Association Rule in Data Mining

Oleh: Dr. Indra, S.Kom, M.T.I

Outline

Definisi Association rule Mining

- Dengan sekumpulan transaksi, temukan aturan yang akan **memprediksi** kemunculan item berdasarkan kemunculan item lainnya pada suatu transaksi
- Prinsipnya: kemunculan bersama, bukan kausalitas !

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

Definisi Frequent Itemset

- Itemset
  - Sekumpulan satu atau lebih item
  - Contoh : {Milk, Bread, Diaper}
  - K-itemset: itemset yang berisi sejumlah k items
- Support count
  - Frekuensi kemunculan dari suatu itemset
  - Contoh: σ ({Milk, Bread, Diaper}) =2
- Support
  - Perbandingan transaksi yang berisi sekumpulan item (itemset) dibandingkan total transaksi
  - Contoh: s({Milk, Bread, Diaper}) =2/5
- Frequent itemset
  - Suatu itemset yang nilai supportnya diatas atau sama dengan batas minimum support

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Definisi Association Rule

- Association Rule
  - Implikasi dari suatu ekspresi/kondisi X --> Y, dimana X dan Y adalah itemsets
  - Contoh: {Milk,Diaper} --> {Beer}
- Rule Evaluation Metrics
  - Support (s)
    - Perbandingan dari suatu transaksi yang mencakup itemset X dan Y
  - Confidence (c)
    - Mengukur seberapa sering item dalam Y muncul dalam transaksi yang mengandung X

Penjabaran Association Rule

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$\{\text{Milk, Diaper}\} \Rightarrow \{\text{Beer}\}$$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

Association Rule Mining Task

- Target akhir dari Rule yang dibangun dari sekumpulan transaksi T harus memenuhi syarat sbb:
  - Nilai support >= nilai ambang batas minsup
  - Confidence >= nilai ambang batas minconf

Memahami Rule yang dihasilkan

Mining Association Rules

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset: {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Makna dari Rule (Slide 8)

- Aturan yang berasal dari itemset yang sama memiliki nilai support yang sama tetapi dapat memiliki kepercayaan (confidence) yang berbeda
- Dengan demikian, dapat dipisahkan persyaratan nilai support dan nilai confidence

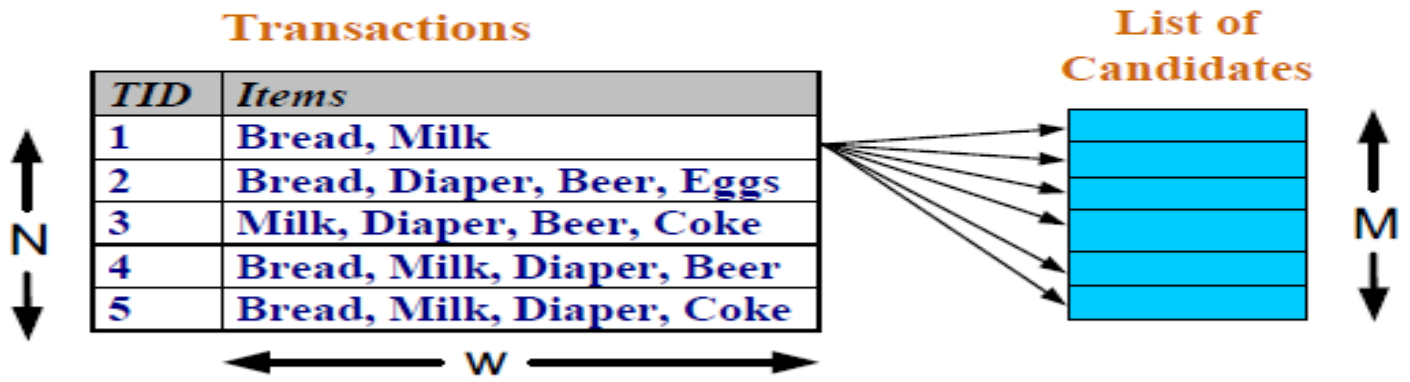
Generate Association Rule

- Dua Tahap membangun Rule:
  - Frequent itemset Generation (membangun frequent itemset)
    - Generate seluruh itemset dimana nilai support >= minsup
  - Rule Generation (Membuat beberapa Rule)
    - Generate rules dengan nilai confidence yang tinggi dari setiap frequent itemset dimana setiap rule adalah percabangan secara biner (cabang 2) pada setiap frequent itemset
- Membangun frequent itemset membutuhkan kompleksitas algoritma yang tinggi (big O)

