern structures

$$\mathbb{K}_{\oplus} = (S_{\oplus}, (D, \sqcap), \delta_{\oplus})$$

$$\mathbb{K}_{\ominus} = (S_{\ominus}, (D, \sqcap), \delta_{\ominus})$$

There is a pattern structure of undetermined examples:

$$\mathbb{K}_{\tau} = (S_{\tau}, (D, \sqcap), \delta_{\tau})$$

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

$$\forall s \in S_{\ominus}(s \in S_{\oplus}) : h \not\sqsubseteq s^{\ominus}(h \not\sqsubseteq s^{\oplus})$$

Hypotheses generation: An example

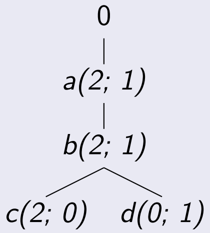
Sequential classification rules

$s_1 : \langle a, b, c \rangle$ - *class 0*

$s_2 : \langle a, b, c \rangle$ - *class 0*

$s_3 : \langle a, b, d \rangle$ - *class 1*

Prefix-tree



Hypotheses

$\langle \{a\}, \{b\}, \{c\} \rangle$ *is a hypothesis of class 0*

$\langle \{a\}, \{b\}, \{d\} \rangle$ *is a hypothesis of class 1*

Classification via hypotheses

$$class(g_\tau) = \begin{cases} \textit{positive} \text{ if } \exists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \nexists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) \\ \textit{negative} \text{ if } \nexists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \exists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) \\ \textit{undetermined} \text{ if } \exists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \exists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) \\ \textit{undetermined} \text{ if } \nexists h_\oplus, h_\oplus \sqsubseteq \delta(g_\tau) \text{ and } \nexists h_\ominus, h_\ominus \sqsubseteq \delta(g_\tau) \end{cases}$$

Emerging patterns based on pattern structures

Growth Rate

$$GrowthRate(g, \mathbb{K}_{\oplus}, \mathbb{K}_{\ominus}) = \frac{Sup_{\mathbb{K}_{\oplus}}(g)}{Sup_{\mathbb{K}_{\ominus}}(g)}$$

Emerging patterns

A pattern is called **emerging pattern** if its growth rate is greater than or equal to Θ_{min}

$$GrowthRate(g, \mathbb{K}_{\oplus}, \mathbb{K}_{\ominus}) > \Theta_{min}$$

Emerging patterns for classification

s is a new object

$$normal\_score_{\oplus}(s) = \frac{\sum_{p \in P_{\oplus}} GrowthRate(p, \mathbb{K}_{\oplus}, \mathbb{K}_{\ominus})}{median(GrowthRate(P_{\oplus}))} : p \sqsubseteq s$$

$$normal\_score_{\ominus}(s) = \frac{\sum_{p \in P_{\ominus}} GrowthRate(p, \mathbb{K}_{\ominus}, \mathbb{K}_{\oplus})}{median(GrowthRate(P_{\ominus}))} : p \sqsubseteq s$$

Classification via emerging patterns

$$class(s) = \begin{cases} positive \;\; if \; normal\_score_{\oplus}(s) > score_{\ominus}(s) \\ negative \;\; if \; normal\_score_{\oplus}(s) < score_{\ominus}(s) \\ undetermined \; if \; normal\_score_{\oplus}(s) = normal\_score_{\ominus}(s) \end{cases}$$

Classification algorithm for gapless prefix-based sequential patterns

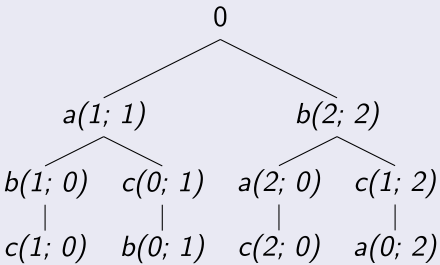
1. Build the prefix tree for the input sequences.
2. For each tree node calculate its Growth Rate.
3. For every new sequence traverse the tree and compute the Score for each class.
4. Compare the Score value for different classes and classify the new sequence.

