MED - Sprawozdanie Końcowe

Bartosz Latosek

Czerwiec 2024

Spis treści

1 Wprowadzenie i definicja problemu

1.1 Cel projektu

Celem projektu jest zbadanie własności zaimplementowanego przez siebie algorytmu: czas wykonania, uzyskiwane wyniki dla różnych wartości parametrów algorytmu oraz kilku zbiorów wejściowych. Zależnie od tematu projekt może być realizowany przez jedną lub dwie osoby.

1.2 Temat projektu

Implementacja algorytmu do odkrywania reguł asocjacyjnych (Apriori, Eclat ...) wraz z implementacją wyliczanie miar: współczynnika podniesienia (lift) oraz 3 wybranych: conviction, cosine, gini, Jaacard / coherence, certainty factor, improvement, mutualInformation, odd ratio. Porównanie zachowania się wybranych miar w odniesieniu do wartości współczynnika podniesienia.

Eksperymenty należy przeprowadzić z wykorzystaniem publicznie dostępnych zbiorów danych, np.:

- https://archive.ics.uci.edu/ml/datasets.php
- http://fimi.uantwerpen.be/data/
- https://github.com/deric/clustering-benchmark
- http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php

1.3 Interpretacja / definicja problemu

Postawione zadanie polega na implementacji wybranego algorytmu a następnie weryfikacji rozwiązania i analiza zachowania w specyficznych warunkach. Zaimplementowany został algorytm Apriori. Językiem użytym do wykorzystania zadnia był *Python* w wersji 3.12.3. Rozwiązanie umożliwia wyznaczanie reguł asocjacyjnych z dyskretnych zbiorów danych a następnie obliczanie wartości wybranych miar:

- **Współczynnik podniesienia (lift)**
  Współczynnik podniesienia (lift) to miara, która ocenia siłę reguły asocjacyjnej. Określa, jak bardzo wystąpienie zdarzenia B jest bardziej prawdopodobne, gdy występuje zdarzenie A, w porównaniu do sytuacji, gdyby zdarzenie A i B były niezależne.

  $$\text{Lift}(A \rightarrow B) = \frac{\text{support}(A \cap B)}{\text{support}(A) \times \text{support}(B)}$$

  Interpretacja:
  - Lift $= 1$: Brak zależności między A i B.
  - Lift > 1: Zdarzenie A zwiększa prawdopodobieństwo wystąpienia zdarzenia B.
  - Lift < 1: Zdarzenie A zmniejsza prawdopodobieństwo wystąpienia zdarzenia B.
- **Wsparcie relatywne (relative_support)**
  Wsparcie relatywne (relative support) to miara, która określa, jak często reguła asocjacyjna występuje w zbiorze danych w porównaniu do całkowitej liczby transakcji.

  $$\text{Relative Support}(A \rightarrow B) = \text{support}(A \cap B)$$

  Interpretacja:
  - Wartość wsparcia relatywnego mieści się w przedziale od 0 do 1.
  - Wysoka wartość oznacza, że reguła jest często spotykana w zbiorze danych.

- **Certainty Factor**
  Certainty factor (CF) to miara, która określa stopień pewności, że zdarzenie B nastąpi, gdy wystąpi zdarzenie A. Jest to miara dodatkowego wsparcia udzielanego przez zdarzenie A zdarzeniu B.

$$\mathrm{CF}(A \to B) = \frac{\mathrm{support}(A \cap B) - \mathrm{support}(A) \times \mathrm{support}(B)}{1 - \mathrm{support}(B)}$$

  Interpretacja:

  - CF $= 0$: Zdarzenie A nie wpływa na prawdopodobieństwo zdarzenia B.
  - CF > 0: Zdarzenie A zwiększa prawdopodobieństwo wystąpienia zdarzenia B.
  - CF < 0: Zdarzenie A zmniejsza prawdopodobieństwo wystąpienia zdarzenia B.