Membangun Frequent itemset dengan Brute Force algorithm

Frequent Itemset Generation

- Brute-force approach:
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database



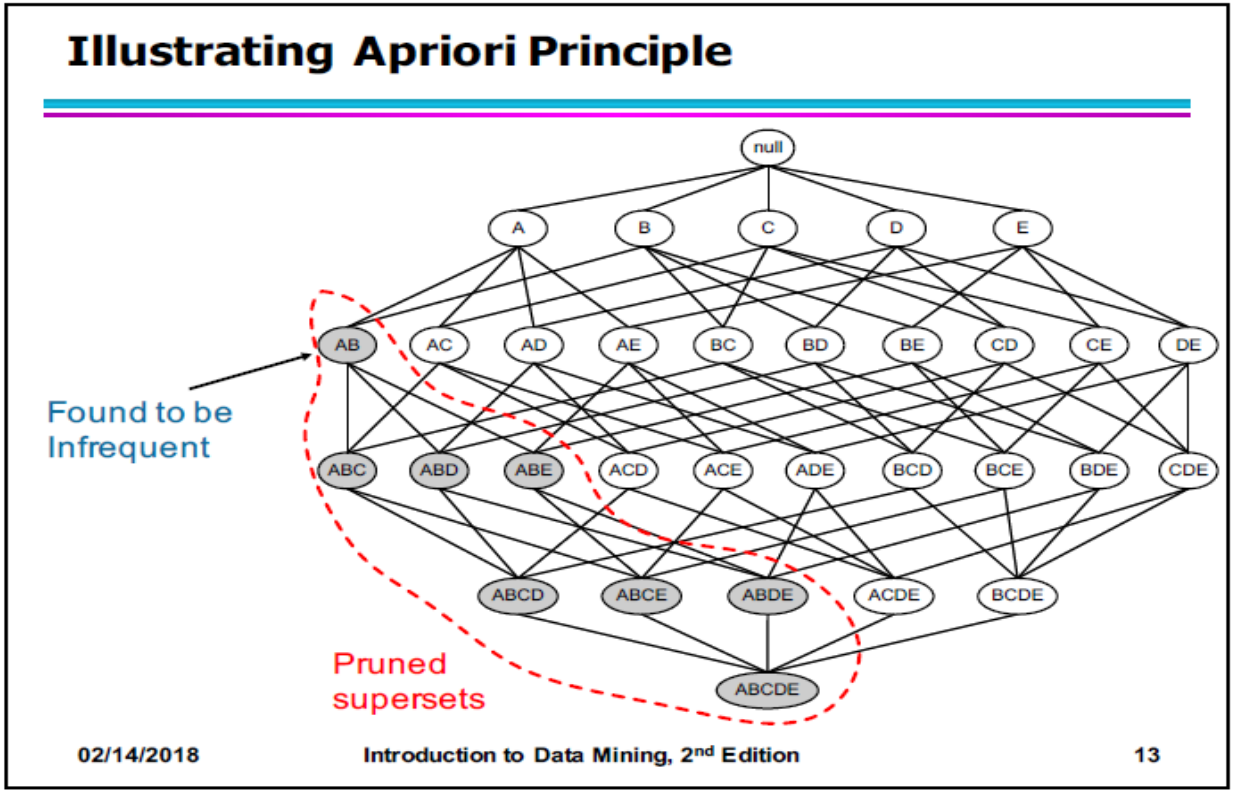
  - Match each transaction against every candidate
  - Complexity ~ O(NMw) => Expensive since $M = 2^d$!!!

