Алгоритм классификация на основе эмерджентных частых последовательностей и узорных структур

Данил Гиздатуллин

5 октября 2016, г. Москва

Возможные события

- Первый опыт работы (job)
- Завершение наивысшей, на момент опроса ступени образования (education)
- Отъезд от родителей (separation)
- Первое партнерство (partner)
- Перое замужество/женитьба (marriage)
- Рождение первого ребенка (children)
- Разрыв отношений (parting)
- Развод (divorce)

Данные и задача

Рспределение по классам

1545 мужчины

3312 женщины

Примеры последовательностей

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

Определения

- **Подпоследовательность** $s = \langle s_1, ..., s_k \rangle$ – это подпоследовательность последовательности $s' = \langle s'_1, ..., s'_k \rangle$, обозначается как $s \preceq s'$, если $k \leq k'$ и существуют $1 \leq r_1 < r_2 < ... < r_k \leq k'$ такие, что $s_j \subseteq s'_{rj}$ для всех $1 \leq j \leq k$.
- **Поддержка** последовательности s на наборе последовательностей D обозначается как Support(s,D) – это количество последовательностей в наборе последовательностей D, для которых s является подпоследовательностью.

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

- **Частая замкнутая последовательность** – последовательный паттерн s такой, что не существует правильной суперпоследовательности s',такой что $s \prec s'$ и

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

Пример

Пусть дан следующий набор последовательностей D:

Таблица: Набор последовательностей D.

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

- $I = \{a, b, c\}$ - алфавит
- $\langle\{a, b\}\{b\}\rangle$ содержится одновременно в последовательности s_1 и последовательности s_3, но не в последовательности s_2
- $Support_D(\langle\{a, b\}\{b\}\rangle) = 2$
- $\{\langle\{a\}\rangle, \langle\{c\}\rangle, \langle\{a\}\{c\}\rangle, \langle\{a, b\}\{b\}\rangle, \langle\{a, c\}\{a\}\rangle\}$ - мн-во замкнутых последоваетльностей.

CAEP: Classification by Aggregating Emerging Patterns

Guozhu Dong, Xiuzhen Zhang, Limsoon Wong, and Jinyan L, 1999

Growth Rate

$$growth\_rate_{D' \to D''}(X) = \begin{cases} \dfrac{supp_{D''}(X)}{supp_{D'}(X)}, \text{ если } supp_{D'}(X) \neq 0 \\ 0, \text{supp}_{D''}(X) = supp(X) = 0 \\ \text{inf, если } supp_{D''}(X) \neq 0 \text{ и } supp_{D'}(X) = 0 \end{cases}$$

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

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

Неразрывные префиксные паттерны

- **Неразрывная префиксная подпоследовательность** – последовательности $s = \langle s_1, ..., s_k \rangle$ есть последовательность $s_1 = \langle s'_1, ..., s'_k \rangle$, обозначается это как $s = s_1*$, если $k \leq k'$ и $\forall i \in k' : s_i = s'_i$.
- **Поддержка в неразрывных префиксных подпоследовательностях**
  T - набор последовательностей

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

Неразрывные префиксные паттерны

- При данной минимальной поддержке $0 < minSup \leq 1$ задача **поиска префиксных неразрывных паттернов** – это задача поиска всех неразрывных префиксных последовательностей s таких, что $Support(s, D) \geq minSup$. Каждая последовательность s такая что $Support(s, D) \geq minSup$ называется **префиксным неразрывным паттерном**.
- Неразрывный префиксный паттерн p называется **замкнутым**, если не существует неразрывного префиксного паттерна d большей длины с такой же поддержкой и $d = p*$

Неразрывные префиксные паттерны

Пример

Таблица: Набор последовательностей D.

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} - алфавит
- $s_1 = s*; s_2 = s*$
- $s_3 \neq s*$
- $Supp_D(s) = \frac{2}{3}$
- $\langle\{a\}\{b\}\rangle$ - замкнутый префиксный паттерн, $\langle\{a\}\rangle$ - нет.