- **Jaacard / coherence**
  Miara Jaccarda (coherence) to współczynnik podobieństwa między dwoma zbiorami, w tym przypadku między zbiorem transakcji zawierających A i B.

$$\mathrm{Jaccard}(A \to B) = \frac{\mathrm{support}(A \cap B)}{\mathrm{support}(A) + \mathrm{support}(B) - \mathrm{support}(A \cap B)}$$

  Interpretacja:

  - Wysoka wartość oznacza dużą współzależność między A i B.

- **Odd Ratio**
  Miara odd ratio określa stosunek szans wystąpienia zdarzenia A w obecności zdarzenia B do szans wystąpienia zdarzenia A w braku zdarzenia B.

$$\text{Odd Ratio}(A \to B) = \frac{\mathrm{support}(A \cap B) \times \mathrm{support}(\neg A \cap \neg B)}{\mathrm{support}(A \cap \neg B) \times \mathrm{support}(\neg A \cap B)}$$

  Interpretacja:

  - Odd Ratio $= 1$: Brak zależności między A i B.
  - Odd Ratio > 1: Zdarzenie A jest bardziej prawdopodobne, gdy występuje zdarzenie B.
  - Odd Ratio < 1: Zdarzenie A jest mniej prawdopodobne, gdy występuje zdarzenie B.

Rozwiązanie działa zarówno z danymi podanymi w formie pliku txt oraz z wybranymi zbiorami dyskretnymi z repozytorium UCIML. W ramach przeprowadzonych testów, przeanalizowana została dokładność rozwiązania zestawiona z innym, bardziej rozbudowanym rozwiązaniem dostępnym w formie biblioteki języka *Python* - **apyori**[5]. Następnie analizie poddane zostało zachowanie się wybranych miar w odniesieniu do wartości współczynnika podniesienia.

2 Charakterystyka proponowanego algorytmu/rozwiązania z odniesieniem do literatury

W ramach implementacji proponowanego algorytmu wykorzystane zostały metodyki i najlepsze rozwiązania opisane w dostarczonej do zadania literaturze.

3 Opis implementacji

W ramach implementacji wydzielone zostały abstrakcje obiektów pełniące określone role i zapewniające spójny interfejs zapewniający ortogonalność elementów oprogramowania.

3.1 Dataclasses

Jest to moduł zawierający abstrakcyjne obiekty reprezentujące pojedynczy zbiór atrybutów, regułę asocjacyjną oraz powiązane z nią metryki. Taki podział jest zgodny z paradygmatem programowania obiektowego i zapewnia możliwość definiowania zachowań / metod obiektów odzwierciedlających logikę biznesową.

3.2 Metrics

Jest to klasa, która enkapsułuje w sobie logikę umożliwiającą wyznaczanie wybranych metryk na podstawie dostarczonych zbiorów reguł i zbiorów atrybutów.

3.3 Apriori

To główna klasa, która skpuia w sobie implementacje algorytmu *Apriori* i zapewnia interfejs zwracający reguły asocjacyjne na podstawie dostarczonego do algorytmu zbioru danych.

3.4 Apriori Utils

To klasa pomocnicza, definiująca w sobie metody używane przez algorytm *Apriori*, ale nie bezpośrednio z nim związane - takie jak łączenie zbiorów czy graficzna prezentacja wyliczonych metryk.

3.5 Data Manager

To klasa, która jest odpowiedzialna za dostarczenie danych do pozostałych obiektów. Umożliwia czerpanie danych zarówno z lokalnego pliku tekstowego oraz zbiorów danych udostępnianych przez repozytorium UCIML.

3.6 run.py

To główny skrypt aplikacji, który łączy w sobie działanie wszystkich zdefiniowanych powyżej klas i umożliwia korzystanie z nich za pomocą interfejsu wiersza poleceń.