Mengurangi jumlah kandidat itemset

- Prinsip Apriori
  - Jika suatu itemset memenuhi nilai minimum support maka semua subset yang berisi itemset tersebut juga memenuhi nilai minimum support
  - Nilai Support dari suatu itemset tidak pernah melebihi dari nilai support dari suatu subsets

Ilustrasi Pengurangan Kandidat Itemset

- Infrequent berarti itemset tersebut kemunculannya dibawah minimum support
- K-itemset yang berisi itemset infrequent dihilangkan karena tidak memenuhi nilai minimum support



Algoritme Apriori

Apriori Algorithm

- F_k: frequent k-itemsets
- L_k: candidate k-itemsets

- Algorithm
  - Let k=1
  - Generate F_1 = {frequent 1-itemsets}
  - Repeat until F_k is empty
    - **Candidate Generation**: Generate L_{k+1} from F_k
    - **Candidate Pruning**: Prune candidate itemsets in L_{k+1} containing subsets of length k that are infrequent
    - **Support Counting**: Count the support of each candidate in L_{k+1} by scanning the DB
    - **Candidate Elimination**: Eliminate candidates in L_{k+1} that are infrequent, leaving only those that are frequent => F_{k+1}

Dataset Ilustrasi Apriori; nilai minimum support= 2 atau 50%

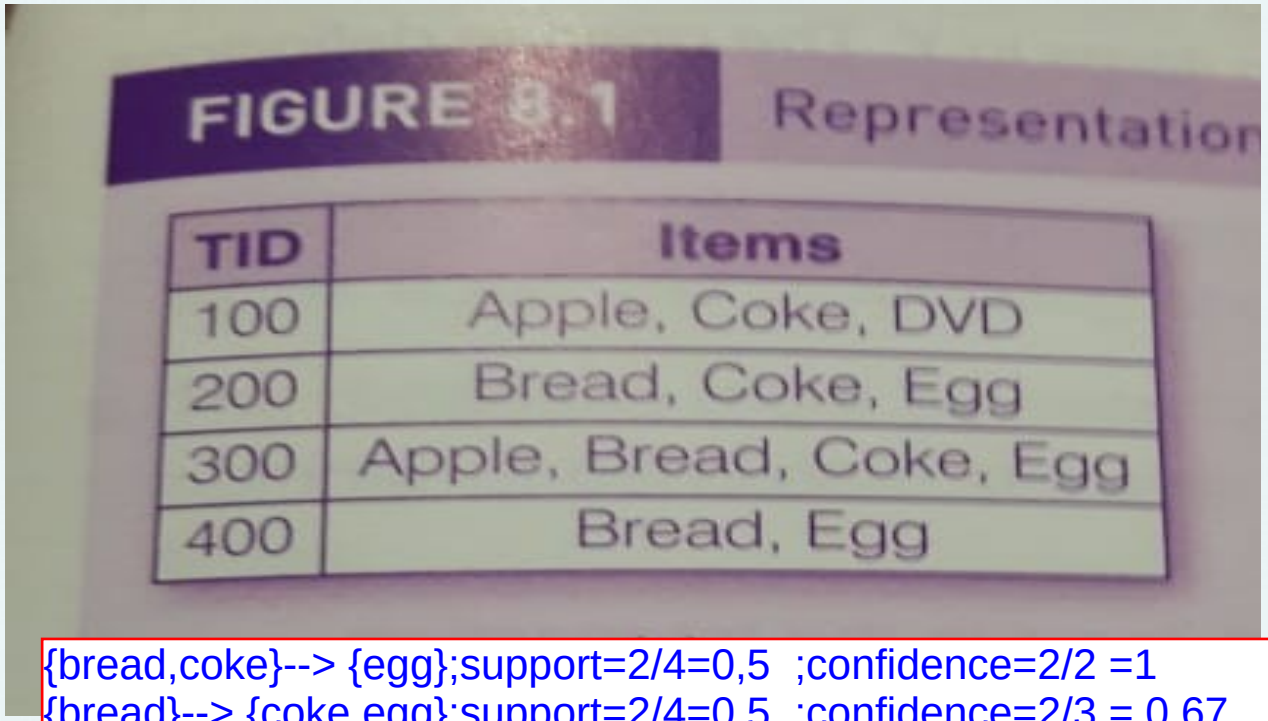


FIGURE 8.1 Representation

TID	Items
100	Apple, Coke, DVD
200	Bread, Coke, Egg
300	Apple, Bread, Coke, Egg
400	Bread, Egg

