Katedra Systemów Informacyjnych
Algorytmy analizy danych
Kierunek Informatyka, WE, sem. VII
Laboratorium nr 3

Grzegorz Cichosz

1. Zapoznaj się z zawartością pliku obliczenia_wlasne_cw.xls oraz wykonaj polecenia w nim zawarte
2. Poszukiwanie reguł asocjacyjnych w strukturze koszykowej danych
a. Zapoznaj się ze strukturą danych w plikach stragan.txt oraz stragan0.csv
Jaka jest różnica pomiędzy tymi strukturami?
Plik stragan.txt ma typową strukturę koszykową. Każda linia odpowiada jednej transakcji i zawiera listę produktów zakupionych razem.
Plik stragan0.csv ma strukturę tabelaryczną. Wiersze odpowiadają koszykom, a kolumny poszczególnym produktom. Wartości (0 lub 1) informują, czy dany produkt znalazł się w koszyku.

b. Jaka jest liczba wymiarów dla danych zawartych w tych plikach?
W przypadku pliku stragan.txt liczba wymiarów nie jest jednoznaczna, ponieważ transakcje mają różną liczbę produktów. W praktyce można ją utożsamiać z liczbą wszystkich unikalnych produktów.
Dla pliku stragan0.csv liczba wymiarów jest równa liczbie kolumn odpowiadających produktom (oraz ewentualnie kolumnie identyfikatora transakcji).

c. Wczytaj dane stragan0.csv do programu Rattle
Zdejmij (odhacz) parametr Partition (opcja nie jest nam potrzebna)
Zapoznaj się z ustawieniami automatycznie przyjętymi w okienku opisu danych (co jest imputem, targetem, itd.)
Po wczytaniu danych sprawdzono automatycznie przypisane role zmiennych. Wszystkie zostały ustawione jako dane wejściowe (Input).

d. Dla przetwarzanego zbioru danych o strukturze tzw. koszykowej, Transakcja jest typu Ident – zatem zmień na tę opcję. Ident to identyfikator koszyka
Ponieważ analizowany zbiór ma strukturę koszykową, ustawiono typ transakcji na Ident, który pełni rolę identyfikatora koszyka. Dzięki temu dane są poprawnie interpretowane przez algorytm.

e. Przejdź do zakładki Associate
Dokonaj oceny opcji i ustawień do wykonania w tej zakładce
Dla wczytanej struktury danych konieczne jest, abyś ustawił opcję Baskets (zagadnienie koszykowe)
Docelowo będziesz mógł ustawiać progi (minimalne wartości) dla wsparcia i zaufania
Spróbuj własnymi słowami opisać, jak rozumiesz Support oraz Confidence
Support (wsparcie) określa, jak często dana reguła występuje w całym zbiorze danych.
Confidence (zaufanie) mówi, z jakim prawdopodobieństwem spełniona jest część prawa reguły, jeśli spełniona jest jej część lewa.

f. Dla wczytanego zbioru danych wygeneruj reguły asocjacyjne (wybierając Wykonaj, a następnie Show Rules) Przeanalizuj raport. Jakie informacje są w nim zawarte – opisz to w raporcie (sprawozdaniu)
Po uruchomieniu algorytmu wygenerowano reguły asocjacyjne.
Raport zawierał m.in.:
postać reguł (część lewa,część prawa),
wartości wsparcia i zaufania,
dodatkowe miary jakości reguł,
liczbę znalezionych reguł.
Raport umożliwia ocenę jakości otrzymanych zależności.

```
      Length        Class         Mode
          14 transactions           S4

Summary of the Apriori Association Rules:

Number of Rules: 12

Summary of the Measures of Interestingness:

    support         confidence          lift
 Min.   :0.1429   Min.   :0.4000   Min.   :1.120
 1st Qu.:0.1429   1st Qu.:0.4000   1st Qu.:1.120
 Median :0.1429   Median :0.4500   Median :1.400
 Mean   :0.1548   Mean   :0.4806   Mean   :1.509
 3rd Qu.:0.1429   3rd Qu.:0.5250   3rd Qu.:1.680
 Max.   :0.2143   Max.   :0.6667   Max.   :2.333
```

g. Zmień progi na Support na 0.4 oraz Confidence na 0.5
Jak wpłynęła zmiana parametrów na wynik poszukiwań reguł asocjacyjnych
Po zwiększeniu progów wsparcia i zaufania liczba wygenerowanych reguł znacznie się zmniejszyła, a w niektórych przypadkach reguły nie zostały znalezione.
Wyższe progi powodują odrzucenie rzadziej występujących zależności i pozostawienie tylko najbardziej ogólnych reguł.

