# *Analisis Asosiasi 2*

CS 4333 Data Mining

Imelda Atastina

# Atribut Kategoris

- Masalah : Bagaimana menerapkan analisisi asosiasi pada atribut yang bertipe kategoris?
- Solusi : Transformasi data menjadi “item-item” (**Binary Item**), sehingga bentuk data seperti data transaksi, baru diterapkan algoritma-algoritma pada metoda analisis asosiasi

# Contoh Data

| Gender | Education | State | Chat Online | Shop Online | Privacy Concern |
|---|---|---|---|---|---|
| Female | Graduate | Illinois | Yes | Yes | Yes |
| Male | College | California | No | No | No |
| Male | Graduate | Michigan | Yes | Yes | Yes |
| Female | College | Virginia | No | Yes | Yes |
| Female | Graduate | California | No | No | Yes |
| Male | College | Minnesota | Yes | Yes | Yes |
| Male | College | Alaska | Yes | Yes | No |
| Male | High School | Oregon | No | No | No |
| Female | Graduate | Texas | Yes | No | No |
| … | … | … | … | … | … |

# Data Hasil Transformasi

| Male | Female | Education =Graduate | Education = College | … | Shop Online = Yes | Shop Online =No |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | … | 1 | 0 |
| 1 | 0 | 0 | 1 | … | 0 | 1 |
| 1 | 0 | 1 | 0 | | 1 | 0 |
| 0 | 1 | 0 | 1 | … | 1 | 0 |
| 0 | 1 | 1 | 0 | … | 0 | 1 |
| 1 | 0 | 0 | 1 | … | 1 | 0 |
| | | | | | | |
| | | | | | | |
| | | | | | | |

# Issue yang harus dipertimbangkan (1)

- Beberapa nilai atribut kurang sering muncul untuk menjadi “frequent pattern”.

  Solusi yang diusulkan munculkan kategori baru sehingga yang kurang sering muncul nilainya menjadi lebih besar.

  Contoh : munculkan kategori “lain-lain”

# Issue yang harus dipertimbangkan (2)

- Beberapa nilai atribut terlalu tinggi frekuensinya dibandingkan nilai atribut lainnya, sehingga dapat memunculkan "redundant pattern".

  Solusi yang diusulkan lakukan preprocessing dengan menghilangkan atribut yang dianggap kurang penting atau redundant

  Contoh : {Computer at home= Yes, Shop Online = Yes}-> {Privacy Concern =Yes}

# Issue yang harus dipertimbangkan (3)

- Waktu proses menjadi lama, terutama jika ada item yang terbentuk dengan frekuensi tinggi, shg candidate itemset yang terbentukpun banyak.
- Solusi : Hindari membentuk candidate itemset yang mengandung lebih dari satu item dari atribut yang sama, krn support countnya pasti nol

  Contoh : {State = X, State = Y, ….}

# Atribut Kontinu

- Dikenal juga dengan nama ***quantitative association rules***
- Preprocessing yang dapat dilakukan :

  - Metoda Diskritisasi
  - Metoda Statistik
  - Metoda Non diskritisasi

# Metoda Diskritisasi

- Diskritisasi pada data kontinu, dilakukan dengan mengelompokkan nilai dari data kontinu tersebut pada sejumlah interval.
- Setelah itu diikuti dengan transformasi binarisasi data diskrit sebelum dapat dilakukan proses analisis asosiasi

# Diskritisasi (2)

- Parameter kunci dalam proses diskritisasi adalah jumlah interval (range interval) yang digunakan untuk membagi data pada tiap atribut.
- Salah menentukan interval dapat berakibat tidak muncul rule yang memenuhi treshold atau justru terlalu banyak rule yang muncul

# Contoh

| Gender | Age | Income | … | Privacy Concern |
|---|---|---|---|---|
| Female | 26 | 90 K | … | Yes |
| Male | 51 | 135 K | … | No |
| Male | 29 | 80 K | … | Yes |
| Female | 45 | 120 K | … | Yes |
| Female | 31 | 95 K | … | Yes |
| Male | 25 | 55 K | … | Yes |
| … | … | … | … | No |

