

DM583

Data Mining

Frequent Itemsets and Association Rules



Outline

DM868 DM870 DSR04

Introduction

Item Sets

Arthur 7imek

Freg. Pattern Mining

Sets and Relations

Definitions and Problem-Statements

Monotonicity-Property of

Frequency The Apriori Algorithm

Example

Efficiency Association Rule

Feature Spaces Clustering - Basics Classification - Basics Bayesian Learning Learning with Dietributione Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

Mining Summary

Frequent Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

References



Recommended Reading

DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining

Sets and Relations

Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Mining Summary

Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Recommended Reading:

- Tan et al. [2006], Chapter 6; Tan et al. [2020], Ch. 4.
- Han et al. [2011], Chapter 6.
- Zaki and Meira Jr. [2020], Chapters 8+9.
- Witten et al. [2011], Chapter 4.5.
- Advanced topics: Aggarwal and Han [2014].



Outline

DM868 DM870 DSR04

Arthur 7imek

Introduction

Freg. Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces

Clustering - Basics

Classification - Basics

Bayesian Learning Learning with

Dietributione Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Frequent Pattern Mining

Sets and Relations



Definition of Sets

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of

Frequency The Apriori Algorithm

Example

Efficiency Association Rule

Minina

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

A set is a collection of objects (e.g., $S = \{1, 4, 9\}$).

The objects are said to be elements of the set (e.g., $1 \in S, 4 \in S, 9 \in S$). Each element is *unique*.

We can define sets

extensionally: by enumerating the elements that define the set (e.g., $S = \{1, 4, 9\}$).

intensionally: by characterizing the elements of the set,

- describing what condition (the "characteristic") function" of the set) holds for all the elements and only for the elements of the set (e.g. $S = \{x | \sqrt{x} \in \mathbb{N} \text{ AND } x < 15\}$ — read '|' as 'for which holds' or 'such that').
- ► The intensional definition typically resorts to a domain over which the set is defined (here: \mathbb{N}).



Properties of Sets

DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining

Sets and Relations

Definitions and Problem-Statements

Problem-Statemen

MonotonicityProperty of

Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with

Distributions
Entropy, Purity, and
(Non-)Linear Sep.
Ensemble Learning

References

A set can be finite (e.g., $S = \{1, 2, 3, ..., n\}$: $1 \in S, 2 \in S, 3 \in S, ..., n \in S$). In this case, the set has a finite size (*cardinality*), that is the number of elements of the set (e.g., |S| = n).

A set can be countably infinite. Example:

$$S = \{1, -1, 3, -3, 5, -5, \ldots\}$$

= $\{x | x = 2k + 1 \ OR \ x = -2k + 1, k \in \mathbb{N}_0\}$

- ightharpoonup A set can be uncountable (e.g., \mathbb{R}).
- ▶ A set can be empty: $S = \{\} = \emptyset$. |S| = 0.

Notations

DM868 DM870 DS804

Arthur Zimek

Introduction

Freg. Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning Learning with

Dietributione Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

logical operators:

\/: or

∧: and

¬: not

 \star : negation of x (e.g., \neq , \notin)

 $A \Rightarrow B$: if A, then B (short for $\neg (A \land \neg B)$)

 $A \leftarrow B$: if B, then A (short for $\neg (B \land \neg A)$)

 $A \iff B: A \Rightarrow B \land A \Leftarrow B$ ("iff", read: "if and only if",

also: \equiv , read: "is equivalent to")

- $ightharpoonup \exists x : p(x), \text{ means: there exists some } x \text{ such that } p(x)$
- $\forall x: p(x), \text{ means: for all } x \text{ holds } p(x)$
- ▶ subset: $T \subseteq S \equiv \forall x \in T : x \in S$
- ▶ proper subset: $T \subset S \equiv (\forall x \in T : x \in S) \land (\exists x \in S : x \notin T)$



Algebra of Sets

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations

Item Sets Definitions and

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining

Summary

Feature Spaces Clustering - Basics

Classification - Basics

Bayesian Learning Learning with Dietributione Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

An algebra is defined over a base set Ω , all sets involved in the algebra are subsets of Ω .

basic operations for $S, T \subseteq \Omega$:

union
$$S \cup T \equiv \{x | x \in S \lor x \in T\}$$

intersection $S \cap T \equiv \{x | x \in S \land x \in T\}$

complement $\bar{S} \equiv S^C \equiv \{x | x \notin S\}$

difference $S \setminus T \equiv \{x | x \in S \land x \notin T\}$

product $S \times T \equiv \{(x, y) | x \in S \land y \in T\}$

Powerset
$$\mathcal{P}(S) \equiv \wp(S) \equiv 2^S \equiv \{S' | S' \subseteq S\}$$

example Let $\Omega = \mathbb{N}$, $S = \{1, 2, 3\}$ and $T = \{2, 3, 4\}$ –

what are the values of all these expressions?



Tuple and Relations

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Minina

Summary Feature Spaces

Clustering - Basics

Classification - Basics

Bayesian Learning Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

n-tuple element of the (Cartesian) product of *n* sets:

$$S_1 \times S_2 \times S_3 \times \ldots \times S_n =$$

$$\{(a_1,a_2,a_3,\ldots,a_n)|a_1\in S_1,a_2\in S_2,a_3\in S_3,\ldots,a_n\in S_n\}$$

If $S_1 = S_2 = S_3 = ... = S_n$, we write:

$$S^n \equiv S_1 \times S_2 \times S_3 \times \ldots \times S_n$$

n-ary relation R is a set of n-tuples:

$$R(x_1,\ldots,x_n)\equiv(x_1,\ldots,x_n)\in R$$

characteristic function of $R \subseteq S_1 \times \ldots \times S_n$ (can also be seen as "predicate": "R is true/false")

$$S_1 \times \ldots \times S_n \to \{\mathsf{true}, \mathsf{false}\}$$

$$t \mapsto \begin{cases} \text{true} & \text{if } t \in R \\ \text{false} & \text{otherwise} \end{cases}$$



Properties of Homogeneous Relations

DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining

Sets and Relations

Item Sets Definitions and

Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Example

Efficiency Association Rule

Association Mining

Summary

Feature Spaces Clustering – Basics

Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Helerences

Relation $R \subseteq S_1 \times S_2 \times S_3 \times ... \times S_n$ is homogeneous iff $S_1 = S_2 = S_3 = ... = S_n$.

A binary homogeneous relation $R \subseteq S \times S$ is

reflexive iff $\forall x \in S : (x, x) \in R$

symmetric iff $\forall x, y \in S : (x, y) \in R \Rightarrow (y, x) \in R$

antisymmetric iff $\forall x, y \in S : (x, y) \in R \land (y, x) \in R \Rightarrow x = y$

transitive iff $\forall x, y, z \in S : (x, y) \in R \land (y, z) \in R \Rightarrow (x, z) \in R$

total iff $\forall x, y \in S : (x, y) \in R \lor (y, x) \in R$

Examples:

- '<' $\subseteq \mathbb{N} \times \mathbb{N}$: $(1,2) \in$ '<' (more customary is the infix notation 1 < 2)
- '= ' $\subseteq \mathbb{N} \times \mathbb{N}$: (361, 361) \in ' = ' or 361 = 361
 - \triangleright ' \subset ' $\subseteq \wp(S)^2$



Partial and Total Order

DM868 DM870 DS804

Arthur Zimek Introduction

Freg. Pattern Mining Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of

Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Minina

Summary

Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning

Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

A relation $R \in S \times S$ can be an order:

R is a partial order if R is antisymmetric and transitive.

R is a total order if R is a partial order and is total.

Are these relations orders? Which kind of order?

- alphanumeric sorting of strings?
- ightharpoonup '<' $\subset \mathbb{N} \times \mathbb{N}$
- ightharpoonup ' \leq ' \subseteq \mathbb{N} \times \mathbb{N}
- \triangleright ' \subset ' $\subset \wp(S)^2$
- \triangleright ' \subset ' $\subset \wp(S)^2$



Lattice of Subsets

DM868 DM870 DS804

Arthur Zimek

Introduction

Freg. Pattern Mining Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces Clustering - Basics

Classification - Basics

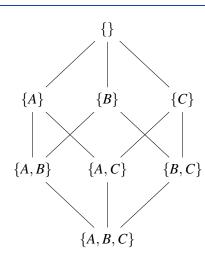
Bayesian Learning

Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning

References





Functions

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-Property of

Frequency The Apriori Algorithm Example

Efficiency Association Rule

Minina Summary

Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

A function is a mapping from some set (the domain) to some set (the image).