{bread,coke}--> {egg};support=2/4=0,5 ;confidence=2/2 =1
{bread}--> {coke,egg};support=2/4=0,5 ;confidence=2/3 = 0,67
{bread,egg}--> {coke};support=2/4=0,5 ;confidence=2/3 = 0,67
{coke}-->{bread,egg};support=2/4=0,5 ;confidence=2/3 = 0,67
{coke,egg}--> {bread};support=2/4=0,5 ;confidence=2/2=1
{egg}-->{coke,bread};support=2/4=0,5 ;confidence=2/3=0,67
misal minimum support=50% ; minimum confidence=70% maka rule yg memenuhi minimum support dan confidence:
{bread,coke}--> {egg};
{coke,egg}--> {bread}

Ilustrasi Apriori

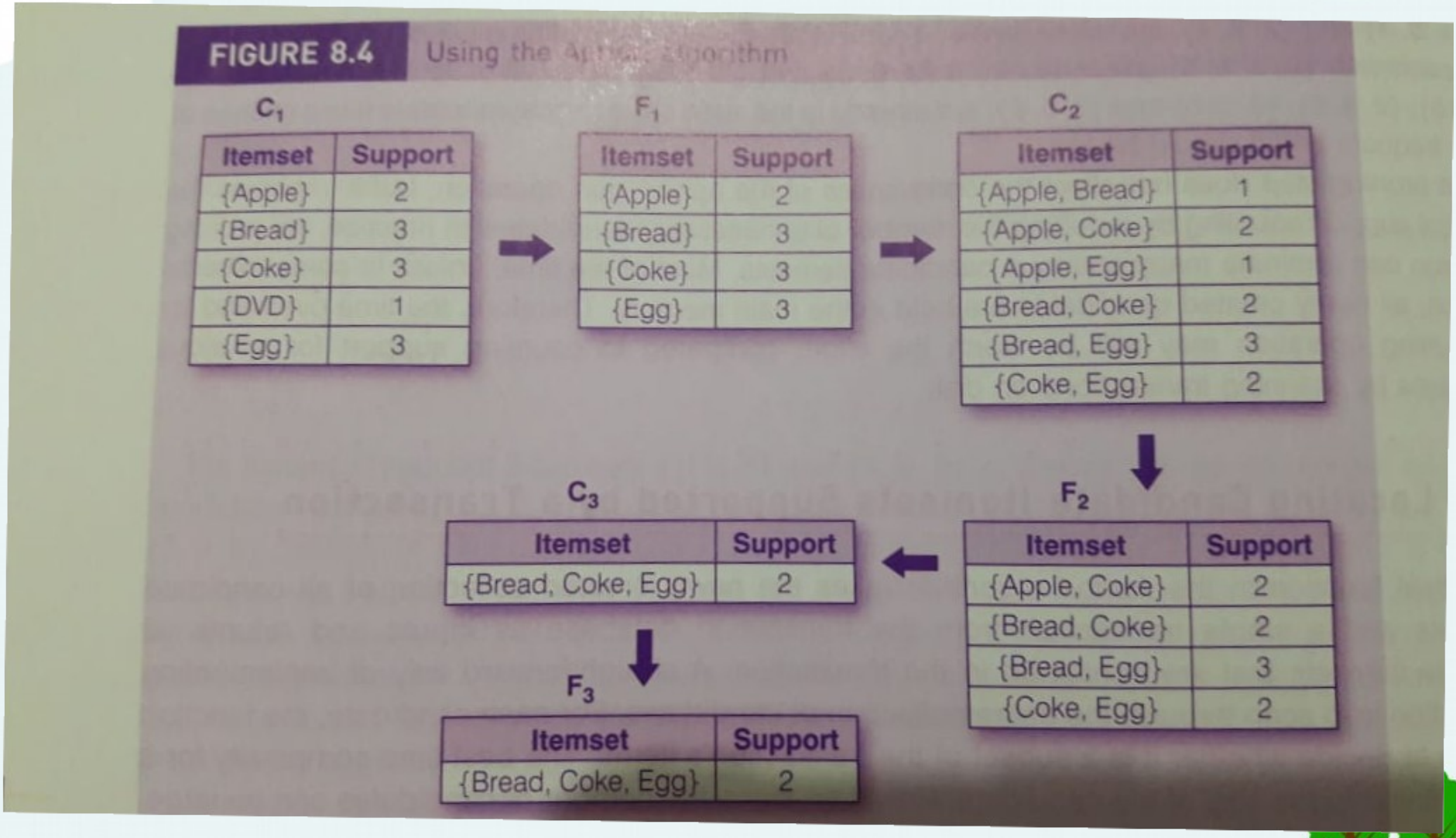


FIGURE 8.4 Using the Apriori Algorithm

C_1

Itemset	Support
{Apple}	2
{Bread}	3
{Coke}	3
{DVD}	1
{Egg}	3

F_1

Itemset	Support
{Apple}	2
{Bread}	3
{Coke}	3
{Egg}	3

C_2

Itemset	Support
{Apple, Bread}	1
{Apple, Coke}	2
{Apple, Egg}	1
{Bread, Coke}	2
{Bread, Egg}	3
{Coke, Egg}	2

F_2

Itemset	Support
{Apple, Coke}	2
{Bread, Coke}	2
{Bread, Egg}	3
{Coke, Egg}	2

C_3

Itemset	Support
{Bread, Coke, Egg}	2

F_3

Itemset	Support
{Bread, Coke, Egg}	2

Penjelasan Ilustrasi Apriori

