Analiza sekwencji

Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Zadania analizy sekwencji

Analiza sekwencji to dziedzina badań zajmująca się badaniem **uporządkowanych** w czasie lub logicznie powiązanych zdarzeń, stanów lub działań, w celu zrozumienia ich struktury, dynamiki oraz zależności między elementami sekwencji.

Celem analizy sekwencji jest identyfikacja regularności, wzorców, przejść oraz różnic między sekwencjami, a także modelowanie i przewidywanie ich przebiegu.

Przykłady zastosowań

- Analiza przebiegów procesów biologicznych (np. sekwencje genów, białek)
- Badanie ścieżek edukacyjnych, karier zawodowych
- Badanie zachowań użytkowników / klientów
- Analiza logów systemowych
- Analiza działań / zdarzeń w procesach technologicznych
- Modelowanie sekwencji zdarzeń w systemach społecznych

Pojęcia podrzędne

Odkrywanie/Wydobywanie wzorców sekwencyjnych (eksploracja sekwencji) – wyszukiwanie powtarzających się, uporządkowanych ciągów zdarzeń lub obiektów (w dużych zbiorach danych) z uwzględnieniem znaczenia relacji czasowych między elementami. Przykłady: wzorce zakupowe, kolejne kroki w obsłudze klienta, badanie postępów choroby / skuteczności leczenia.

Modelowanie sekwencji – technika do identyfikacji struktury i zależności między zdarzeniami oraz prognozowania kolejnych kroków. Przykłady: przetwarzanie języka naturalnego (NLP), rozumienie i generowanie tekstu (tłumaczenie, pisanie emaili).

Analiza predykcyjna sekwencji – prognozowanie przyszłych zdarzeń lub stanów na podstawie dotychczasowej sekwencji. Przykłady: przewidywanie awarii / katastrof, analiza zachowań, diagnostyka medyczna.

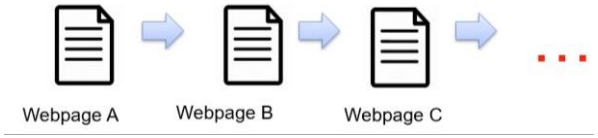
Pojęcie sekwencji

Sekwencja to uporządkowana lista symboli, zdarzeń, obiektów, przedmiotów (rzeczy). Porządek może być związany z linią czasu lub inną strukturą sekwencyjną. Można interpretować sekwencję jako listę koszyków.

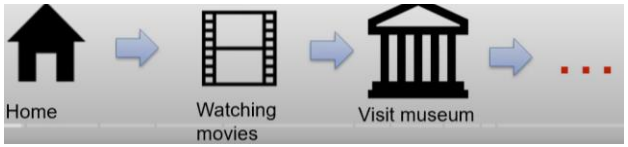
Przykłady związane z czasem

- Sekwencja zakupów

- Sekwencja kliknięć na strony internetowe



- Sekwencje aktywności osoby



Przykład związany z sekwencją wyrazów (konstrukcja zdania)

Definicje pomocnicze

Element, symbol (*item*), zdarzenie w analizie asocjacji lub sekwencji to jednostka danych, która jest przedmiotem analizy. Zbiór takich elementów oznacza się przez *I* .

Przykład:

I = {a ,b, c, d, e, f, g}, gdzie poszczególne symbole identyfikują towar: a = jabłko, b = chleb, c = ciasto, d = daktyle, e = jajka, f = rybę, g = winogrona.

Zbiór elementów (*itemset*) jest podzbiorem *I*. Nie może zawierać tego samego elementu więcej niż raz. O zbiorze zawierającym *k* elementów mówi się, że jest *k*-elementowy.