- We can see functions as relations with particular properties: a univalent (or right-unique) relation over domain and image.
- Formally, a function f is a binary relation over $D \times I$: $f \subseteq D \times I$, for which holds:

$$(x, y) \in f \land (x, z) \in f \Rightarrow y = z$$

i.e., for each $d \in D$, f maps to at most one $i \in I$.

Notation:

$$(x,y) \in f \iff y = f(x) \iff f(x) = y \iff f: x \mapsto y$$



Outline

DM868 DM870 DS804 Arthur Zimek

. . .

Introduction
Freq. Pattern Mining

Sets and Relations

Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of

Frequency
The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces Clustering – Basics

Classification – Basics

Bayesian Learning Learning with

Distributions Entropy, Purity, and (Non-)Linear Sep.

(Non-)Linear Sep. Ensemble Learning

References

Introduction

Frequent Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering – Basics and k-means

Classification - Basics and a Basic Classifier

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

Distributions and Learning with Distributions

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation

Ensemble Learning



Motivation

DM868 DM870 DS804

Arthur Zimek

Introduction

Freg. Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of

Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References









Learning: Data Mining,



What You Need to Know

Modeling about Data Mining and...) May Kuhn





> Yaser S. Abu-Mostafa



Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann...



on the webpage of some online bookseller:

Frequently Bought Together







Total price: \$188.39 Add all three to Cart Add all three to List

This item: Data Mining and Analysis: Fundamental Concepts and Algorithms by Mohammed J. Zaki Hardcover \$67.89

Data Mining: The Textbook by Charu C. Aggarwal Hardcover \$85.49

✓ Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking by Foster Provost Paperback \$35.01

Customers Who Bought This Item Also Bought











Data Science for Business:





食食食食 17 108

Page 1 of 17



Market Basket Analysis

DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining

Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining Summary

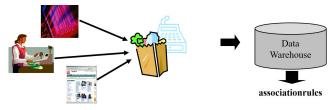
Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References



application examples:

- layout of supermarket: optimize the arrangement of items bought together
- online seller: recommend related items
- cross-marketing, add-on sales, targeted attached mailings

read about the infamous example of beer and diapers:

http://www.theregister.co.uk/2006/08/15/beer_diapers/





Frequent Patterns

DM868 DM870 DS804

Arthur Zimek Introduction

Freg. Pattern Mining Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning

Learning with Dietributione Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Transaction database:

```
T1:
     {bread, butter, milk, sugar}
T2:
     {butter, flour, milk, sugar}
T3:
     {butter, eggs, milk, salt}
T4:
     {eqqs}
T5:
     {butter, flour, milk, salt, sugar}
```

If we observe patterns, can we conclude on associations between items?



Associations

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining

Sets and Relations

Item Sets Definitions and

Problem-Statements

MonotonicityProperty of

Frequency
The Apriori Algorithm

THE APHOLIA

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces

Clustering - Basics

Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

a,**b,c,d**,e **b,c,d** a,**b,c,d** c,d,f a,**b,c,d**,e a,c,d a,c,e,f c,d,e,f a,**b,c,d**,f a,**b,c,d**,f

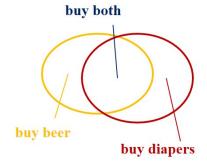
In 5 out of 10 (50%) cases, b,c,d

occur together.

In 5 cases we have **b,c**, and in all those 5 cases we also have **d**:

Rule with 100% confidence:

If **b,c** are in the set, then also **d** is in the set.





Outline

DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining Sets and Relations

Item Sets
Definitions and

Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

mo ripnom

Example

Efficiency

Association Bule

Mining

Summary

Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.

(Non-)Linear Sep. Ensemble Learning

References

Introduction

Frequent Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering – Basics and k-means

Classification - Basics and a Basic Classifie

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

Distributions and Learning with Distributions

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation

Ensemble Learning

Association Rules

```
DM868 DM870
   DS804
```

Arthur 7imek

Introduction

Freg. Pattern Mining Sets and Relations

Item Sets Definitions and

```
Problem-Statements
```

```
Monotonicity-
```

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

```
T1:
     {bread, butter, milk, sugar}
```

T2: {butter, flour, milk, sugar}

T3: {butter, eggs, milk, salt}

T4: {eaas}

T5: {butter, flour, milk, salt, sugar}

items bread, butter, eggs, milk etc.

transaction a database entry as a set of items

rule $L \Rightarrow R, L \cap R = \emptyset$ (disjoint sets of items)

l left-hand-side or antecedent

R right-hand-side or consequent

► {butter, flour} ⇒ {milk}



Definition: Frequent Itemsets

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements

Flobletti-Staterile

Monotonicity-Property of

Frequency

The Apriori Algorithm

Example

Efficiency

Efficiency

Association Rule

Mining Summary

Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

items: $I = \{i_1, \dots, i_m\}$ a set of literals (e.g., items in a shop)

itemset: $X \subseteq I$ (e.g., the items in a basket)

transaction: $T = (tid, X_{tid})$ designates a specific itemset

transaction database \mathcal{D} : a set of transactions

order: items in an itemset are ordered by some strict total order (e.g., alphabetical order of the literals), i.e.:

$$X = (x_1, x_2, \dots, x_k) \Rightarrow x_1 < x_2 < \dots < x_k$$

length of an itemset: number of elements contained in the itemset

k-itemset: an itemset of length k (e.g., T1 is a 4-itemset, T4 is a 1-itemset)



Definition: Cover, Support, Frequency

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations Item Sets

Definitions and Problem-Statements

Monotonicity-Property of

Frequency The Apriori Algorithm

Example

Efficiency

Association Rule

Minina

Summary Feature Spaces

Clustering - Basics

Classification - Basics Bayesian Learning Learning with

Dietributione Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

cover of an itemset: set of all transactions that contain the itemset: $cover(X) = \{(tid, X_{tid}) | (tid, X_{tid}) \in \mathcal{D} \land X \subseteq X_{tid}\}$ support of an itemset: the support s of an itemset X (S(X)) is the number of transactions containing X (i.e., the size of the cover set): $s(X) = |\operatorname{cover}(X)|$

frequency of an itemset: the frequency of an itemset X is its support relative to the database size $f(X) = \frac{s(X)}{|\mathcal{D}|}$

frequent itemset: given some support threshold σ , an itemset X is frequent (w.r.t. σ) iff: $s(X) > \sigma$ or equivalently $f(X) \geq \frac{\sigma}{|\mathcal{D}|}$



Problem 1: Frequent Itemset Mining

DM868 DM870 DS804

Arthur 7imek Introduction

Freg. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements

Monotonicity-Property of

Frequency The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering - Basics

Classification - Basics Bayesian Learning

Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning

References

Given:

- a set of items I
- ▶ a transaction database D over I
- ightharpoonup a support threshold σ

Find all frequent itemsets in \mathcal{D} , i.e., $\{X|X\subseteq I \land s(X)\geq \sigma\}$

example: which itemsets are frequent with $\sigma = 3$ in \mathcal{D} :

- T1: {bread, butter, milk, sugar}
- T2: {butter, flour, milk, sugar}
- T3: {butter, eggs, milk, salt}
- T4: {eqqs}
- T5: {butter, flour, milk, salt, sugar}



Definition: Association Rule

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets

Definitions and

Problem-Statements

Monotonicity-Property of

Frequency
The Apriori Algorithm

The Aprion Aigo

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces Clustering – Basics

Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

Ensemble Lear References

i telefelices

association rule: expresses an implication of the form $X \Rightarrow Y$, where X and Y are itemsets, $X \cap Y = \emptyset$

implication: describes a co-occurrence, not a causality

An association rule does not necessarily need to hold in all cases. We can describe its strength (or weakness), based on the observed cases:

support: The support of an association rule in $\mathcal D$ is the support of the union of its components:

$$s(X \Rightarrow Y) = s(X \cup Y)$$

frequency: Analogously, $f(X \Rightarrow Y) = f(X \cup Y)$

confidence: $conf(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$



Problem 2: Association Rule Mining

DM868 DM870 DS804 Arthur Zimek

Introduction

Freq. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements

Monotonicity-Property of

Frequency The Apriori Algorithm

Example Efficiency

Efficiency Association Rule

Mining Summary

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning

Given:

- a set of items I
- ightharpoonup a transaction database \mathcal{D} over I
- lacktriangle a support threshold σ and a confidence threshold c

Find all association rules $X \Rightarrow Y$ in \mathcal{D} with a support of at least σ and a confidence of at least c, i.e.:

$${X \Rightarrow Y | s(X \Rightarrow Y) \ge \sigma \land \operatorname{conf}(X \Rightarrow Y) \ge c}$$

T1: {bread, butter, milk, sugar}

T2: {butter, flour, milk, sugar}

T3: {butter, eggs, milk, salt}

T4: {eggs}

T5: {butter, flour, milk, salt, sugar}



Problem 1 ⊂ Problem 2

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining Sets and Relations

Item Sets
Definitions and

Problem-Statements

Monotonicity-Property of

Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Summary

Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

References

Problem 1 is part of Problem 2:

- Itemset A is frequent w.r.t. σ
- Given
 - $ightharpoonup A = X \cup Y$
 - $X \cap Y = \emptyset$
 - $ightharpoonup X = A \setminus Y$
 - $ightharpoonup Y = A \setminus X$
- ' $X \Rightarrow Y$ ' is frequent w.r.t. σ

Two-step approach:

- 1. find all frequent itemsets w.r.t. σ
- 2. generate rules with confidence $\geq c$ from each frequent itemset, where each rule is a binary partition of the itemset



Find Frequent Itemsets: Naïve Algorithm (Brute Force)

DM868 DM870 DS804

Arthur Zimek Introduction

Freg. Pattern Mining Sets and Relations

Item Sets Definitions and

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule

Mining Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning Learning with

Dietributione Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

T4: {eqqs}

Each possible itemset is a candidate for being a frequent itemset.

- We need to check the database to count if the itemset is actually frequent.
- Complexity roughly amounts to: number of candidates × number of transactions
- How many candidates do we have, given n items?

```
т1:
     {bread, butter, milk,
                             sugar}
T2:
     {butter, flour, milk, sugar}
T3:
     {butter, eggs, milk, salt}
```

T5: {butter, flour, milk, salt, sugar}



Outline

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Belations

Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of

Frequency
The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Mining Summary

Feature Spaces Clustering – Basics

Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.

References

Introduction

Frequent Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering – Basics and k-means

Classification - Basics and a Basic Classifie

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

Distributions and Learning with Distributions

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation

Ensemble Learning



Lattice of Itemsets

A(5)

AD(3)

ABE(2)

AC(4)

ABC(2) ABD(1)

AB(3)

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations Item Sets Definitions and

Problem-Statements

Monotonicity-Property of

Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining

Summary Feature Spaces

Clustering - Basics

Classification - Basics Bayesian Learning Learning with

Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

ABCD(1) ABCE(2)

B(4)

AÈ(4)

ACD(3)

ABDE(1)

D(4)

BE(2)

ACE(4) ADE(3) BCD(1) BCE(2) BDE(1) CDE(3)

CE(4)

ACDE(3) BCDE(1)

CD(4)

TID

4

6

E(4)

Α В

0

0 0

0

0

DE(3)

0 0 0

0

0

ABCDE(1)

{}(7)

C(5)

BD(1)

BČ(2)

71



DM868 DM870 DS804

Arthur Zimek

Freg. Pattern Mining Item Sets Definitions and Problem-Statements Monotonicity-

Property of

Frequency

Efficiency Association Rule Minina Summary Feature Spaces

Clustering - Basics Classification - Basics Bayesian Learning Learning with Dietributione

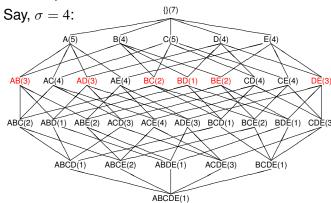
Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

Introduction

Sets and Relations

The Apriori Algorithm Example

- If X is frequent, all subsets $X' \subseteq X$ are frequent as well.
- If X is not frequent, neither any superset $X' \supset X$ can be frequent.





DM868 DM870 DS804

Introduction

Arthur Zimek

Freq. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements

Monotonicity-Property of

Frequency

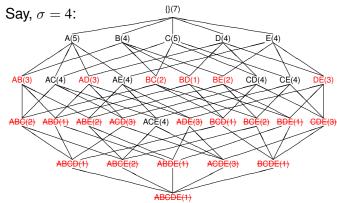
Efficiency
Association Rule
Mining
Summary
Feature Spaces

Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with
Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

The Apriori Algorithm Example

- ▶ If *X* is frequent, all subsets $X' \subseteq X$ are frequent as well.
- ▶ If *X* is not frequent, neither any superset $X' \supseteq X$ can be frequent.





DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
Monotonicity-

Property of

Frequency

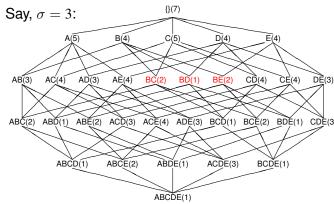
Efficiency
Association Rule
Mining
Summary
Feature Spaces

Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with
Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

The Apriori Algorithm Example

- ▶ If *X* is frequent, all subsets $X' \subseteq X$ are frequent as well.
- ▶ If *X* is not frequent, neither any superset $X' \supseteq X$ can be frequent.





DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
Monotonicity-

The Apriori Algorithm Example

Property of

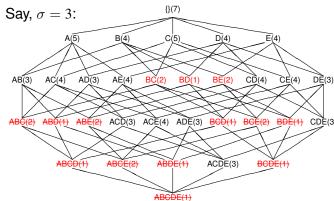
Frequency

Efficiency
Association Rule
Mining
Summary
Feature Spaces

Clustering – Basics
Classification – Basics
Bayesian Learning
Learning with
Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

- ▶ If *X* is frequent, all subsets $X' \subseteq X$ are frequent as well.
- ▶ If *X* is not frequent, neither any superset $X' \supseteq X$ can be frequent.





Maximal Frequent Itemsets

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning Due to the anti-monotonicity, we can summarize solutions by their *border* in the lattice:

- ▶ An itemset *X* belongs to the border (w.r.t. some σ), if:
 - ▶ $\forall Y \subset X : Y \text{ is frequent (w.r.t. } \sigma)$
 - ▶ $\forall Z \supset X : Z \text{ is not frequent (w.r.t. } \sigma)$
- positive border: X itself is frequent
- ▶ such frequent itemsets are called **maximal frequent itemsets**
- ▶ they are frequent itemsets with no frequent supersets

Maximal frequent itemsets can be used as a condensed representation of a solution, as all frequent itemsets can be derived from the maximal frequent itemsets.

Note that:

The anti-monotonicity property of support is also called downward-closure property.



Maximal Frequent Itemsets

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule Minina

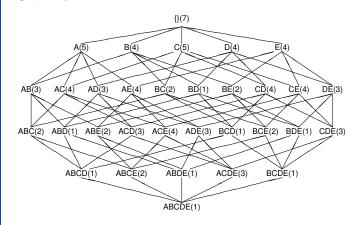
Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Example: find the positive border (maximal frequent itemsets) for $\sigma = 3$





Maximal Frequent Itemsets

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Minina

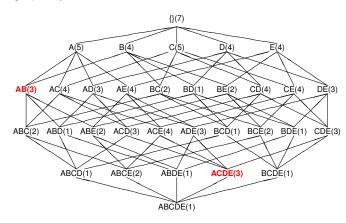
Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Example: find the positive border (maximal frequent itemsets) for $\sigma = 3$





Closed (Frequent) Itemsets

DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining Sets and Relations Item Sets Definitions and

Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm Example Efficiency

Association Rule Mining

Summary Feature Spaces

Clustering – Basics Classification – Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

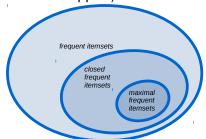
closed itemset: An itemset *X* is *closed* if none of its immediate supersets has exactly the same support as *X*.

closed frequent itemset: An itemset is a closed frequent itemset (w.r.t. some σ) if it is a closed itemset and is frequent (w.r.t. σ).