3. Poszukiwanie reguł o charakterze opisowym zmiennej decyzyjnej - 1
a. Zapoznaj się ze struktura danych z pliku contact-lenses0.csv
Jaka jest charakterystyka danych zawartych w tym zbiorze. Opisz w sprawozdaniu
Czego dotyczą dane? Jak je rozumiesz? Opisz w sprawozdaniu
Opis oryginalny danych znajdziesz pod adresem https://archive.ics.uci.edu/ml/datasets/lenses
Zbiór contact-lenses0.csv dotyczy doboru soczewek kontaktowych dla pacjentów.
Dane opisują cechy pacjentów, takie jak wiek, wada wzroku czy obecność astygmatyzmu, a zmienna decyzyjna określa zalecany typ soczewek.

Wszystkie zmienne mają charakter jakościowy, dzięki czemu dane dobrze nadają się do analizy reguł asocjacyjnych.

b. Wczytaj dane z pliku contact-lenses0.csv
c. Te dane nie są o strukturze koszykowej
W opcjach zakładki Data wszystkie zmienne muszą mieć opcję Input
Zatem nie ustawiamy też opcji Baskets w zakładce Associate
Dane wczytano do programu Rattle.

Zbiór nie ma struktury koszykowej, dlatego wszystkie zmienne ustawiono jako Input i nie zaznaczano opcji Baskets.

d. Dla wczytanego zbioru danych wygeneruj reguły asocjacyjne

Dla zbioru wygenerowano reguły asocjacyjne opisujące zależności pomiędzy cechami pacjentów a rodzajem zalecanych soczewek.

```
    support         confidence          lift
 Min.   :0.1250   Min.   :0.2000   Min.   :0.400
 1st Qu.:0.1250   1st Qu.:0.5000   1st Qu.:1.000
 Median :0.1667   Median :0.5000   Median :1.000
 Mean   :0.1751   Mean   :0.5851   Mean   :1.297
 3rd Qu.:0.2083   3rd Qu.:0.7500   3rd Qu.:1.500
 Max.   :0.5000   Max.   :1.0000   Max.   :6.000
```

e. Dokonaj wstępnej oceny reguł

Znajdź reguły użyteczne, które pozwolą ustalić zasady przypisywanie w diagnostyce medycznej danego rodzaju

szkieł kontaktowych pacjentowi

W sprawozdaniu umieść te reguły oraz uzasadnij dlaczego można je nazwać regułami użytecznymi

Za reguły użyteczne uznano te, które:

pozwalają określić typ soczewek na podstawie cech pacjenta,

mają stosunkowo wysokie zaufanie,

są logiczne i łatwe do interpretacji.

Reguły te mogą być pomocne w procesie diagnostycznym.

4. Poszukiwanie reguł o charakterze opisowym zmiennej decyzyjnej - 2

a. Zapoznaj się ze struktura danych z pliku iris0.csv

Jaka jest charakterystyka danych zawartych w tym zbiorze. Opisz w sprawozdaniu

Czego dotyczą dane? Jak je rozumiesz? Jak jest ich charakterystyka? Opisz w sprawozdaniu

Opis oryginalny danych znajdziesz pod adresem https://archive.ics.uci.edu/ml/datasets/iris

Czy na podstawie tych danych jest możliwe szukanie reguł asocjacyjnych? Czy może trzeba będzie dokonać

pewnego ich przetworzenia, a jeśli tak to jakiego?

Jak powinny wyglądać reguły aby były one użyteczne?

Zbiór iris0.csv zawiera dane dotyczące trzech gatunków irysa.

Cechy opisują wymiary działek kielicha i płatków kwiatów, natomiast zmienna decyzyjna określa gatunek rośliny.

Dane zawierają zmienne numeryczne, dlatego przed wyszukiwaniem reguł konieczne jest ich odpowiednie przetworzenie.