4 Instrukcja użytkownika (jak uruchomić/korzystać z implementacji)

Poniższy skrypt umożliwia uruchomienie algorytmu Apriori, który służy do wykrywania często współwystępujących wzorców w zbiorze danych. Aby korzystać ze skryptu, należy uruchomić go z odpowiednimi flagami w linii poleceń. Poniżej omówione są dostępne flagi oraz przykłady ich użycia:

- **-f, –input-file**: Ta opcja pozwala na wskazanie pliku wejściowego zawierającego zbiór danych. Plik musi istnieć. Przykład użycia:

  ```
  python script.py -f path/to/dataset.csv
  ```

- **-u, –UCI-dataset**: Ta opcja umożliwia pobranie zadanego zbioru danych z bazy UCI Machine Learning Repository. Domyślnie ustawione jest "car_evaluation". Pozostałe możliwe zbiory danych to "tic_tac_toe_endgame" oraz "nursery". Przykład użycia:

  ```
  python script.py -u nursery
  ```

- **-s, –min-support**: Ta opcja określa minimalną wartość wsparcia (support), która jest używana do filtrowania rzadkich elementów. Domyślnie jest to 0.15. Przykład użycia:

  ```
  python script.py -s 0.2
  ```

- **-c, –min-confidence**: Ta opcja określa minimalną wartość ufności (confidence), która jest używana do wyznaczania silnych reguł asocjacyjnych. Domyślnie jest to 0.6. Przykład użycia:

```
python script.py -c 0.7
```

Przykład uruchomienia skryptu z kilkoma flagami jednocześnie:

```
python script.py -f path/to/dataset.csv -s 0.2 -c 0.7
```

W przypadku, gdy nie zostanie podany plik wejściowy, skrypt spróbuje pobrać domyślny zbiór danych z UCI, chyba że zostanie określony inny zbiór przy użyciu flagi `-u`.

5 Charakterystyka wykorzystywanych zbiorów danych

5.1 Komentarz ogólny

Wykorzystane zostały trzy zbiory danych z repozytorium UCIML o charakterze dystkretnym. W przyszłości możliwa jest implementacja dodatkowej funkcjonalności, która miałaby na celu dyskretyzacje zbiorów danych z wartościami rzeczywistymi na podstawie - na przykład kategoryzacji niska / średnia / wysoka wartość atrybutu.

5.2 Opis zbiorów danych z repozytorium UCIML

1. **Car Evaluation**: - Zbiór danych dotyczący oceny samochodów. Zawiera informacje na temat cech samochodu (np. cena, liczba drzwi, pojemność bagażnika, bezpieczeństwo) oraz decyzji dotyczących oceny samochodu (klasyfikacja jako: unacceptable, acceptable, good, very good). - Zbiór ten zawiera 1728 rekordów, z czego każdy opisuje pojedynczy egzemplarz samochodu. - Dane są wyraźnie zdefiniowane i sklasyfikowane, co ułatwia analizę i zrozumienie danych.

2. **Tic Tac Toe Endgame**: - Zbiór danych dotyczący końcówki gry w kółko i krzyżyk. Zawiera informacje o ułożeniu planszy gry (np. pozycje poszczególnych pól: top-left, top-middle, top-right, itd.) oraz decyzji dotyczących oceny końcowego stanu gry (klasyfikacja jako: win, loss, draw). - Zbiór ten zawiera 958 rekordów, z czego każdy opisuje pojedynczą sytuację w końcówce gry w kółko i krzyżyk. - Dane są w pełni dyskretne i reprezentują wszystkie możliwe kombinacje ułożenia pól na planszy w końcowej fazie gry.

3. **Nursery**: - Zbiór danych dotyczący oceny przedszkoli. Zawiera informacje o cechach przedszkola (np. ilość dzieci, ilość osób dorosłych, poziom opieki) oraz decyzji dotyczących oceny przedszkola (klasyfikacja jako: recommend, priority, not recommend, very_recom). - Zbiór ten zawiera 12960 rekordów, z czego każdy opisuje pojedyncze przedszkole. - Dane są zróżnicowane i uwzględniają różne aspekty funkcjonowania przedszkoli, co pozwala na wszechstronną analizę.