Closed frequent itemsets (including their support) represent a solution (all frequent itemsets and their support).

CFIs can also generate all frequent itemsets since they necessarily **contain** the MFIs

The supports of all frequent itemsets can also be retrieved from the CFIs' supports





Outline

DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining

Sets and Relations

Definitions and

Monotonicity-Property of

Frequency
The Apriori Algorithm

The Aprion Aig

Example

Efficiency Association Rule

Mining

Summary Feature Spaces

Clustering - Basics

Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Fnsemble Learning

References

Introduction

Frequent Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering – Basics and k-means

Classification - Basics and a Basic Classifie

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

Distributions and Learning with Distributions

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation

Ensemble Learning



Apriori Algorithm [Srikant and Agrawal, 1996]: Idea

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

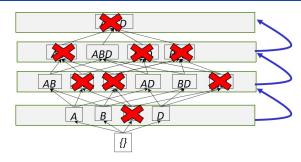
The Apriori Algorithm Example

Example Efficiency Association Rule Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References



- 1. find frequent 1-itemsets first, then 2-itemsets, 3-itemsets etc. (breadth-first search in the lattice)
- 2. for finding (k+1)-itemsets C_{k+1} : consider only those as candidates, where *all* k-itemsets $C_k \subset C_{k+1}$ are frequent
- 3. count frequency of all *k*-itemset candidates in a single database scan (hashing of the candidate itemsets)



Apriori Algorithm [Srikant and Agrawal, 1996]: Pseudo Code

DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining Sets and Relations

Sets and Relations
Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of

Frequency
The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

....

 C_k : k-itemset candidates, S_k : frequent k-itemsets (solution)

Algorithm 2.1 (Apriori [Srikant and Agrawal, 1996])

```
Apriori(I, \mathcal{D}, \sigma)
 S_1 = \{frequent 1-itemsets\};
 k=2:
 while S_{k-1} \neq \emptyset do
    C_k = AprioriGenerateCandidates(S_{k-1});
     for each transaction T \in \mathcal{D} do
        C_T = \{c \in C_k | c \subseteq T\};
        for each c \in C_T do
           c.count++;
    S_k = \{c \in C_k | c.count > \sigma\};
    k++;
 return \bigcup_{\iota} S_{\iota}:
```



Apriori Candidate Generation

```
DM868 DM870
  DS804
```

Arthur Zimek Introduction

Freg. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Minina

Summary Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

1. *join:*

two frequent (k-1)-itemsets $p, q \in S_{k-1}$ are joined if they

are identical in the **first** (order!) k-2 items:

 $p \in S_{k-1}$: (A, B, C)

 $q \in S_{k-1}$: (A, B, D)

Algorithm 2.2 (AprioriGenerateCandidates(S_{k-1}))

 $\Rightarrow (A, B, C, D) \in C_k$

2. pruning:

remove all k-itemsets from Ck that contain any

(k-1)-itemset $\not\in S_{k-1}$

example: $S_3 = \{(1,2,3), (1,2,4), (1,3,4), (1,3,5), (2,3,4)\}$ 1. join: $C_4 = \{(1,2,3,4), (1,3,4,5)\}$

2. pruning: remove (1, 3, 4, 5) result: $C_4 = \{(1, 2, 3, 4)\}$



Outline

DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining

Sets and Relations

Item Sets Definitions and

Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency
Association Rule
Mining

Summary Feature Spaces

Clustering - Basics

Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Introduction

Frequent Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering – Basics and k-means

Classification - Basics and a Basic Classifie

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

Distributions and Learning with Distributions

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation

Ensemble Learning



Apriori Example: Find Frequent Itemsets X with $f(X) \ge 0.3$

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-Property of

Frequency
The Apriori Algorithm

Example

Efficiency
Association Rule
Mining
Summary

Feature Spaces
Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

SDU 6 C1 and S1

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

Α	
В	
С	
D	
Е	
F	
G	
Н	
I	
J	
K	
L	



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH ABCEGH 2 3 ABCEFH BCDEFGHL 4 ABEKH 5 BEFGHIK 6 ABDGH 8 ABDG 9 BDFG 10 CEFACEFH 11 ABEG 12

C1

Α R ח E G Н K



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets

Definitions and Problem-Statements Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database



O i		
Α	1	
В	2	
С	1	
D		
Е	2	
F		
G	2	
Н	2	
I		
J		
K		
Ĺ		

SDU 6 C1 and S1

DM868 DM870 DS804

Arthur 7imek

Introduction Freg. Pattern Mining Sets and Belations Item Sets

Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary

Feature Spaces Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH ABCEFH BCDEFGHL	
3		
4		
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

Α	2
В	3 2
С	2
D	
Е	3
F	1
G	2
Н	3
I	
J K	
K	
L	

SDU 6 C1 and S1

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations

Item Sets Definitions and

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

Α	2
В	2 4
С	3
D	1
E F	4
F	2
G	3
Н	4
I	
J K	
K	
L	1

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency
Association Rule

Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH ABCEGH 2 3 ABCEFH BCDEFGHL 4 ABEKH 5 BEFGHIK 6 ABDGH 8 ABDG 9 BDFG 10 CEFACEFH 11 ABEG 12

C1

_	-
Α	3
В	5 3
С	3
D	1
Е	5
F	
G	3 5
Н	5
I	
J	
K	1
Ĺ	1

SDU 6 C1 and S1

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Belations Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

_	
Α	3
В	6
С	3
D	1
Е	6
F	ფ
G	4
Н	6
ı	1
J	
K	2
L	1

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example

Efficiency Association Rule Mining

Mining Summary Feature Spaces

Clustering – Basics Classification – Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH ABCEGH 2 3 ABCEFH 4 BCDEFGHL ABEKH 5 BEFGHIK 6 ABDGH 8 ABDG 9 BDFG 10 CEFACEFH 11 ABEG 12

C₁

4 Α R 3 2 D 6 3 5 G Н K 2

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining

Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning

Database

BEGH 2 ABCEGH 3 ABCEFH 4 BCDEFGHL ABEKH 5 BEFGHIK 6 ABDGH 8 ABDG 9 BDFG 10 CEFACEFH 11 ABEG 12

C1

5 Α R 8 3 D 3 6 3 G 6 Н K 2

SDU 6 C1 and S1

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Belations

Item Sets Definitions and

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Database

	1	B E G H A B C E G H	
	2		
	3	ABCEFH	
	4	BCDEFGHL	
	5	ABEKH	
	6	BEFGHIK	
	7	ABDGH	
	8	ABDG	
þ	9	B D F G C E F	
	10		
	11	ACEFH	
	12	ABEG	

Α	5
В	5 9 3
С	3
D	4
Е	6
F	4
G	7
Н	7
I	1
J	
K	2
L	1

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example
Efficiency
Association Rule
Mining

Mining Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Helerences

Database

	1	BEGH
	2	ABCEGH
	3	ABCEFH
	4	BCDEFGHL
	5	ABEKH
	6	BEFGHIK
	7	ABDGH
	8	ABDG
	9	BDFG
\rightarrow	10	CEF
	11	ACEFH
	12	ABEG

Α	5 9
В	9
С	4
D	4
Ε	7
F	5
G	7
Н	7
I	1
J	
K	2
L	1

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example
Efficiency
Association Rule

Mining Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep.
Ensemble Learning

neierences

Database

BEGH 2 ABCEGH 3 ABCEFH 4 BCDEFGHL ABEKH 5 BEFGHIK 6 ABDGH 8 ABDG 9 BDFG 10 CEFACEFH 11 ABEG 12

C1

6 Α R 9 5 D 4 8 F 6 G Н 8 K 2

SDU 6 C1 and S1

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Belations Item Sets Definitions and Problem-Statements Monotonicity-Property of

