Odkrywanie reguł asocjacyjnych

W tym ćwiczeniu skupiamy się na problemie wzorców zakupowych występujących w koszykach zakupów. Analizujemy własności algorytmów FP-Growth i GSP i sprawdzamy, jak parametry minimalnego wsparcia (*minsup*) i minimalnej ufności (*minconf*) wpływają na uzyskiwane wyniki.

Rapid Miner

- Uruchom narzędzie RapidMiner
- Utwórz nowy przepływ. Użyj operatora `Read CSV` aby załadować plik marketbasket.csv. Jako separator między kolumnami wskaż przecinek. Zaznacz także, że pierwszy wiersz w pliku zawiera nazwy atrybutów. Po załadowaniu pliku uruchom przepływ i obejrzyj metadane.
- Dodaj do przepływu operator `Numerical to Binominal` aby zamienić każdy atrybut na atrybut binarny. Następnie umieść na przepływie operator `FP-Growth` i ustal liczbę poszukiwanych zbiorów częstych na 100 (parametr *find min number of itemsets*). Ustaw też maksymalną liczbę powtórzeń na 100 (paramter *max number of retries*). Uruchom przepływ i obejrzyj wyniki. Porównaj wyniki z sytuacją w której wyłączysz parametr *find min number of itemsets* i ustalisz próg minimalnego wsparcia na *min support=0.01*.
- Dodaj do przepływu operator `Create Association Rules`, wskazując jako kryterium selekcji reguł miarę ufności (ang.*confidence*) z wartością progową 0.8. Uruchom przepływ. Obejrzyj znalezione reguły asocjacyjne, uruchom wizualizację reguł (zakładka *Graph View*), porównaj kilka sposobów wyświetlania reguł. Znajdź reguły tłumaczące, dlaczego ludzie kupują jajka i biały chleb.
- Użyj operatora `Item Sets do Data` w celu przetransformowania znalezionego zbioru zbiorów częstych do postaci danych

- Utwórz nowy przepływ i użyj operatora `Generate Sales Data` do wygenerowania 10 000 syntetycznych transakcji. Obejrzyj uzyskane zbiory dane.
- Przy pomocy operatora `Date to Numerical` zmień sposób przedstawiania czasu na numer dnia w ramach epoki.
- Dokonaj binaryzacji atrybutu *product_category* i obejrzyj wynik
- Wybierz atrybuty *customer_id*, *date* oraz nowo utworzone flagi binarne dla poszczególnych kategorii, i tylko te atrybuty wyślij do kolejnego operatora
- Korzystając z operatora `Generalized Sequential Patterns` znajdź wzorce pokazujące długoterminowe wzorce zakupowe. Przyjmij, że interesują Cię transakcje zawarte w przeciągu roku, wszystkie zakupy dokonane w ramach tygodnia potraktuj jako pojedynczą transakcję, przyjmij też że między poszczególnymi transakcjami danego klienta musiały upłynąć co najmniej dwa tygodnie

Orange Data Mining

- W menu Options wybierz opcję Add-ons i zainstaluj wtyczkę *Associate*. Uruchom ponownie narzędzie
- Dodaj do przepływu operator `Datasets` i pobierz zbiór *Foodmart 2000*.
- Przy użyciu operatora `Select Columns` pozbądź się kolumn reprezentujących identyfikator sklepu
- Użyj operatora `Python Script` aby zmienić typ atrybutów z numerycznych na dyskretne. Posłuż się poniższym kodem

```
from Orange.data import Domain, DiscreteVariable, Table

lst = []

for row in in_data:
    lst.append(['1' if attr > 0 else None for attr in row])

vars = [DiscreteVariable.make(name=attr.name, values=['0','1'])
     for attr in in_data.domain.attributes]

domain = Domain(vars)
out_data = Table(domain, lst)

print(out_data)
```

- Obejrzyj uzyskany wynik w operatorze `Data Table`
- Dodaj operator `Frequent Itemsets`. Porównaj liczbę zbiorów częstych odkrywanych na poziomie minimalnego wsparcia 1%, 0.1%, 0.05%, 0.005%
- Dodaj operator `Association Rules`. Poszukaj reguł asocjacyjnych tłumaczących, jakie zakupy w koszyku "powodują" zakup pizzy.
- Sprawdź, jaki jest maksymalny poziom wsparcia, na którym można znaleźć co najmniej trzy takie reguły posiadające ufność powyżej 80%
- Sprawdź, jaka jest maksymalna ufność, dla której można znaleźć co najmniej trzy takie reguły posiadające wsparcie co najmniej 0.05%