- C1 adalah candidate kemunculan 1-item pada beberapa transaksi
- F1 adalah Frequent 1-itemset dimana jumlah kemunculan transaksinya diatas nilai minimum support
- C2 adalah candidate kemunculan 2-item (penggabungan dari 1-item) pada beberapa transaksi
- F2 adalah Frequent 2-itemset (berisi gabungan 2 item) dimana jumlah kemunculan transaksinya diatas nilai minimum support

Membangun Rule dari Apriori

- Hasil kombinasi itemset dari F3 digunakan untuk membuat rule
- Rule dibangun dengan syarat nilai support dan confidence diatas atau sama dengan minsupp atau minconfidence berikut:
  - {bread,coke} --> {egg}
  - {bread}--> {coke,egg}
  - Total rule = $2^3 - 2 = 6$ (6 rule terbentuk)

{bread,coke}--> {egg};
{bread}--> {coke,egg};
{bread,egg}--> {coke}
{coke}-->{bread,egg}
{coke,egg}--> {bread};
{egg}-->{coke,bread};

Membangun Association Rule dari Data Kuantitatif

Dataset dengan atribut kuantitatif (atribut: recordID, Age, Gender, NoofMobilesUsed)

RecordID	Age	Gender	NoOfMobilesUsed
1	25	Male	1
2	23	Female	1
3	29	Male	0
4	33	Female	2
5	37	Female	2