Frequency The Apriori Algorithm

Example Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

C1

O i	
Α	7
В	10
С	5
D	4
Е	9
F	6
G	8
Н	8
ı	1
J	
K	2
Ĺ	1

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example

Efficiency
Association Rule
Mining

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

Database

BEGH 2 ABCEGH 3 ABCEFH 4 BCDEFGHL 5 ABEKH 6 BEFGHIK ABDGH 8 ABDG 9 BDFG 10 $C \in F$ ACEFH 11 ABEG 12

S1

 $\sigma = 30\% \Leftrightarrow \text{support} \ge 4$

SDU C2 and S2

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

S₁ В D Ε G

C2			
AB		CE	
AC		CF	
AD		CG	
AE		СН	
AF		DE	
AG		DF	
AH		DG	
ВС		DH	
BD		EF	
BE		EG	
BF		EH	
BG		FG	
ВН		FH	
CD		GH	

DM868 DM870 DS804

Arthur 7imek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and

Problem-Statements Monotonicity-Property of

Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH 2 ABCEGH 3 ABCEFH BCDEFGHL 4 5 ABEKH BEFGHIK 6 ABDGH 8 ABDG 9 BDFG CEF 10 ACEFH 11 12 ABEG

S1Α В C D Ε

F G Н

C2			
AB		CE	
AC		CF	
AD		CG	
AE		СН	
AF		DE	
AG		DF	
AH		DG	
ВС		DH	
BD		EF	
BE	1	EG	1
BF		EH	1
BG	1	FG	
BH	1	FH	
CD		GH	1



DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining Sets and Relations

Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example
Efficiency
Association Rule

Mining Summary

Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH 2 ABCEGH 3 ABCEFH BCDEFGHL 4 5 ABEKH BEFGHIK 6 ABDGH 8 ABDG 9 BDFG CEF 10 ACEFH 11 12 ABEG

S1
A
B
C
D
E
F
G
H

G2			
AB	1	CE	1
AC	1	CF	
AD		CG	1
AE	1	СН	1
AF		DE	
AG	1	DF	
AH	1	DG	
BC	1	DH	
BD		EF	
BE	2	EG	2
BF		EH	2
BG	2	FG	
BH	2	FH	
CD		GH	2

 \sim



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements

Problem-Stateme Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

.....

Database

BEGH
ABCEGH
ABCEFH
BCDEFGHL
ABEKH
BEFGHIK
ABDGH
ABDG
BDFG
CEF
ACEFH
ABEG

S₁ Α В D Ε G Н

C2			
AB	2	CE	2
AC	2	CF	1
AD		CG	1
AE	2	СН	2
AF	1	DE	
AG	1	DF	
АН	2	DG	
ВС	2	DH	
BD		EF	1
BE	3	EG	2
BF	1	EH	3
BG	2	FG	
ВН	3	FH	1
CD		GH	2

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of

Frequency
The Apriori Algorithm

The Apriori Algo Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
) 4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

.

S1
A
B
C
D
E
F
G

C2

AB	2	CE	3
AC	2	CF	2
AD		CG	2
AE	2	СН	3
AF	1	DE	1
AG	1	DF	1
АН	2	DG	1
ВС	3	DH	1
BD	1	EF	2
BE	4	EG	3
BF	2	EH	4
BG	3	FG	1
ВН	4	FH	2
CD	1	GH	3

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of

Frequency
The Apriori Algorithm

Example
Efficiency
Association Rule
Mining

Mining Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

(Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

S1
A
B
C
D
E
F
G
H

C2			
AB	3	CE	3
AC	2	CF	2
AD		CG	2
AE	3	СН	3
AF	1	DE	1
AG	1	DF	1
AH	3	DG	1
ВС	3	DH	1
BD	1	EF	2
BE	5	EG	3
BF	2	EH	5
BG	3	FG	1
ВН	5	FH	2
CD	1	GH	3

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency
The Apriori Algorithm

Example Efficiency Association Rule

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	
	2 3 4 5 6 7 8 9 10	

S1
A
B
C
D
E
F
G
H

C2			
AB	3	CE	3
AC	2	CF	2
AD		CG	2
AE	3	СН	3
AF	1	DE	1
AG	1	DF	1
AH	3	DG	1
ВС	3	DH	1
BD	1	EF	3
BE	6	EG	4
BF	3	EH	6
BG	4	FG	2
BH	6	FH	3
CD	1	GH	4

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database

BEGH	
ABCEGH	
ABCEFH	
BCDEFGHL	
ABEKH	
BEFGHIK	
ABDGH	
ABDG	
BDFG	
CEF	
ACEFH	
ABEG	

 C_2

S₁

Α

В

D

Ε

F

G

Н

		_	
AB	4	CE	3
AC	2	CF	2
AD	1	CG	2
AE	3	СН	3
AF	1	DE	1
AG	2	DF	1
AH	4	DG	2
ВС	3	DH	2
BD	2	EF	3
BE	6	EG	4
BF	3	EH	6
BG	5	FG	2
BH	7	FH	3
CD	1	GH	5

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of

Frequency
The Apriori Algorithm

Example
Efficiency
Association Rule
Mining

Summary Feature Spaces

Clustering – Basics
Classification – Basics
Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

	1	BEGH	
	2	ABCEGH	
	3	ABCEFH	
	4	BCDEFGHL	
	5	ABEKH	
	6	BEFGHIK	
	7	ABDGH	
\rightarrow	8	ABDG	
	9	BDFG	
	10	CEF	
	11	ACEFH	
	12	ABEG	

S1
A
B
C
D
E
F
G
H

C2			
AB	5	CE	3
AC	2	CF	2
AD	2	CG	2
AE	3	СН	3
AF	1	DE	1
AG	3	DF	1
AH	4	DG	3
BC	3	DH	2
BD	3	EF	3
BE	6	EG	4
BF	3	EH	6
BG	6	FG	2
BH	7	FH	3
CD	1	GH	5

~~

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example
Efficiency
Association Rule
Mining

Mining Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning

References

Database

BEGH	
ABCEGH	
ABCEFH	
BCDEFGHL	
ABEKH	
BEFGHIK	
ABDGH	
ABDG	
BDFG	
CEF	
ACEFH	
ABEG	

S1
A
B
C
D
E
F

Н

C2			
AB	5	CE	3
AC	2	CF	2
AD	2	CG	2
AE	3	СН	3
AF	1	DE	1
AG	3	DF	2
АН	4	DG	4
ВС	3	DH	2
BD	4	EF	3
BE	6	EG	4
BF	4	EH	6
BG	7	FG	3
ВН	7	FH	3
CD	1	GH	5

~~

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

	1	BEGH
	2	ABCEGH
	3	ABCEFH
	4	BCDEFGHL
	5	ABEKH
	6	BEFGHIK
	7	ABDGH
	8	ABDG
	9	BDFG
\rightarrow	10	CEF
	11	ACEFH
	12	ABEG

S₁ Α В

D Ε G Н C2

AB	5	CE	4
AC	2	CF	3
AD	2	CG	2
AE	3	СН	3
AF	1	DE	1
AG	3	DF	2
AH	4	DG	4
ВС	3	DH	2
BD	4	EF	4
BE	6	EG	4
BF	4	EH	6
BG	7	FG	3
ВН	7	FH	3
CD	1	GH	5

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of

Frequency
The Apriori Algorithm

Example Efficiency

Association Rule Mining Summary

Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Distributions
Entropy, Purity, and
(Non-)Linear Sep.
Ensemble Learning

References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
) 11	ACEFH
12	ABEG

S1

A B C D E F G C2

AB	5	CE	5
AC	3	CF	4
AD	2	CG	2
AE	4	СН	4
AF	2	DE	1
AG	3	DF	2
AH	5	DG	4
ВС	3	DH	2
BD	4	EF	5
BE	6	EG	4
BF	4	EH	7
BG	7	FG	3
ВН	7	FH	4
CD	1	GH	5

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example
Efficiency
Association Rule

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

.....