6 Wyniki eksperymentów pokazujących właściwości proponowanego rozwiązania

6.1 Porównanie z innym dostępnym algorytmem

W ramach tej części testów, stworzona implementacja algorytmu *Apriori* była porównywana z dostępną w języku python biblioteką - **apyori** możliwą do pobrania przez:

```
pip install apyori
```

Porównane zostały takie własności jak procent odnalezionych reguł asocjacyjnych, procent poprawnie wyznaczonych wartości wsp. confidence (w odniesieniu do wyników zwracanych przez implementację **apyori**) oraz czasy wykonania obydwu rozwiązań. Całość dostępna jest w skrypcie *apyori_comparison.py*, dostępnym w kodzie źródłowym rozwiązania. Porównanie odbyło się przy uwzględnieniu różnych zbiorów danych i wartości współczynników *min_support* oraz *min_confidence*. Czasy egzekucji były mierzone dla 1 000 000 wywołań funkcji wyznaczającej reguły asocjacyjne dla każdego zestawu parametrów wejściowych.

Dataset	Min Support	Min Confidence	Rules Match	Confidence Match	My Exec Time	Apyori Exec Time
car_evaluation	0.15	0.15	100.00%	100.00%	0.913 ms	0.095 ms
car_evaluation	0.15	0.5	100.00%	100.00%	0.861 ms	0.095 ms
car_evaluation	0.15	0.8	100.00%	100.00%	0.901 ms	0.095 ms
car_evaluation	0.5	0.15	100.00%	100.00%	0.847 ms	0.094 ms
car_evaluation	0.5	0.5	100.00%	100.00%	0.832 ms	0.095 ms
car_evaluation	0.5	0.8	100.00%	100.00%	0.859 ms	0.095 ms
car_evaluation	0.8	0.15	100.00%	100.00%	0.838 ms	0.095 ms
car_evaluation	0.8	0.5	100.00%	100.00%	0.849 ms	0.094 ms
car_evaluation	0.8	0.8	100.00%	100.00%	0.845 ms	0.096 ms
tic_tac_toe_endgame	0.15	0.15	100.00%	100.00%	0.897 ms	0.095 ms
tic_tac_toe_endgame	0.15	0.5	100.00%	100.00%	0.888 ms	0.096 ms
tic_tac_toe_endgame	0.15	0.8	100.00%	100.00%	0.903 ms	0.095 ms
tic_tac_toe_endgame	0.5	0.15	100.00%	100.00%	0.862 ms	0.096 ms
tic_tac_toe_endgame	0.5	0.5	100.00%	100.00%	0.828 ms	0.096 ms
tic_tac_toe_endgame	0.5	0.8	100.00%	100.00%	0.836 ms	0.095 ms
tic_tac_toe_endgame	0.8	0.15	100.00%	100.00%	0.847 ms	0.097 ms
tic_tac_toe_endgame	0.8	0.5	100.00%	100.00%	0.852 ms	0.096 ms
tic_tac_toe_endgame	0.8	0.8	100.00%	100.00%	0.848 ms	0.096 ms
nursery	0.15	0.15	100.00%	100.00%	1.125 ms	0.097 ms
nursery	0.15	0.5	100.00%	100.00%	1.156 ms	0.097 ms
nursery	0.15	0.8	100.00%	100.00%	1.162 ms	0.096 ms
nursery	0.5	0.15	100.00%	100.00%	0.905 ms	0.096 ms
nursery	0.5	0.5	100.00%	100.00%	0.914 ms	0.097 ms
nursery	0.5	0.8	100.00%	100.00%	0.930 ms	0.096 ms
nursery	0.8	0.15	100.00%	100.00%	0.900 ms	0.097 ms
nursery	0.8	0.5	100.00%	100.00%	0.897 ms	0.096 ms
nursery	0.8	0.8	100.00%	100.00%	0.908 ms	0.096 ms