Zbiór elementów informuje jakie elementy występują razem w jakimś punkcie czasowym. Taki zbiór elementów może być interpretowany **jako koszyk.**

Przykłady:

{a, b, c} = {jabłko, chleb, ciasto} – zbiór zawierając 3 elementy (koszyk 3-elementowy)

{d, e} = {daktyle, jajka} – zbiór 2-elementowy

Formalna definicja sekwencji

Dyskretna sekwencja *S* jest uporządkowaną listą zbioru elementów I („koszyków”):

$S = \langle X_1, X_2, \ldots, X_n \rangle$, gdzie $X_j \subseteq$ I dla każdego $j \in \{1, 2, \ldots, n\}$

Przykłady (znaczenie symboli – por. poprzednie slajd):

< {*a*, *b*}, {c}> jest sekwencją zawierającą dwa zbiory elementów (dwa „koszyki”). Oznacza, że klient kupił jabłka i chleb w tym samym czasie (nie ma znaczenia porządek / kolejność elementów), a potem kupił ciasto.

< {*a*}, {*a*}, {c}> jest sekwencją zawierająca trzy zbiory elementów (trzy „koszyki”); klient kupił w oddzielnych transakcjach następujących po sobie jabłka, jabłka i ciasto.

Baza danych sekwencji to zbiór danych składający się z co najmniej jednej sekwencji (jednej lub więcej). Każda sekwencji ma identyfikator (odpowiednik klucza głównego w tabeli bazy danych).

IDS	Sekwencje
1	<{a}, {a, b, c }, {a, c}, {d}, {c, f}>
2	<{a, d}, {c}, { b, c }, {a, e}>
3	<{e, f}, {a, b}, {d, f}, {c}, {b}>
4	<{e}, {g}, {a, f}, {c}, {b}, {c}>

Koniec części I

Wzorce sekwencji

Wzorzec sekwencji to powtarzający się układ elementów (itemów) w określonej kolejności (występuje w bazie sekwencji częściej niż pewien ustalony próg).
Odsetek (lub liczba) wystąpień wzorca to jego **wsparcie**.

Przykład
Baza danych sekwencji składa się z czterech rekordów.

IDS	Sekwencje
1	<{a}, {a, b, c}, {a, c}, {d}, {c, f}>
2	<{a, d}, {c}, {b, c}, {a, e}>
3	<{e, f}, {a, b}, {d, f}, {c}, {b}>
4	<{e}, {g}, {a, f}, {c}, {b}, {c}>

- <{a}, {f}> jest wzorcem sekwencji o wsparciu 2/4=0,5 - pojawia się w 50% sekwencji.
- <{a}, {b, c}> jest wzorcem sekwencji o wsparciu 0,5.
- W czwartej sekwencji elementy {a, f} występują jednym koszyku; nie tworzą układu: „a następnie f” , nie wchodzą więc do drugiego wzorca sekwencji.

Reguła sekwencyjna

Reguła sekwencyjna to zależność postaci: **A → B**
Oznacza, że jeżeli w pewnej sekwencji wystąpi uporządkowany zestaw elementów (zdarzeń) **A**, **to po nim** wystąpi uporządkowany zestaw elementów (zdarzeń) **B.**

Poprzednik A i następnik B są też interpretowane jako wzorce sekwencji (podsekwencje).

Reguła sekwencyjna to odpowiednik reguły asocjacyjnej, ale z dodatkowym warunkiem kolejności zdarzeń w czasie i tym, że w poprzedniku występują nie elementy tylko wzorce sekwencji.

Przykład

Interpretacja reguły sekwencyjnej: <{a}, {b, c}> → <{d}, {e}>

- W sekwencji jest obecny wzorzec: „występuje element a, po którym występują łącznie elementy b i c".
- W sekwencji jest obecny wzorzec: „występuje element d, po którym występuje element e".
- Reguła sekwencyjna mówi, że jeżeli w sekwencji wystąpi wzorzec po a występują łącznie b i c to po nim nastąpi wzorzec po d wystąpi e.