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

S1
A
B
C
D
E
F
G
H

C2 CE AB 6 5 AC 3 CF 4 AD 2 CG 2 AΕ 5 CH 4 AF DE AG DF 2 4 AΗ 5 DG 4 BC DH 3 2 EF 5 BD 4 BE EG 5 BF 4 EΗ BG 8 FG 3 BH FΗ 4 5 GH

SDU 6 C2 and S2

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

S₁ Α В D Ε G Н

S2				
AB	6	CE	5	
AC	3	CF	4	
AD	2	CG	2	
AE	5	СН	4	
AF	2	DE	4	
AG	4	DF	2	
АН	5	DG	4	
ВС	3	DH	2	
BD	4	EF	5	
BE	7	EG	5	
BF	4	EH	7	
BG	8	FG	3	
ВН	7	FH	4	
CD	4	GH	5	



S2 to C3

DM868 DM870 DS804

Arthur Zimek
Introduction

Freq. Pattern Mining Sets and Relations

Item Sets
Definitions and

Problem-Statements
MonotonicityProperty of

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining

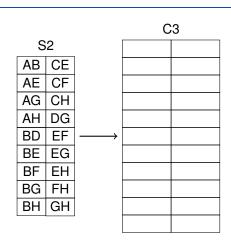
Summary Feature Spaces

Clustering - Basics

Classification – Basics Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining

Sets and Relations Item Sets Definitions and

Problem-Statements
MonotonicityProperty of

Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule Mining

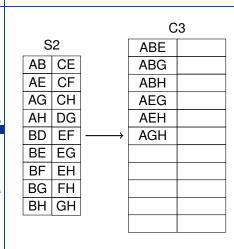
Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

References



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets

Definitions and Problem-Statements Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency
Association Rule
Mining

Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning References

i telefelices

S2 CE AB CF AΕ AG CH AΗ DG BD EF EG BE BF EΗ BG FΗ BH GH

C3 ABE BEG **ABG BEH ABH BFG AEG BFH BGH** AEH **AGH BDE BDF BDG BDH** BEF

DM868 DM870 DS804

Arthur Zimek
Introduction

Freq. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency
Association Rule
Mining

Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

S2
AB CE
AE CF
AG CH
AH DG
BD EF
BE EG

EΗ

FΗ

GH

BF

BG

BH

C3 ABE BEG ABG **BEH ABH BFG AEG BFH BGH** AEH **AGH CEF BDE** CEH **BDF CFH BDG BDH BEF**

DM868 DM870 DS804

Arthur Zimek Introduction

Freg. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

S2 AB CE CF AΕ CH AG AΗ DG **EF** BD BE EG BF EΗ FΗ BG GH

BH

C3 BEG ABE ABG BEH **ABH BFG BFH** AEG AEH **BGH** CEF **AGH BDE** CEH **BDF CFH EFG BDG EFH BDH BEF EGH**

DM868 DM870 DS804

Arthur Zimek
Introduction

Freq. Pattern Mining Sets and Relations

Item Sets

Definitions and

Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

S2 AB CE CF AΕ CH AG AΗ DG **EF** BD BE EG BF EΗ FΗ BG GH BH

C3 BEG ABE ABG BEH **ABH BFG BFH** AEG AEH **BGH** CEF **AGH** BDE CEH **BDF CFH BDG EFG EFH BDH BEF EGH**



DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

ABE	BEH	
ABG	BFH	
ABH	BGH	
AEG	CEF	
AEH	CEH	
AGH	CFH	
BDG	EFH	
BEF	EGH	
BEG		



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering – Basics Classification – Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

Ensemble Lea References

Database



ABE		BEH	1
ABG		BFH	
ABH		BGH	1
AEG		CEF	
AEH		CEH	
AGH		CFH	
BDG		EFH	
BEF		EGH	1
BEG	1		



DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database

BEGH	
ABCEGH	
ABCEFH	
BCDEFGHL	
ABEKH	
BEFGHIK	
ABDGH	
ABDG	
BDFG	
CEF	
ACEFH	
ABEG	

ABE	1	BEH	2
ABG	1	BFH	
ABH	1	BGH	2
AEG	1	CEF	
AEH	1	CEH	1
AGH	1	CFH	
BDG		EFH	
BEF		EGH	2
BEG	2		



DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and

Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH		
ABCEGH		
ABCEFH		
BCDEFGHL		
ABEKH		
BEFGHIK		
ABDGH		
ABDG		
BDFG		
CEF		
ACEFH		
ABEG		

	•	•	
ABE	2	BEH	3
ABG	1	BFH	1
ABH	2	BGH	2
AEG	1	CEF	1
AEH	2	CEH	2
AGH	1	CFH	1
BDG		EFH	1
BEF	1	EGH	2
BEG	2		



DM868 DM870 DS804

Arthur Zimek

Introduction

Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH	
ABCEGH	
ABCEFH	
BCDEFGHL	
ABEKH	
BEFGHIK	
ABDGH	
ABDG	
BDFG	
CEF	
ACEFH	
ABEG	

ABE	2	BEH	4
ABG	1	BFH	2
ABH	2	BGH	3
AEG	1	CEF	2
AEH	2	CEH	3
AGH	1	CFH	2
BDG	1	EFH	2
BEF	2	EGH	3
BEG	3		

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

ABE	3	BEH	5
ABG	1	BFH	2
ABH	3	BGH	3
AEG	1	CEF	2
AEH	3	CEH	3
AGH	1	CFH	2
BDG	1	EFH	2
BEF	2	EGH	3
BEG	3		

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining Summary

Feature Spaces Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database

BEGH
ABCEGH
ABCEFH
BCDEFGHL
ABEKH
BEFGHIK
ABDGH
ABDG
BDFG
CEF
ACEFH
ABEG

ABE	3	BEH	6
ABG	1	BFH	3
ABH	3	BGH	4
AEG	1	CEF	2
AEH	3	CEH	3
AGH	1	CFH	2
BDG	1	EFH	3
BEF	3	EGH	4
BEG	4		

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database

	1	BEGH
	2	ABCEGH
	3	ABCEFH
	4	BCDEFGHL
	5	ABEKH
	6	BEFGHIK
÷	7	ABDGH
	8	ABDG
	9	BDFG
	10	CEF
	11	ACEFH
	12	ABEG

ABE	3	BEH	6
ABG	2	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	2	EFH	3
BEF	3	EGH	4
BEG	4		

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example

Efficiency Association Rule Mining

Summary Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning References Database

	1	BEGH
	2	ABCEGH
	3	ABCEFH
	4	BCDEFGHL
	5	ABEKH
	6	BEFGHIK
	7	ABDGH
÷	8	ABDG
	9	BDFG
	10	CEF
	11	ACEFH
	12	ABEG
→	7 8 9 10 11	ABDGH ABDG BDFG CEF ACEFH

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	3	EFH	3
BEF	3	EGH	4
BEG	4		

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example Efficiency Association Rule Mining

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1 BEGH 2 ABCEGH 3 ABCEFH 4 BCDEFGHL 5 ABEKH 6 BEFGHIK
3 ABCEFH 4 BCDEFGHL 5 ABEKH
4 BCDEFGHL 5 ABEKH
5 A B E K H
6 BEEGHIK
0 DEI GIIIK
7 ABDGH
8 ABDG
9 BDFG
10 CEF
11 ACEFH
12 ABEG

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	2
AEH	3	CEH	3
AGH	2	CFH	2
BDG	4	EFH	3
BEF	3	EGH	4
BEG	4		



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of

Frequency The Apriori Algorithm

Example
Efficiency
Association Rule
Mining

Mining Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

Ensemble Lear References

Database

BEGH
ABCEGH
ABCEFH
BCDEFGHL
ABEKH
BEFGHIK
ABDGH
ABDG
BDFG
CEF
ACEFH
ABEG

ABE	3	BEH	6
ABG	3	BFH	3
ABH	4	BGH	5
AEG	1	CEF	3
AEH	3	CEH	3
AGH	2	CFH	2
BDG	4	EFH	3
BEF	3	EGH	4
BEG	4		



DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

	1	BEGH
	2	ABCEGH
	3	ABCEFH
	4	BCDEFGHL
	5	ABEKH
	6	BEFGHIK
	7	ABDGH
	8	ABDG
	9	BDFG
	10	CEF
\rightarrow	11	ACEFH
	12	ABEG

3	BEH	6
3	BFH	3
4	BGH	5
1	CEF	4
4	CEH	4
2	CFH	3
4	EFH	4
3	EGH	4
4		
	3 4 1 4 2 4 3	3 BFH 4 BGH 1 CEF 4 CEH 2 CFH 4 EFH 3 EGH



DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics Classification - Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

	1	BEGH
	2	ABCEGH
	3	ABCEFH
	4	BCDEFGHL
	5	ABEKH
	6	BEFGHIK
	7	ABDGH
	8	ABDG
	9	BDFG
	10	CEF
	11	ACEFH
\rightarrow	12	ABEG