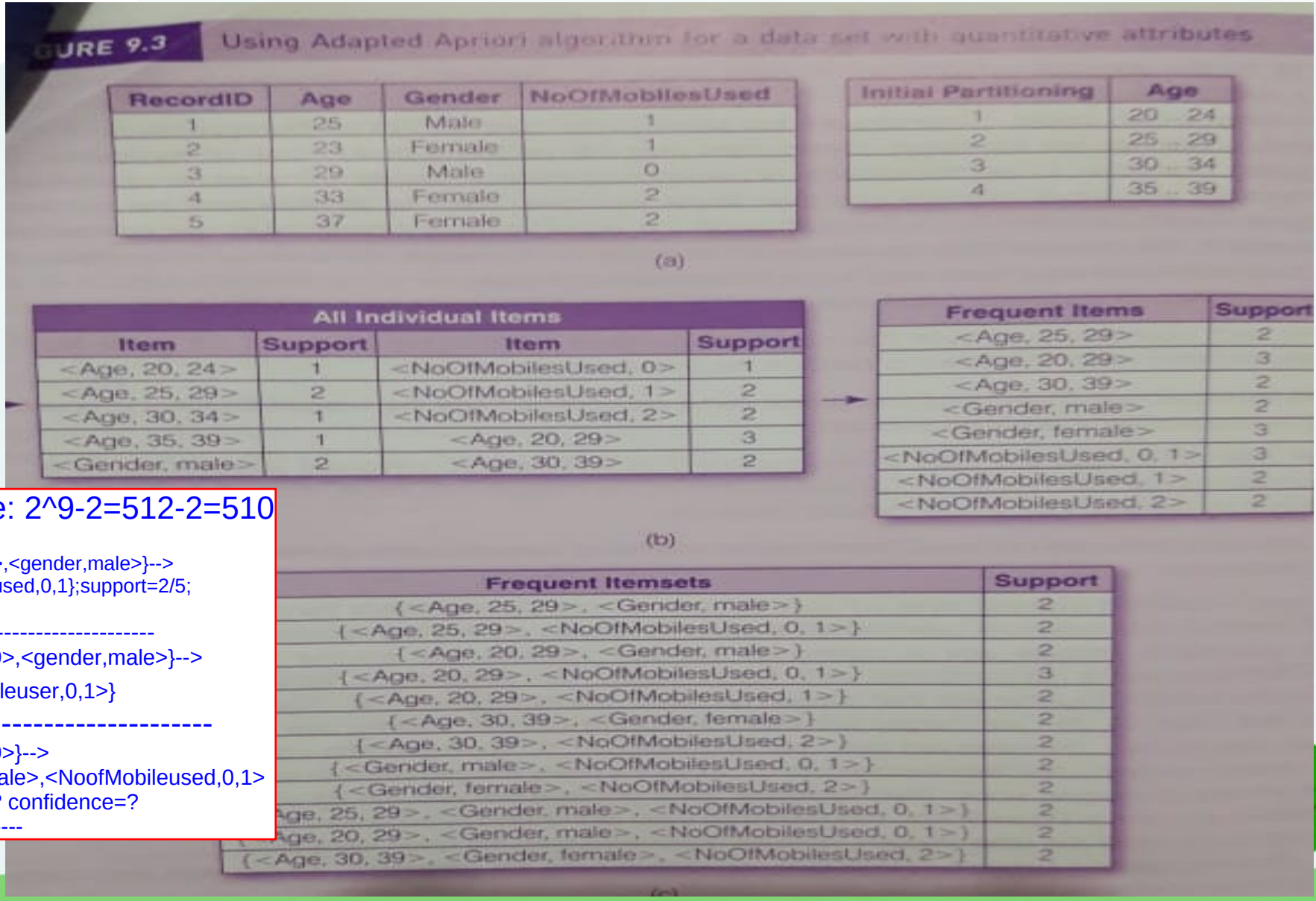
(a)

RecordID	Age: 20–29	Age: 30–39	Gender: Female	Gender: Male	NoOfMobilesUsed: 0–1	NoOfMobilesUsed: 2
1	1	0	0	1	1	0
2	1	0	1	0	1	0
3	1	0	0	1	1	0
4	0	1	1	0	0	1
5	0	1	1	0	0	1

(b)

RecordID	Items
1	<Age, 20, 29>, <Gender, Male>, <NoOfMobilesUsed, 0, 1>
2	<Age, 20, 29>, <Gender, Female>, <NoOfMobilesUsed, 0, 1>
3	<Age, 20, 29>, <Gender, Male>, <NoOfMobilesUsed, 0, 1>
4	<Age, 30, 39>, <Gender, Female>, <NoOfMobilesUsed, 2>
5	<Age, 30, 39>, <Gender, Female>, <NoOfMobilesUsed, 2>

Penerapan Apriori pada Atribut Kuantitatif



GURE 9.3 Using Adapted Apriori algorithm for a data set with quantitative attributes

RecordID	Age	Gender	NoOfMobilesUsed
1	25	Male	1
2	23	Female	1
3	29	Male	0
4	33	Female	2
5	37	Female	2

Initial Partitioning	Age
1	20 .. 24
2	25 .. 29
3	30 .. 34
4	35 .. 39

(a)