- Data Age dibagi dalam interval sbb:

  Age $\in$[16,20)

  Age $\in$[20,24)

  Age $\in$[24,32)

  …

  Age $\in$[50,60)

# Isu penting dalam proses diskritisasi (1)

- Proses komputasi menjadi mahal, perhatikan jika data terbagi menjadi k interval, maka akan ada k(k-1)/2 binary item yang harus dibentuk, dan jika item-item tersebut frequent, maka candidate itemset yang terbentukpun banyak.
- Solusi yang diusulkan : gunakan nilai treshold support yang maksimum

# Isu penting dalam proses diskritisasi (2)

- Banyak *redundant rule* yang terbentuk
- Contoh :

  R1 : {Age $\in$[16,20), Gender = Male}-> {Chat Online = Yes}

  R2 : {Age $\in$[16,24), Gender = Male}-> {Chat Online = Yes}

  Seharusnya yang digunakan R2 saja , karena *coverage* nya lebih besar, jadi R1 merupakan *redundant rule*

# Metoda Statistik

- Dalam pembentukan rule statistik yang digunakan adalah statistik deskriptif seperti rata-rata,median, variansi, dsb. Digunakan untuk menggambarkan populasi atribut yang dianggap menarik
- Untuk memvalidasi rule yang terpilih digunakan uji hipotesa

# Metoda Non Diskritisasi

- Contoh penggunaan metoda non diskritisasi adalah text mining yang dikenal dengan algoritma Min-Apriori

# Min-Apriori (Han et al)

**Document-term matrix:**

| TID | W1 | W2 | W3 | W4 | W5 |
|---|---|---|---|---|---|
| D1 | 2 | 2 | 0 | 0 | 1 |
| D2 | 0 | 0 | 1 | 2 | 2 |
| D3 | 2 | 3 | 0 | 0 | 0 |
| D4 | 0 | 0 | 1 | 0 | 1 |
| D5 | 1 | 1 | 1 | 0 | 2 |

**Contoh :**

**W1 dan W2 cenderung muncul bersamaan**

# Min-Apriori

| TID | W1 | W2 | W3 | W4 | W5 |
|---|---|---|---|---|---|
| D1 | 2 | 2 | 0 | 0 | 1 |
| D2 | 0 | 0 | 1 | 2 | 2 |
| D3 | 2 | 3 | 0 | 0 | 0 |
| D4 | 0 | 0 | 1 | 0 | 1 |
| D5 | 1 | 1 | 1 | 0 | 2 |

- Solusi yang potensial :

  Ubah menjadi matriks 0/1 dan gunakan algoritma yang ada ( mis. Apriori)
- Diskritisasi tidak diperlukan karena user justru menginginkan asosiasi antar kata bukan pada range kata-kata

# Min-Apriori

- Bagaimana menentukan nilai support sbh kata?
  - Normalisasi vektor kata – mis. menggunakan $L_1$ norm
    - Setiap kata harus mempunyai nilai support 1.0

| TID | W1 | W2 | W3 | W4 | W5 |
|---|---|---|---|---|---|
| D1 | 2 | 2 | 0 | 0 | 1 |
| D2 | 0 | 0 | 1 | 2 | 2 |
| D3 | 2 | 3 | 0 | 0 | 0 |
| D4 | 0 | 0 | 1 | 0 | 1 |
| D5 | 1 | 1 | 1 | 0 | 2 |

**Normalize** →

| TID | W1 | W2 | W3 | W4 | W5 |
|---|---|---|---|---|---|
| D1 | 0.40 | 0.33 | 0.00 | 0.00 | 0.17 |
| D2 | 0.00 | 0.00 | 0.33 | 1.00 | 0.33 |
| D3 | 0.40 | 0.50 | 0.00 | 0.00 | 0.00 |
| D4 | 0.00 | 0.00 | 0.33 | 0.00 | 0.17 |
| D5 | 0.20 | 0.17 | 0.33 | 0.00 | 0.33 |