	ABE	4	BEH	6
	ABG	4	BFH	3
	ABH	4	BGH	5
	AEG	2	CEF	4
Ī	AEH	4	CEH	4
	AGH	2	CFH	3
	BDG	4	EFH	4
	BEF	3	EGH	4
Ī	BEG	5		



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

S3

4	BEH	6
4	BFH	3
4	BGH	5
2	CEF	4
4	CEH	4
2	CFH	3
4	EFH	4
3	EGH	4
5		
	4 4 2 4 2 4 3	4 BFH 4 BGH 2 CEF 4 CEH 2 CFH 4 EFH 3 EGH

DM868 DM870 DS804

Arthur Zimek
Introduction

Freq. Pattern Mining Sets and Relations Item Sets

Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining Summary

Summary Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning
Learning with
Distributions

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

References

S3 ABE **ABG ABH** C4 **AEH** BDG **BEG BEH BGH CEF** CEH **EFH EGH**

EGH

DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining Sets and Relations

Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of

Property of Frequency The Apriori Algorithm

Example

Efficiency
Association Rule
Mining

Mining Summary Feature Spaces

Clustering – Basics

Classification – Basics Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

References

S3 ABE **ABG ABH** C4 **AEH ABEG** BDG **ABEH BEG ABGH** BEH **BGH CEF** CEH **EFH**

EFH

EGH

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations

Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of

Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Distributions
Entropy, Purity, and
(Non-)Linear Sep.
Ensemble Learning

References

S3 ABE **ABG ABH** C4 **AEH ABEG** BDG **ABEH BEG ABGH** BEH **BEGH BGH CEF** CEH

EGH

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations

Item Sets Definitions and Problem-Statements

Monotonicity-Property of

Frequency The Apriori Algorithm

Example

Efficiency Association Rule Mining

Summary Feature Spaces

Clustering - Basics

Classification - Basics Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning References

S3 ABE **ABG ABH** C4 AEH **ABEG** BDG **ABEH** BEG **ABGH** BEH **BEGH** BGH CEFH **CEF** CEH **EFH**

DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining

Sets and Relations Item Sets Definitions and

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

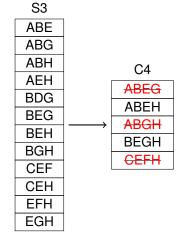
Summary
Feature Spaces

Clustering - Basics

Classification – Basics Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References



AEG, AGH and CFH not frequent!

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

ABEH	
BEGH	



C4 and S4

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of

Frequency The Apriori Algorithm

Example
Efficiency
Association Rule

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database



ABEH	
BEGH	1



C4 and S4

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and

Problem-Statements

MonotonicityProperty of

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining

Summary
Feature Spaces
Clustering – Basics

Classification – Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database



ABEH	1
BEGH	2



DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Belations Item Sets

Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary

Feature Spaces Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

Ensemble Learning References

(Non-)Linear Sep.

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG

ABEH	2
BEGH	2

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Database

BEGH
ABCEGH
ABCEFH
BCDEFGHL
ABEKH
BEFGHIK
ABDGH
ABDG
BDFG
CEF
ACEFH
ABEG

ABEH	2
BEGH	3

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Belations

Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

BEGH
ABCEGH
ABCEFH
BCDEFGHL
ABEKH
BEFGHIK
ABDGH
ABDG
BDFG
CEF
ACEFH
ABEG

ABEH	3
BEGH	3

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1 BEGH 2 ABCEGH 3 ABCEFH 4 BCDEFGHL 5 ABEKH
3 ABCEFH 4 BCDEFGHL
4 BCDEFGHL
5 ABEKH
6 BEFGHIK
7 ABDGH
8 ABDG
9 BDFG
10 CEF
11 ACEFH
12 ABEG

ABEH	3
BEGH	4

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations

Item Sets Definitions and Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining Summary

Feature Spaces Clustering - Basics Classification - Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning References

Database

1	BEGH
2	ABCEGH
3	ABCEFH
4	BCDEFGHL
5	ABEKH
6	BEFGHIK
7	ABDGH
8	ABDG
9	BDFG
10	CEF
11	ACEFH
12	ABEG
	2 3 4 5 6 7 8 9 10

ABEH	3
BEGH	4

C4 and S4

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency

The Apriori Algorithm

Efficiency
Association Rule
Mining

Mining Summary Feature Spaces

Clustering – Basics
Classification – Basics
Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

Database

1	BEGH	
2	ABCEGH	
3	ABCEFH	
4	BCDEFGHL	
5	ABEKH	
6	BEFGHIK	
7	ABDGH	
8	ABDG	
9	BDFG	
10	CEF	
11	ACEFH	
12	ABEG	

S4

ABEH	3
BEGH	4

Only one frequent 4-Itemset remaining.



Maximal Frequent Itemsets and **Closed Frequent Itemsets**

DM868 DM870 DS804

Introduction

Arthur Zimek Freg. Pattern Mining

Sets and Relations

Item Sets Definitions and **S1**

> В 10

D

G 8

Н 8

5

4

9

6

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Minina

Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

S2

AB ΑE 5 AG

ΑН 5 BD

RF BF BG 8

7 BH CE CF

CH DG

EF 5 FG

EH FH

GH

BEH

CFF 4 CFH 4

EFH 4 **EGH** 4

S3

ABF 4 ABG 4 ABH 4 **AFH** 4 BDG 4 **BFG** 5 6 **BGH** 5

S4

BEGH 4

MFI and CFI CFI only



Maximal Frequent Itemsets and **Closed Frequent Itemsets**

DM868 DM870 DS804

Introduction

Arthur Zimek Freg. Pattern Mining

Sets and Relations

Item Sets Definitions and

S1

5

4

9

6

В 10

D

G 8

Н 8

Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Minina Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

S2

AB ΑE 5 AG

ΑН 5 BD

RF BF BG 8

7 BH CE

CF CH DG

EF 5 FG

EH

FH GH

S3

ABF 4 ABG 4 ABH 4 **AFH** 4 BDG 4 **BFG** 5 **BEH** 6

BGH 5 CFF 4 CFH 4

EFH 4 **EGH** 4 **S4**

BEGH 4

MFI and CFI CFI only



Maximal Frequent Itemsets and **Closed Frequent Itemsets**

DM868 DM870 DS804

Arthur Zimek Freg. Pattern Mining

Sets and Relations

Introduction

Item Sets Definitions and Problem-Statements **S1** Monotonicity-Property of Frequency B The Apriori Algorithm Example Efficiency

Association Rule Minina Summary Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning References

ΑE AG ΑН BD RF BF BG вн CE CF CH DG EF FG

10

5

9

6

D 4

G 8

Н 8

S2 AB 5 5 7 5 EΗ FH GH

S3**ABE** 4 ABG 4 ABH 4 **AFH** 4 4 **S4 BFG** 5 **BEGH** 4 **BEH** 6 **BGH** 5 CFF 4 **CFH** 4 **EFH** 4 MFI and CFI **EGH** CFI only



S2

DM868 DM870 DS804

Arthur Zimek
Introduction
Freg. Pattern Mining

Sets and Relations

Classification - Basics

Bayesian Learning

Entropy, Purity, and

Ensemble Learning

(Non-)Linear Sep.

Learning with

References

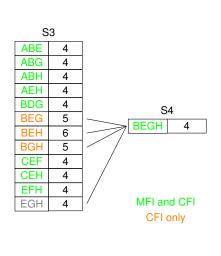
Item Sets Definitions and Problem-Statements S1Monotonicity-Property of Frequency B 10 The Apriori Algorithm Example 5 Efficiency Association Rule D 4 Mining 9 Summary Feature Spaces 6 Clustering - Basics

G | 8

8

AB ΑE 5 AG ΑН 5 BD RF BF BG вн CE CF CH DG EF 5 FG EΗ

> FH GH





DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining Summary

Feature Spaces

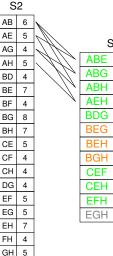
Clustering – Basics Classification – Basics

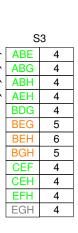
Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

S1
A 7
B 10
C 5
D 4
E 9
F 6
G 8
H 8









DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and

Problem-Statements
Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary

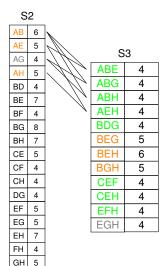
Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

S1
A 7
B 10
C 5
D 4
E 9
F 6
G 8
H 8







DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary

Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References



S1

5

9

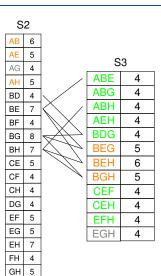
6

B 10

D | 4

G | 8

H | 8







DM868 DM870 DS804

Introduction

Arthur Zimek Freg. Pattern Mining

Clustering - Basics

Bayesian Learning

Entropy, Purity, and

Ensemble Learning

(Non-)Linear Sep.