Miary jakości reguły sekwencyjnej

W regule sekwencyjnej postaci: **A → B** są spełnione założenia:

- części z A i B nie mogą się pokrywać: $A \cap B = \emptyset$
- oba elementy reguły są niepuste, tzn. $A \neq \emptyset$ i $B \neq \emptyset$
- zdarzenia z A muszą wystąpić wcześniej niż zdarzenia z B

Wsparcie (*support*) to prawdopodobieństwo wystąpienia reguły w bazie sekwencji. Jest estymowane przez odsetek sekwencji w bazie danych sekwencji S, które zawierają całą regułę:

$$support(A \rightarrow B) = |\{s_i \in S: A \text{ i } B \text{ występują w } s_i\}| \,/\, |S|$$

Ufność (*confidence*, pewność, wiarygodność) to prawdopodobieństwo warunkowe wystąpienia reguły w zbiorze sekwencji zawierających wzorzec A. Jest estymowane przez stosunek liczby sekwencji zawierających całą regułę do liczby sekwencji zawierających jej poprzednik:

$$confidence(A \rightarrow B) = support(A \rightarrow B) \,/\, support(A)$$

Lift (współczynnik wzrostu, siła reguły) to wskaźnik informujący ile razy wystąpienie wzorca A zwiększa lub zmniejsza prawdopodobieństwo późniejszego wystąpienia wzorca B w stosunku do sytuacji, gdyby A i B występowały niezależnie w czasie:

$$lift(A \rightarrow B) = confidence(A \rightarrow B) \,/ support(B)$$
$$= support(A \rightarrow B) \,/\, (support(A) \cdot support(B))$$

Interpretacja miary lift

$$lift(A \rightarrow B) = support(A \rightarrow B) / (support(A) \cdot support(B))$$

support(A → B) proporcja sekwencji, w których A występuje przed B
support(A) i support(B) to proporcje sekwencji zawierających odpowiednio A i B (niezależnie od kolejności).

Lift w regule sekwencyjnej mierzy siłę zależności między A i B, uwzględniając następstwo zdarzeń w czasie („A przed B”).

Wartość lift	Interpretacja sekwencyjna
> 1	Wystąpienie A zwiększa prawdopodobieństwo, że B pojawi się później – A jest „prekursorem” B.
= 1	Wystąpienie A nie wpływa na prawdopodobieństwo późniejszego wystąpienia B – A i B są niezależne czasowo.
< 1	Wystąpienie A zmniejsza prawdopodobieństwo późniejszego wystąpienia B – może istnieć relacja negatywna lub wzajemne wykluczanie w czasie.

Wyliczenie miar dla przykładowej reguły sekwencyjnej

<{a}, {c}> → <{f}>

IDS	Sekwencje
1	<{a}, {a, b, c }, {a, c}, {d}, {c, f}>
2	<{a, d}, {c}, { b, c }, {a, e}>
3	<{e, f}, {a, b}, {d, f}, {c}, {b}>
4	<{e}, {g}, {a, f}, {c}, {b}, {c}>

1. Wyszukanie w sekwencji 1 wszystkich transakcji (koszyków), w których występuje a:
   Koszyk 1 → {a}; Koszyk 2 → {a, b, c}; Koszyk 3 → {a, c}
2. Dla każdego wystąpienia a w kolejnych koszykach sekwencji 1, poszukiwanie późniejszych koszyków, w których występuje c
   W sekwencji 1 → c w koszykach 2, 3, 5 → 3 wystąpienia
   W sekwencji 2 → c w koszykach 3, 5 → 2 wystąpienia
   W sekwencji 3 → c w koszyku 4 → 1 wystąpienie
   W sekwencji 4 → c w koszykach 4, 6 → 2 wystąpienia
3. Obliczenie sumarycznej liczby wystąpień: 3+2+1+2 = 8. Wzorzec ⟨{a}, {c}⟩ występuje 8 razy w badanej bazie sekwencji, jeśli są liczone wszystkie pary. W analizie sekwencji jeżeli wzorzec występuje co najmniej raz w sekwencji (raz lub więcej razy), to liczy się go jako jednokrotne wystąpienie w tej sekwencji.
4. We wszystkich sekwencjach wzorzec <**{a}, {c}**> występuje co najmniej 1.Stąd:
   Support(<**{a}, {c}**>) = 4/4 = 1.