# Min-Apriori

- Definisi support :

$$\sup(C) = \sum_{i \in T} \min_{j \in C} D(i, j)$$

| TID | W1 | W2 | W3 | W4 | W5 |
|---|---|---|---|---|---|
| D1 | 0.40 | 0.33 | 0.00 | 0.00 | 0.17 |
| D2 | 0.00 | 0.00 | 0.33 | 1.00 | 0.33 |
| D3 | 0.40 | 0.50 | 0.00 | 0.00 | 0.00 |
| D4 | 0.00 | 0.00 | 0.33 | 0.00 | 0.17 |
| D5 | 0.20 | 0.17 | 0.33 | 0.00 | 0.33 |

**Example:**

**Sup(W1,W2,W3)**

**= 0 + 0 + 0 + 0 + 0.17**

**= 0.17**

# Sifat Anti-monotone Support

| TID | W1 | W2 | W3 | W4 | W5 |
|---|---|---|---|---|---|
| D1 | 0.40 | 0.33 | 0.00 | 0.00 | 0.17 |
| D2 | 0.00 | 0.00 | 0.33 | 1.00 | 0.33 |
| D3 | 0.40 | 0.50 | 0.00 | 0.00 | 0.00 |
| D4 | 0.00 | 0.00 | 0.33 | 0.00 | 0.17 |
| D5 | 0.20 | 0.17 | 0.33 | 0.00 | 0.33 |

**Contoh:**

**Sup(W1) = 0.4 + 0 + 0.4 + 0 + 0.2 = 1**

**Sup(W1, W2) = 0.33 + 0 + 0.4 + 0 + 0.17 = 0.9**

**Sup(W1, W2, W3) = 0 + 0 + 0 + 0 + 0.17 = 0.17**

# Latihan

| No | Weather Condition | Driver'sCondition | Traffic Violation | Seat Bealt | Crash Severity |
|---|---|---|---|---|---|
| 1 | Good | Alcohol-impaired | Exceed speed limit | No | Major |
| 2 | Bad | Sober | None | Yes | Minor |
| 3 | Good | Sober | Disobey stop sign | Yes | Minor |
| 4 | Good | Sober | Exceed speed limit | Yes | Major |
| 5 | Bad | Sober | Disobey Traffing signal | No | Major |
| 6 | Good | Alcohol-impaired | Disobey stop sign | Yes | Minor |
| 7 | Bad | Sober | None | Yes | Major |
| 8 | Good | Alcohol-impaired | Disobey Traffing signal | Yes | Major |
| 9 | Good | Alcohol-impaired | None | No | Major |
| 10 | Bad | Sober | Disobey Traffing signal | No | Major |
| 11 | Good | Alcohol-impaired | Exceed speed limit | Yes | Major |
| 12 | Bad | Sober | Disobey stop sign | Yes | Minor |

a. Show binarized version of the data

b. Assuming support treshold 30%, how many candidate and frequent itemset will be generated?

c. Assuming min confidence 70%, generate the rules!

# Latihan

| A | B | C |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 1 |
| 3 | 1 | 0 |
| 4 | 1 | 0 |
| 5 | 1 | 1 |
| 6 | 0 | 1 |
| 7 | 0 | 0 |
| 8 | 1 | 1 |
| 9 | 0 | 0 |
| 10 | 0 | 0 |
| 11 | 0 | 0 |
| 12 | 0 | 1 |

**The first attribute is continuous while the remaining two attributes are asymmetric binary. A Rule is considered strong if its support exceeds 15% and its confidence exceeds 60%. The given supports the following two rules**

**a. {(1≤ A ≤ 2), B =1} ->{C=1}**

**b. {(5≤ A ≤ 8), B =1} ->{C=1}**

**(i) Compute the support and confidence for both rules**

**(ii) Use discertize with bind-width = 2,3, 4, state wether the above two rules discovered by the Apriori algorithm**