Rysunek 1: Porównanie wyników własnej implementacji i biblioteki apyori

Na powyższym rysunku widać, że w każdym przypadku implementacja wykonana w ramach niniejszego projektu osiągnęła **100%** pokrycia odkrytych reguł asocjacyjnych i wsp. *confidence* dla różnych parametrów wejściowych. Oznacza to, że nie pomija ona żadnej z reguł i jest całkowicie spójna z rozwiązaniem dostarczonym przez bibliotekę *apyori*. Mniej korzystnie wyglądają natomiast czasy egzekucji obydwu implementacji. Własna implementacja algorytmu zajmuje średnio **10** razy dłużej, niż dostarczone rozwiązanie biblioteczne. Jest to jednak akceptowalna różnica - biblioteka *apyori* była tworzona znacznie dokładniej i ze znacznie większym naciskiem na doprecyzowanie niż niniejszy projekt. Po osiągnięciu satysfakcjonujących wyników i dokładności działania algorytmu, twórcy mogli skupić się na aspektach wydajnościowych. W ramach dalszego rozwoju niniejszego projektu, również istniałaby możliwość spróbowania poprawienia wyników czasu egzekucji dostarczonego rozwiązania, ale wykracza to poza zakres projektu.

6.2 Porównanie zachowania się wybranych miar w odniesieniu do wartości współczynnika podniesienia

W ramach niniejszej analizy, własna implementacja algorytmu *Apriori* została użyta do wyznaczenia reguł asocjacyjnych i powiązanych z nimi metryk na trzech dyskretnych zbiorach danych pochodzących z UCIML repo: "car_evaluation", "tic_tac_toe_endgame" oraz "nursery". Uzyskane wyniki zostały posortowane na podstawie wsp. uniesienia - lift a następnie zaprezentowane w postaci tabeli ogólnej i tabeli uśrednionych wartości metryk w zależności od wartości wsp. lift. W związku z nieczytelną i długą listą wyznaczonych reguł asocjacyjnych w dwóch ostatnich zbiorach danych, w ninejszym raporcie zamieszczę jedynie przykładową pełną tabelę dla zbioru danych *Car evaluation* (pozostałe można obejrzeć pobierając kod i uruchamiając skrypt *metrics_comparison.py*).

Wszystkie poniższe wyniki zostały uzyskane przy wartości $min\_support = 0.15$ oraz $min\_confidence = 0.3$.

6.2.1 Car Evaluation

Lift Range	Rule	Relative Support	Lift	Certainty	Jaccard	Odds Ratio
0.6-0.7	('safety_high',) ==> ('target_unacc',)	0.16	0.69	1.44	0.87	2.06
0.7-0.8	('persons_more',) ==> ('target_unacc',)	0.19	0.80	1.68	1.18	2.40
0.7-0.8	('persons_4',) ==> ('target_unacc',)	0.18	0.77	1.63	1.10	2.32
0.8-0.9	('maint_med',) ==> ('target_unacc',)	0.16	0.89	2.48	1.88	3.54
0.8-0.9	('safety_med',) ==> ('target_unacc',)	0.21	0.89	1.86	1.50	2.66
0.8-0.9	('buying_med',) ==> ('target_unacc',)	0.16	0.89	2.48	1.88	3.54
0.8-0.9	('maint_low',) ==> ('target_unacc',)	0.16	0.89	2.48	1.88	3.54
0.9-1.0	('target_unacc',) ==> ('lug_boot_med',)	0.23	0.97	0.46	0.46	1.39
0.9-1.0	('lug_boot_med',) ==> ('target_unacc',)	0.23	0.97	2.04	1.93	2.92
0.9-1.0	('doors_3',) ==> ('target_unacc',)	0.17	0.99	2.78	2.71	3.97
0.9-1.0	('doors_4',) ==> ('target_unacc',)	0.17	0.97	2.70	2.46	3.86
0.9-1.0	('doors_5more',) ==> ('target_unacc',)	0.17	0.97	2.70	2.46	3.86
0.9-1.0	('target_unacc',) ==> ('lug_boot_big',)	0.21	0.91	0.43	0.42	1.30
0.9-1.0	('lug_boot_big',) ==> ('target_unacc',)	0.21	0.91	1.92	1.62	2.74
1.0-1.1	('doors_2',) ==> ('target_unacc',)	0.19	1.08	3.02	3.86	4.31
1.0-1.1	('maint_high',) ==> ('target_unacc',)	0.18	1.04	2.91	3.25	4.15
1.0-1.1	('buying_high',) ==> ('target_unacc',)	0.19	1.07	3.00	3.75	4.28
1.1-1.2	('maint_vhigh',) ==> ('target_unacc',)	0.21	1.19	3.33	7.13	4.76
1.1-1.2	('buying_vhigh',) ==> ('target_unacc',)	0.21	1.19	3.33	7.13	4.76
1.1-1.2	('target_unacc',) ==> ('lug_boot_small',)	0.26	1.12	0.53	0.56	1.59
1.1-1.2	('lug_boot_small',) ==> ('target_unacc',)	0.26	1.12	2.34	3.10	3.35
1.4-1.5	('persons_2',) ==> ('target_unacc',)	0.33	1.43	3.00	29.79	4.28
1.4-1.5	('target_unacc',) ==> ('persons_2',)	0.33	1.43	0.68	0.85	2.04
1.4-1.5	('safety_low',) ==> ('target_unacc',)	0.33	1.43	3.00	29.79	4.28
1.4-1.5	('target_unacc',) ==> ('safety_low',)	0.33	1.43	0.68	0.85	2.04