All Individual Items			
Item	Support	Item	Support
<Age, 20, 24>	1	<NoOfMobilesUsed, 0>	1
<Age, 25, 29>	2	<NoOfMobilesUsed, 1>	2
<Age, 30, 34>	1	<NoOfMobilesUsed, 2>	2
<Age, 35, 39>	1	<Age, 20, 29>	3
<Gender, male>	2	<Age, 30, 39>	2

Frequent Items	Support
<Age, 25, 29>	2
<Age, 20, 29>	3
<Age, 30, 39>	2
<Gender, male>	2
<Gender, female>	3
<NoOfMobilesUsed, 0, 1>	3
<NoOfMobilesUsed, 1>	2
<NoOfMobilesUsed, 2>	2

(b)

Frequent Itemsets	Support
{<Age, 25, 29>, <Gender, male>}	2
{<Age, 25, 29>, <NoOfMobilesUsed, 0, 1>}	2
{<Age, 20, 29>, <Gender, male>}	2
{<Age, 20, 29>, <NoOfMobilesUsed, 0, 1>}	3
{<Age, 20, 29>, <NoOfMobilesUsed, 1>}	2
{<Age, 30, 39>, <Gender, female>}	2
{<Age, 30, 39>, <NoOfMobilesUsed, 2>}	2
{<Gender, male>, <NoOfMobilesUsed, 0, 1>}	2
{<Gender, female>, <NoOfMobilesUsed, 2>}	2
{<Age, 25, 29>, <Gender, male>, <NoOfMobilesUsed, 0, 1>}	2
{<Age, 20, 29>, <Gender, male>, <NoOfMobilesUsed, 0, 1>}	2
{<Age, 30, 39>, <Gender, female>, <NoOfMobilesUsed, 2>}	2

(c)

total rule: 2^9-2=512-2=510

rule:

{<age,25,29>,<gender,male>}--> {NoofMobileused,0,1};support=2/5; conf=2/2=1

{<age,20,29>,<gender,male>}--> {<noOfMobileuser,0,1>}

{<age,25,29>}--> {<gender,male>,<NoofMobileused,0,1>}; support=? confidence=?

Apriori dengan Atribut Kuantitatif

- Atribut kuantitiatif dapat diolah dengan Apriori dengan syarat dilakukan pengelompokkan atribut
- Untuk atribut kategorial A dengan nilai $a_1,a_2,..,a_k$ dituliskan <A,a_i> dan diasosiasikan sebagai item
  - Contoh: <Gender, Female>
- Untuk atribut kuantitatif B dengan nilai yang memiliki range, range dibagi dalam beberapa partisi dan dituliskan <B,l,u> dimana l dan u adalah lower dan upper dari suatu nilai.
  - Contoh: <Age, 20,29> artinya: umur dari 20-29 tahun masuk dalam kriteria ini

Tahapan Apriori pada data Kuantitatif

- Untuk setiap atribut kuantitatif, jumlah partisi ditentukan dan dibuat diskritisasi dengan range nilai yang sama
- Seluruh data kuantitatif dan kategorial dikumpulkan dan nilai support dihitung berdasarkan jumlahnya. Kemudian frequent items terpilih
- Untuk menghindari rule dengan nilai support yang kecil maka nilai dari interval yang berisi nilai kecil maka digabung dengan interval yang lebih panjang

Permasalahan pada Apriori dengan Atribut Kuantitatif

- Jika atribut dibagi dengan banyak partisi dengan interval yang kecil maka support untuk setiap partisi kecil sehingga rule tidak dihasilkan dari nilai support yang kecil ini.
- Ketika interval terlalu besar, maka informasi dapat hilang ketika nilai setiap individu digabung dalam satu interval sehingga munculnya ketidak pastian dan mengurangi nilai confidence
-

• Solusi: menggunakan partial-completeness

- Partisi sebagian dibuat menggan interval yang pendek
- Sebagian partisi dibuat menggunakan interval yang lebih panjang
- Kemudian tentukan frekuensi yang sama setiap interval

Daftar Pustaka

- Data mining techniques and Application An Introduction (Hongbo Du, 2010)
- Introduction to Data mining (Pang Ning Tan, Michael Steinbach, Vipin Kumar, 2006)