Пример

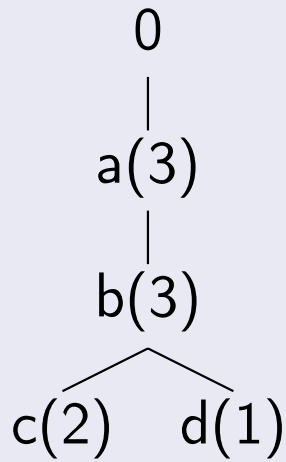
Пример хранения

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

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

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

Дерево



Пример получения гипотез

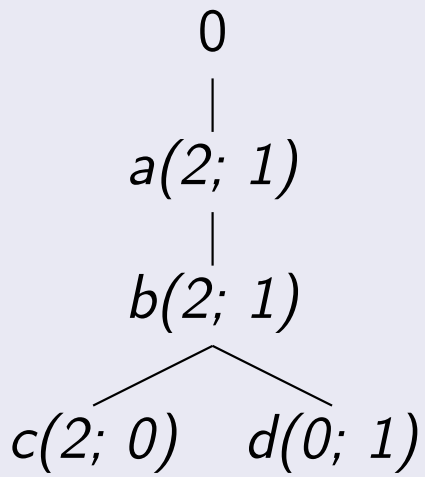
Пример последовательностей

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

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

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

Префиксное дерево



Гипотезы

$\langle \{a\}, \{b\}, \{c\} \rangle$ - *гипотеза класса 0*

$\langle \{a\}, \{b\}, \{d\} \rangle$ - *гипотеза класса 1*

Использование гипотез для классификации

$$\text{Class} = \begin{cases} +, \exists h_{\oplus}, h_{\oplus} \sqsubseteq g_{\theta} \text{ и } \nexists h_{\ominus}, h_{\ominus} \sqsubseteq g_{\theta} \\ -, \nexists h_{\oplus}, h_{\oplus} \sqsubseteq g_{\theta} \text{ и } \exists h_{\ominus}, h_{\ominus} \sqsubseteq g_{\theta} \\ \textit{неопределен}, \exists h_{\oplus}, h_{\oplus} \sqsubseteq g_{\theta} \text{ и } \exists h_{\ominus}, h_{\ominus} \sqsubseteq g_{\theta} \\ \textit{неопределен}, \nexists h_{\oplus}, h_{\oplus} \sqsubseteq g_{\theta} \text{ и } \nexists h_{\ominus}, h_{\ominus} \sqsubseteq g_{\theta} \end{cases}$$

Эмерджентные паттерны на узорных структурах

Growth Rate

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

Эмерджентные паттерны

Паттерн называется **эмерждентным**, если он удовлетворяет минимальному значению Growth Rate(Θ_{min}), заданному заранее

$$GrowthRate(g, K_{\oplus}, K_{\ominus}) > \Theta_{min}$$

Использование эмерджентных паттернов для классификации

s - новый объект

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

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

Классификация с использованием эмерджентных паттернов

$$\text{Class} = \begin{cases} +, \textit{если } \text{normal\_score}_{\oplus}(s) > score_{\ominus}(s) \\ -, \textit{если } \text{normal\_score}_{\oplus}(s) < score_{\ominus}(s) \\ \textit{неопределен}, \ \textit{если } \text{normal\_score}_{\oplus}(s) = normal\_score_{\ominus}(s) \end{cases}$$