Rysunek 2: Pełna tabela wyników dla zbioru danych Car Evaluation

Lift Range	No. Rules	Relative Support [Avg. / Std. dev.]	Certainty [Avg. / Std. dev.]	Jaccard [Avg. / Std. dev.]	Odds Ratio [Avg. / Std. dev.]
0.6-0.7	1	[0.160 / 0.000]	[1.443 / 0.000]	[0.870 / 0.000]	[2.060 / 0.000]
0.7-0.8	2	[0.183 / 0.000]	[1.651 / 0.001]	[1.140 / 0.001]	[2.358 / 0.001]
0.8-0.9	4	[0.168 / 0.000]	[2.326 / 0.073]	[1.785 / 0.027]	[3.322 / 0.148]
0.9-1.0	7	[0.199 / 0.001]	[1.863 / 0.898]	[1.723 / 0.778]	[2.862 / 1.120]
1.0-1.1	3	[0.186 / 0.000]	[2.975 / 0.002]	[3.619 / 0.069]	[4.249 / 0.005]
1.1-1.2	4	[0.234 / 0.001]	[2.385 / 1.309]	[4.479 / 7.824]	[3.615 / 1.696]
1.4-1.5	4	[0.333 / 0.000]	[1.840 / 1.346]	[15.323 / 209.370]	[3.162 / 1.260]

Rysunek 3: Tabela wyników uśrednionych dla zbioru danych Car Evaluation

UWAGA! Wyliczone miary wartości średnich oraz odchyleń standardowych mogą wydawać się błędne, w odniesieniu do tabeli z pełnymi wynikami, jednak ma to związek z różnymi precyzjami wyświetlania wyników i zostało zweryfikowane.

Na podstawie powyższej tabeli można stwierdzić, że współczynnik **wsparcia relatywnego (*relative support*)** jest tym większy, im wyższy współczynnik podniesienia lift. Widać też, że wartości w poszczególnych kubełkach nie odbiegają znacząco od siebie (minimalne wartości odchylenia standardowego).

W przypadku współczynnika **Certanity**, nie widać dokładnej zależności pomiędzy wartościami wsp. lift, jednak można dostrzec znacznie wyższą wartość odchylenia standardowego.

Współczynnik Jaccarda rośnie wykładniczo w stosunku do wartości współczynnika podniesienia. To samo dotyczy odchylenia standardowego jego wartości w zadanym przedziale.

W przypadku współczynnika **Odds Ratio** również widać tendencje wzrostową w stosunku do wartości wsp. lift, jednak nie jest to silnie widoczne zarówno w stosunku do wartości średniej jak i odchylenia standardowego.