b. Wczytaj dane z pliku iris0.csv

c. Te dane nie są o strukturze koszykowej

W opcjach zakładki Data wszystkie zmienne muszą mieć opcję Input

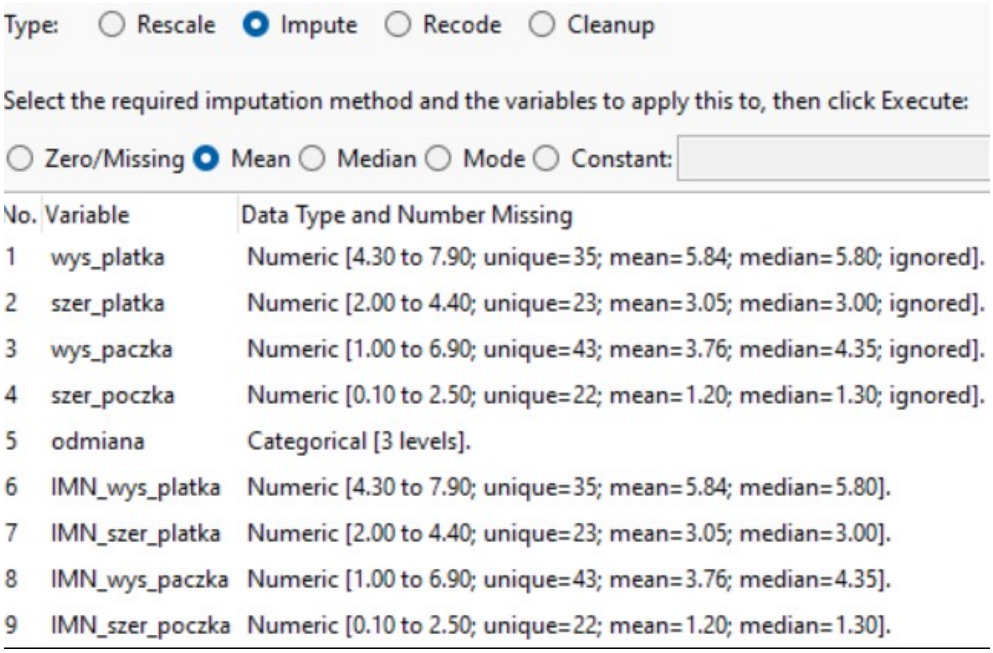
Zatem nie ustawiamy też opcji Baskets w zakładce Associate

Dane wczytano do programu Rattle.

Zbiór nie ma struktury koszykowej, dlatego wszystkie zmienne ustawiono jako Input, bez użycia opcji Baskets.

d. UWAGA! Zanim zaczniemy poszukiwać reguł, musimy dokonać ich dyskretyzacji. Zrobimy to w zakładce
Transform, gdzie dokonamy (najprościej) z opcją Recode oraz Equal With Number 6 (możesz ustawić na 8, ale
jaki będzie to miało skutek na transformację wczytanych danych, odpowiedź zawrzyj w sprawozdaniu)
Zaznacz wszystkie dane numeryczne i wykonaj przetworzenie danych (wybierz Wykonaj)
Jaki jest efekt tego przetworzenia? Co się wydarzyło? Opisz w sprawozdaniu
Przed analizą reguł wykonano dyskretyzację danych w zakładce Transform, korzystając z opcji Recode oraz metody Equal Width z liczbą przedziałów równą 6.

W wyniku tego procesu wartości numeryczne zostały zamienione na przedziały.
Zwiększenie liczby przedziałów (np. do 8) powoduje większą szczegółowość danych, ale jednocześnie zmniejsza wsparcie reguł.



No.	Variable	Data Type and Number Missing
1	wys_platka	Numeric [4.30 to 7.90; unique=35; mean=5.84; median=5.80; ignored].
2	szer_platka	Numeric [2.00 to 4.40; unique=23; mean=3.05; median=3.00; ignored].
3	wys_paczka	Numeric [1.00 to 6.90; unique=43; mean=3.76; median=4.35; ignored].
4	szer_poczka	Numeric [0.10 to 2.50; unique=22; mean=1.20; median=1.30; ignored].
5	odmiana	Categorical [3 levels].
6	IMN_wys_platka	Numeric [4.30 to 7.90; unique=35; mean=5.84; median=5.80].
7	IMN_szer_platka	Numeric [2.00 to 4.40; unique=23; mean=3.05; median=3.00].
8	IMN_wys_paczka	Numeric [1.00 to 6.90; unique=43; mean=3.76; median=4.35].
9	IMN_szer_poczka	Numeric [0.10 to 2.50; unique=22; mean=1.20; median=1.30].

e. Teraz w zakładce Associate poszukamy reguł asocjacyjnych
Dokonaj odpowiedniego ustawienie parametrów progowych algorytmu. Dla jakich ustawień znajdziesz reguły
użyteczne? Dlaczego uznałeś je za użyteczne?
W sprawozdaniu umieść te reguły z komentarzem

Po dyskretyzacji danych wygenerowano reguły asocjacyjne.
Za reguły użyteczne uznano te, które:
prowadzą do określenia gatunku irysa,
mają wysokie zaufanie,
są krótkie i czytelne.
Takie reguły pozwalają w prosty sposób opisać cechy charakterystyczne poszczególnych gatunków.

```
    lhs                                   rhs                                 support   confidence
[1] {tear.prod.rate=reduced}           => {contact.lenses=none}               0.5000000  1.0000000
[2] {contact.lenses=none}              => {tear.prod.rate=reduced}            0.5000000  0.8000000
[3] {astigmatism=yes}                  => {contact.lenses=none}               0.3333333  0.6666667
[4] {contact.lenses=none}              => {astigmatism=yes}                   0.3333333  0.5333333
[5] {spectacle.prescrip=hypermetrope} => {contact.lenses=none}                0.3333333  0.6666667
[6] {contact.lenses=none}              => {spectacle.prescrip=hypermetrope}   0.3333333  0.5333333
```