Алгоритм поиска неразрывных префиксных паттернов и классификации

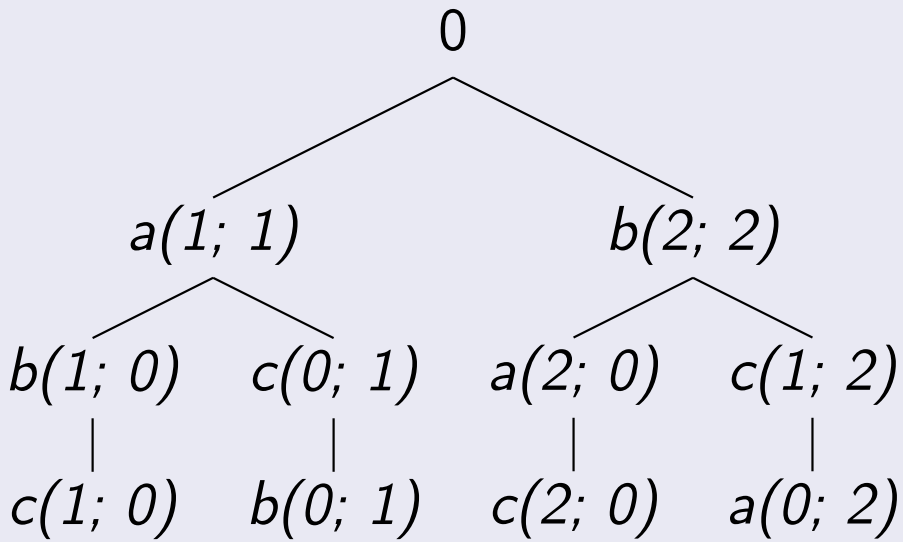
1. Строим префиксное дерево на основе всех наших последовательностей.
2. Для каждой вершины расчитываем Growth Rate.
3. Для классификации новой последовательности идем по нашему префиксному дереву и считаем Score для кажого класса.
4. Сравниваем значение Score для разных классов и классифицируем новую последовательность.