6.2.2 Tic Tac Toe endgame

Lift Range	No. Rules	Relative Support [Avg. / Std. dev.]	Certainty [Avg. / Std. dev.]	Jaccard [Avg. / Std. dev.]	Odds Ratio [Avg. / Std. dev.]
0.6-0.7	1	[0.154 / 0.000]	[1.227 / 0.000]	[0.760 / 0.000]	[1.877 / 0.000]
0.8-0.9	32	[0.178 / 0.000]	[0.920 / 0.096]	[0.800 / 0.053]	[1.960 / 0.099]
0.9-1.0	36	[0.181 / 0.001]	[0.976 / 0.060]	[0.920 / 0.044]	[2.200 / 0.098]
1.0-1.1	56	[0.192 / 0.003]	[1.234 / 0.317]	[1.334 / 0.445]	[2.658 / 0.392]
1.1-1.2	56	[0.173 / 0.000]	[1.164 / 0.069]	[1.360 / 0.120]	[2.927 / 0.127]
1.2-1.3	46	[0.183 / 0.002]	[1.578 / 0.958]	[-15.597 / 3176.056]	[3.688 / 1.279]
1.3-1.4	16	[0.185 / 0.000]	[1.349 / 0.026]	[2.072 / 0.141]	[3.612 / 0.047]
1.6-1.7	2	[0.200 / 0.000]	[1.630 / 0.002]	[4.413 / 0.080]	[4.647 / 0.003]

Rysunek 4: Tabela wyników uśrednionych dla zbioru danych Tic Tac Toe endgame

Na podstawie powyższej tabeli można stwierdzić, że współczynnik **wsparcia relatywnego (*relative support*)** jest również tym większy, im wyższy współczynnik podniesienia lift. Ponownie widać również, że wartości w poszczególnych kubełkach nie odbiegają znacząco od siebie (minimalne wartości odchylenia standardowego).

W przypadku współczynnika **Certanity**, podobnie jak w przypadku poprzedniego zbioru danych - nie widać dokładnej zależności w stosunku do wartości wsp. lift.

Współczynnik Jaccarda ponownie zdaje się rosnąć wykładniczo w stosunku do wsp. lift, jednak na uwagę zasługuje wartość znacznie odbiegająca od reszty - dla kubełka wartości 1.2-1.3. Wahania powodowane są przez poniższe 4 reguły asocjacyjne pobrane z ogólnej tabeli:

1.2-1.3	('bottom-right-square_x', 'middle-middle-square_x') ==> ('target_positive',)	0.15	1.28	4.66	-198.21	7.14
1.2-1.3	('top-left-square_x', 'middle-middle-square_x') ==> ('target_positive',)	0.15	1.28	4.66	-198.21	7.14
1.2-1.3	('bottom-left-square_x', 'middle-middle-square_x') ==> ('target_positive',)	0.15	1.28	4.66	-198.21	7.14
1.2-1.3	('middle-middle-square_x', 'top-right-square_x') ==> ('target_positive',)	0.15	1.28	4.66	-198.21	7.14

Rysunek 5: Wartości specjalne

W przypadku współczynnika **Odds Ratio** ponownie widać tendencje wzrostową w stosunku do wartości wsp. lift.

6.2.3 Nursery

Lift Range	No. Rules	Relative Support [Avg. / Std. dev.]	Certainty [Avg. / Std. dev.]	Jaccard [Avg. / Std. dev.]	Odds Ratio [Avg. / Std. dev.]
0.9-1.0	58	[0.166 / 0.000]	[1.082 / 0.173]	[1.079 / 0.173]	[2.496 / 0.250]
1.0-1.1	4	[0.171 / 0.000]	[1.175 / 0.243]	[1.290 / 0.340]	[2.733 / 0.368]
1.4-1.5	2	[0.156 / 0.000]	[1.503 / 0.010]	[3.040 / 0.160]	[4.654 / 0.024]
1.6-1.7	2	[0.186 / 0.000]	[1.696 / 0.000]	[5.591 / 0.054]	[5.121 / 0.001]
1.8-1.9	2	[0.190 / 0.000]	[1.833 / 0.015]	[12.399 / 22.473]	[5.675 / 0.035]
2.9-3.0	10	[0.200 / 0.004]	[3.600 / 4.140]	[-1.400 / 1.440]	[12.600 / 19.440]