Execution example

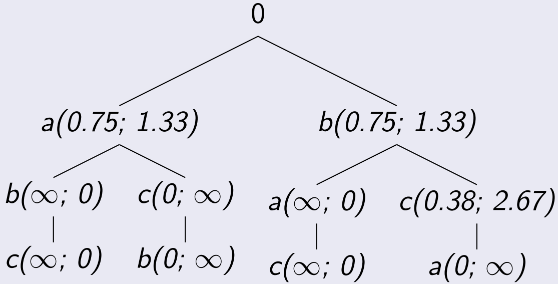
Input sequences

$class\_0 : \{\langle\{a\}\{b\}\{c\}\rangle, \langle\{b\}\{a\}\{c\}\rangle, \langle\{b\}\{a\}\{c\}\rangle, \langle\{b\}\{c\}\rangle\}$
$class\_1 : \{\langle\{a\}\{c\}\{b\}\rangle, \langle\{b\}\{c\}\{a\}\rangle, \langle\{b\}\{c\}\{a\}\rangle\}$

Prefix tree



Counting Growth Rate



New sequence

$$\langle\{b\};\{c\};\{a\}\rangle - ???$$

$$Score_0 = 0$$

$$Score_1 = 2.67 + \infty = \infty$$

Comparison of closed and non-closed patterns

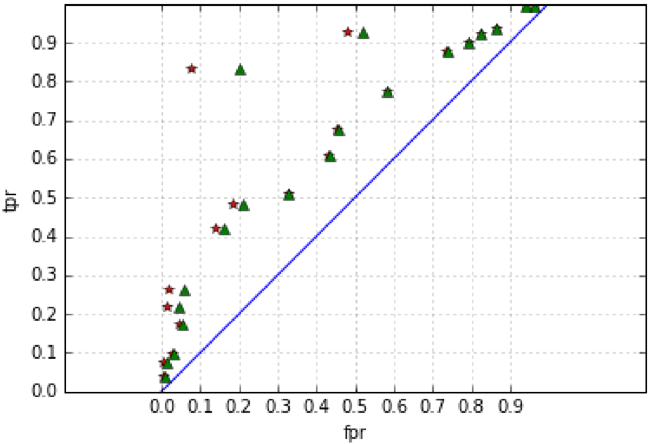


Figure: TPR vs FPR for closed and non-closed patterns

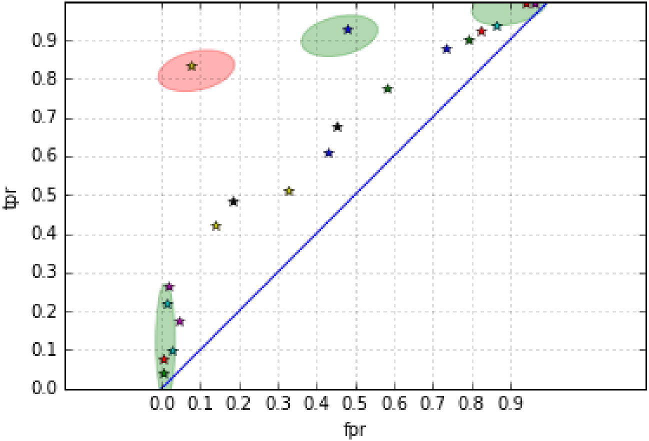


Figure: TPR-FPR for classification via gapless prefix-based patterns

Interesting patterns (women)

$$(\langle\{work, separation\}, \{marriage\}, \{children\}, \{education\}\rangle, [inf, 0.006])$$

$$(\langle\{separation, partner\}, \{marriage\}\rangle, [inf, 0.006])$$

$$(\langle\{work, separation\}, \{marriage\}, \{children\}\rangle, [inf, 0.008])$$

$$(\langle\{work, separation\}, \{marriage\}\rangle, [inf, 0.009])$$

Interesting patterns (men)

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

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

$$(\langle\{educ\},\{work\},\{part\},\{mar\},\{sep\},\{ch\}\rangle,[10.6,0.006])$$

Conclusion

1. We have studied several pattern mining techniques for demographic sequences including pattern-based classification in particular.
2. We have fitted existing approaches for sequence mining of a special type (gapless and prefix-based ones).
3. The results for different demographic groups (classes) have been obtained and interpreted.
4. In particular, a classifier based on emerging sequences and pattern structures has been proposed.

Conclusion

Thank you!

Questions?