Пример работы алгоритма

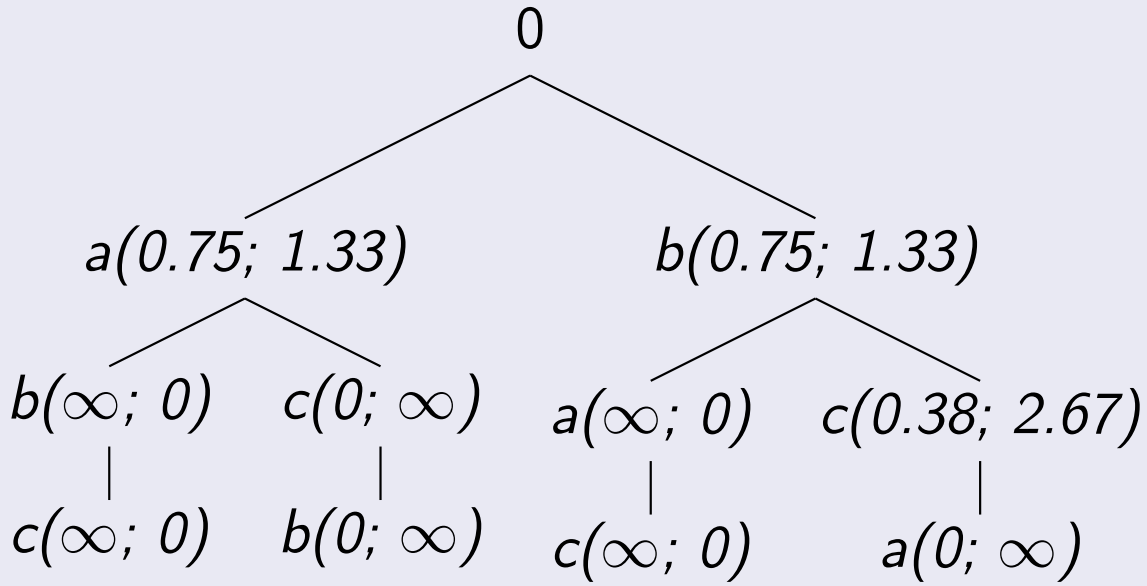
Последовательности данных

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

Префиксное дерево



Считаем Growth Rate



Новая последовательнотсть

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

$$Score_0 = 0$$

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

Эксперименты и результаты

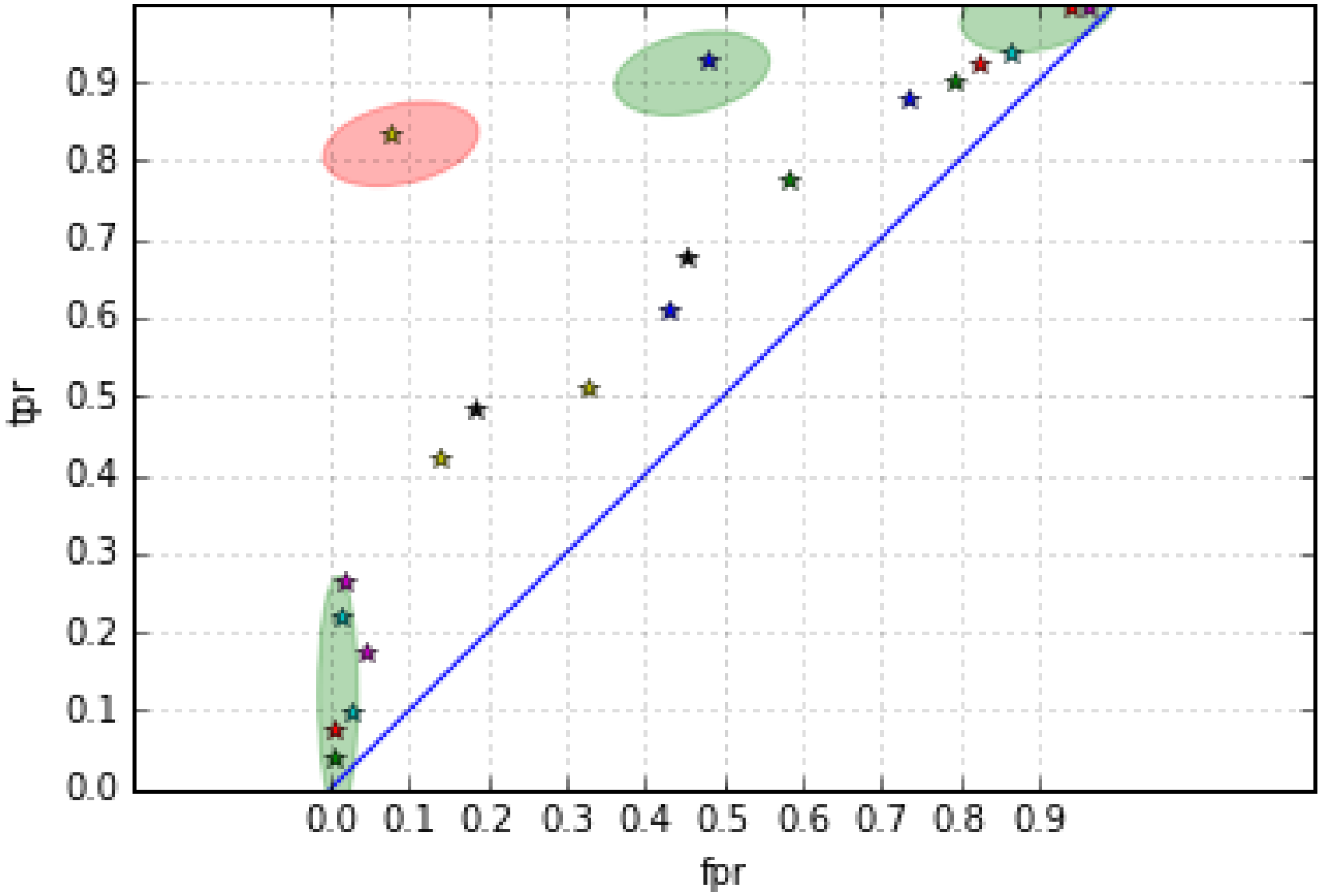


Рис.: TPR – FPR классификация на неразрывных префиксных паттернах

Паттерны для классификации (Женщины)

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

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

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

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

Паттерны для классификации (Мужчины)

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

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

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

Выводы

1. В данной работе было изучено применение методов анализа последовательностей к задачам демографического направления. В частности задаче поиска паттернов, характеризующих отдельные классы данных.
2. Был разработан и реализован новый метод анализа паттернов специального типа (неразрывных и префиксных).
3. Получены и проинтерпретированы паттерны поведения для разных демографических групп (классов).
4. Разработан и протестирован классификатор на языке Python на основе эмерджентных частых последовательностей и узорных структур