Rysunek 6: Tabela wyników uśrednionych dla zbioru danych Nursery

Na podstawie powyższej tabeli można ponownie stwierdzić, że współczynnik **wsparcia relatywnego (*relative support*)** jest tym większy, im wyższy współczynnik podniesienia lift. Lekkie wahania można zignorować, przez niewielką ilość reguł asocjacyjnych w środkowych zakresach wartości wsp. lift.

W przypadku współczynnika **Certanity**, podobnie jak w przypadku dwóch poprzednich zbiorów nie widać dokładnej zależności pomiędzy wartościami wsp. lift, jednak można dostrzec znacznie wyższą wartość odchylenia standardowego.

Współczynnik Jaccarda ponownie zdaje się rosnąć wykładniczo w stosunku do wsp. lift. Nie jest to tak dobrze widoczne jak w przypadku poprzednich zbiorów danych przez niezbalansowany rozkład reguł w poszczególnych przedziałach wartości wsp. lift.

W przypadku współczynnika **Odds Ratio** również widać tendencje wzrostową w stosunku do wartości wsp. lift, która zdaje się być lekko zaburzona przez duże odchylenia wartości w ostatnim przedziale.

7 Wnioski

Na podstawie niniejszego raportu można wysnuć kilka istotnych wniosków:

1. Implementacja algorytmu Apriori w ramach projektu osiągnęła pełne pokrycie z wynikami uzyskanymi za pomocą popularnej biblioteki *apyori*, co świadczy o poprawności jej działania.
2. Pomimo pełnego pokrycia, własna implementacja działa średnio 10 razy wolniej niż biblioteka *apyori*. W przyszłości istnieje możliwość optymalizacji kodu w celu poprawy wydajności.
3. Analiza zachowania się wybranych miar w odniesieniu do wartości współczynnika podniesienia wykazała, że większość miar (współczynnik Jaccarda, Wsparcie Relatywne, Certainty Factor, Odds Ratio) wykazuje tendencję do wzrostu wraz ze wzrostem wartości współczynnika podniesienia. Istnieją jednak niejednoznaczności, zwłaszcza w przypadku wartości odstających, które mogą wpływać na interpretację wyników.
4. Dokładność i wiarygodność wyników uzyskanych przez algorytm Apriori zależy od odpowiedniego dobrania parametrów, takich jak minimalne wsparcie i minimalne zaufanie. Warto przeprowadzić analizę wrażliwości, aby określić optymalne wartości tych parametrów dla konkretnego zestawu danych.
5. Analiza wybranych miar asocjacyjnych może dostarczyć cennych informacji o relacjach między zbiorami elementów w danych. Niemniej jednak, interpretacja wyników wymaga ostrożności i głębszej analizy uzyskanych wyników.

Ogólnie rzecz biorąc, projekt dostarcza solidnej implementacji algorytmu Apriori oraz umożliwia przeprowadzenie analizy skuteczności reguł asocjacyjnych i wybranych miar w kontekście różnych zbiorów danych. Istnieje również potencjał do dalszego rozwoju w celu optymalizacji wydajności i rozszerzenia funkcjonalności.

8 Literatura

[1] Roberto J. Bayardo, Rakesh Agrawal, and Dimitrios Gunopulos. Constraint-based rule mining in large, dense databases. In *Proceedings of the 15th International Conference on Data Engineering (ICDE)*, pages 188–197. IEEE, 2000.

[2] Fernando Berzal, Ignacio Blanco, Daniel Sanchez, and Maria-Amparo Vila. Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6(3):221–235, 2002.

[3] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 255–264. ACM, 1997.

[4] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, 2004.

[5] Kiyoto Yabe. apyori: A simple implementation of apriori algorithm. https://github.com/ymoch/apyori, 2015. Accessed: 2024-05-30.