Support(<**{a}, {c}> → <{f}**>) = 1 / 4 = 0,25
Confidence ((<**{a}, {c}> → <{f}**>)) = Support(<**{a}, {c}> → <{f}**>) / Support(<**{a}, {c}**>) = 0,25
Lift(Support(<**{a}, {c}> → <{f}**>) = Support(<**{a}, {c}> → <{f}**>) / (Support(<**{a}, {c}**>) · Support(<**{f}**>))
= 0,25 / (1 · 0,25) = 1

Częściowo uporządkowana reguła sekwencyjna

Częściowo uporządkowana reguła sekwencyjna to reguła postaci A → B, gdzie co najmniej jeden z jej składowych (poprzednik, następnik) jest nieuporządkowanym zbiorem elementów (itemsets). Reguła musi spełniać takie same warunki jak standardowa reguła sekwencyjna:

- części z A i B nie mogą się pokrywać: $A \cap B = \emptyset$
- oba elementy reguły są niepuste, tzn. $A \neq \emptyset$ i $B \neq \emptyset$
- zdarzenia z A muszą wystąpić wcześniej niż zdarzenia z B

W takiej regule część elementów ma określoną kolejność, a część może występować w dowolnym porządku względem siebie, przy zachowaniu globalnych relacji porządkowych wobec pozostałych elementów wzorca.

Miary jakości częściowo uporządkowanych reguł sekwencyjnych (wsparcia – *support*, ufności - *confidence* i siły reguły – *lift*) wyznacza się analogicznie jak w standardowych regułach sekwencyjnych.

Kilka standardowych reguł sekwencyjnych można przedstawić za pomocą jednej częściowo uporządkowanej reguły sekwencyjnej.

Częściowo uporządkowana reguła sekwencyjna - przykład

IDS	Sekwencje
1	⟨ {a, b}, {c}, {f}, {g}, {e} ⟩
2	⟨ {a, d}, {c}, {b}, {a, b, e, f} ⟩
3	⟨ {a}, {b}, {f}, {e} ⟩
4	⟨ {b}, {f, g} ⟩

W regule sekwencyjnej: {a, b} → {e, f} wzajemna kolejność **a i b** jest nie istotna, ważne aby były przed **e i f**, których kolejność też jest nieistotna.

Miary jakości ww. częściowo uporządkowanej reguły sekwencyjnej:

- Support ({a, b} → {e, f}) = 3 / 4 = 0,75
- Confidence ({a, b} → {e, f}) = (3 / 4) / (3 / 4) = 1
- Support(B) = 3 / 4 = 0,75 stąd: Lift({a, b} → {e, f}) = 1 / 0,75 = 1,33

Analiza asocjacji vs. analiza sekwencji

Cecha	Analiza asocjacji	Analiza sekwencji
Cel	Identyfikacja współwystępujących elementów (np. „klienci kupujący X często kupują też Y – w tym samym czasie”)	Identyfikacja reguł uwzględniających kolejność zdarzeń (np. „po zakupie X klienci często kupują Y, a następnie Z”)
Wymiar czasowy	Brak – dane analizowane są jako zbiory. Kolejność elementów i krotność ich wystąpienia w „koszyku” nie ma znaczenia.	Obecny – istotne jest następstwo zdarzeń w czasie. Krotność elementu w jednym punkcie czasowym nie ma znaczenia[1)]
Wynik	Reguły asocjacyjne	Reguły sekwencyjne
Interpretacja	Odpowiada na pytanie: „Co często pojawia się razem?”	Odpowiada na pytanie: „Co zwykle następuje po czym?”

[1)] W sekwencji ({A, B}, {C}) w pierwszym kroku wystąpiły A i B, a w kolejnym C. Powtarzanie elementu w ramach jednego itemsetu, np. ({A, A, B}, {C}), nie ma znaczenia – ważne jest tylko, że A się pojawiło. W klasycznych algorytmach nadmiarowe wystąpienia w tym samym czasie są traktowane jako jeden element. Jeżeli jednak krotność ma znaczenie (np. analiza liczby leków podanych jednocześnie), trzeba rozszerzyć model i uwzględniać te krotności (*quantitative sequential patterns*) albo kodować je jako odrębne zdarzenia (np. {A1, A2, B} zamiast {A, A, B}).