Learning with Dietributione

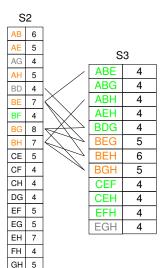
References

Classification - Basics

Sets and Relations Item Sets Definitions and Problem-Statements **S1** Monotonicity-Property of Frequency B 10 The Apriori Algorithm Example 5 Efficiency Association Rule D 4 Mining 9 Summary Feature Spaces 6

> G 8

> Н 8



S4 BEGH 4



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and

Problem-Statements
Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining Summary

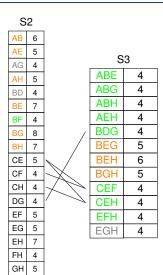
Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Distributions
Entropy, Purity, and
(Non-)Linear Sep.
Ensemble Learning

References

S1
A 7
B 10
C 5
D 4
E 9
F 6
G 8
H 8



S4
BEGH 4



DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining Summary

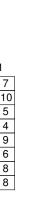
Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with Dietributione

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References



5

9

6

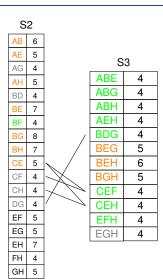
S1

B

D 4

G 8

Н 8







DM868 DM870 DS804

Arthur Zimek

Introduction Freq. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

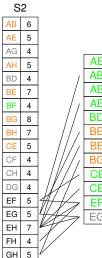
Mining Summary

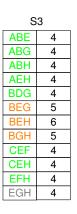
Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References





S4 BEGH 4



DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency
Association Rule
Mining

Summary Feature Spaces

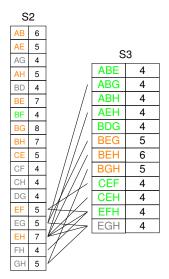
Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

S1
A 7
B 10
C 5
D 4
E 9
F 6
G 8
H 8



S4
BEGH 4

CFI only





Arthur Zimek

Artiful Zillio

Introduction
Freq. Pattern Mining

Sets and Relations Item Sets

Definitions and

Problem-Statements

Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary

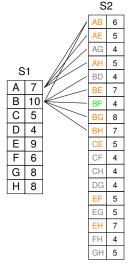
Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References



S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH 4





Arthur Zimek

Artiful Zillie

Introduction Freq. Pattern Mining

Sets and Relations

Definitions and

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining Summary

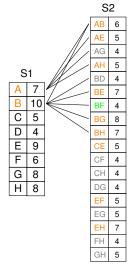
Feature Spaces

Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References



S3

ABE	4
ABG	4
ABH	4
AEH	4
BDG	4
BEG	5
BEH	6
BGH	5
CEF	4
CEH	4
EFH	4
EGH	4

S4

BEGH 4





Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of

Frequency

The Apriori Algorithm Example

Efficiency Association Rule Mining

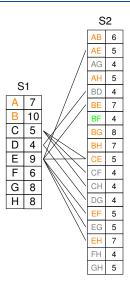
Summary Feature Spaces

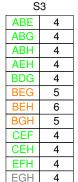
Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

References











Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and

Problem-Statements
Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule Mining

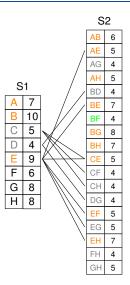
Summary Feature Spaces

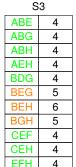
Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References





EGH

4



MFI and CFI CFI only

S4

4





Arthur Zimek

Introduction Freq. Pattern Mining

Sets and Relations Item Sets Definitions and

Problem-Statements
Monotonicity-

Monotonicity-Property of Frequency

The Apriori Algorithm

The Apriori Algo Example

Efficiency
Association Rule
Mining

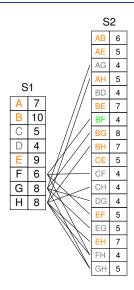
Summary Feature Spaces

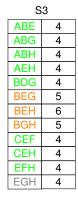
Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Distributions
Entropy, Purity, and
(Non-)Linear Sep.
Ensemble Learning

References











Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations

Sets and Relation Item Sets

Problem-Statements Monotonicity-

Property of Frequency

The Apriori Algorithm

Example Efficiency

Association Rule Mining

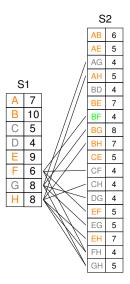
Summary Feature Spaces

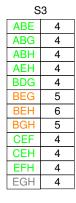
Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References









DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and

Problem-Statements

MonotonicityProperty of

Frequency
The Apriori Algorithm

Example

Efficiency Association Rule

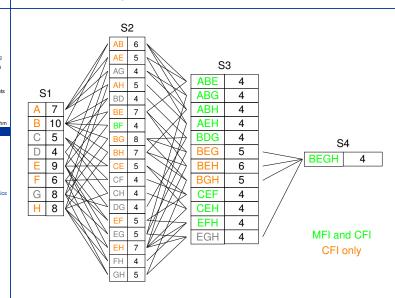
Mining Summary

Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

References





Outline

DM868 DM870 DS804

Arthur Zimek

Introduction

Freq. Pattern Mining Sets and Belations

Sets and Relation

Definitions and

Problem-Stateme Monotonicity-

Property of Frequency

The Apriori Algorithm Example

Citatan

Efficiency Association Bule

Mining

Summary Feature Spaces

Clustering - Basics

Classification – Basics Bayesian Learning Learning with

Distributions
Entropy, Purity, and
(Non-)Linear Sep.

(Non-)Linear Sep. Ensemble Learning

110101011000

Introduction

Frequent Pattern Mining

Sets and Relations

Item Sets

Definitions and Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering – Basics and k-means

Classification - Basics and a Basic Classifie

Basic Probability Theory, Bayes' Rule, and Bayesian Learning

Distributions and Learning with Distributions

Entropy, Purity, and Separation: Linear vs. Non-Linear Separation

Ensemble Learning



In Theory...

DM868 DM870 DS804

Arthur Zimek

- Freq. Pattern Mining
 Sets and Relations
 Item Sets
 Definitions and
- Definitions and Problem-Statements Monotonicity-Property of Frequency

The Apriori Algorithm Example Efficiency

Association Rule Mining Summary

Feature Spaces Clustering – Basics

Classification – Basics Bayesian Learning Learning with Distributions Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning References ▶ naïve algorithm: count frequency of all k-itemsets, for all $\binom{I}{k}$ k-itemsets, for all k

- ▶ number of possible itemsets $0 \le k \le |I|$?
- ▶ Apriori: one database scan for all frequent *k*-itemset *candidates* of a given *k*
- ▶ reduction of number of candidates by the anti-monotonicity principle of frequency: generate only candidates that have a chance to be frequent (join of frequent (k - 1)-itemsets and pruning)



Experiments

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations

Item Sets Definitions and

Problem-Statements Monotonicity-

Property of Frequency

Example

Efficiency

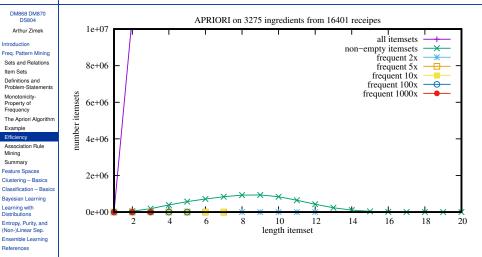
Association Rule Mining

Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

References





Experiments

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining Sets and Relations

Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency The Apriori Algorithm

Example

Efficiency

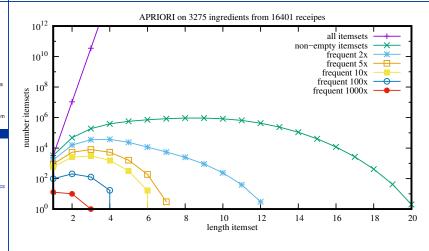
Association Rule Mining Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning







Experiments

DM868 DM870 DS804

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations Item Sets Definitions and

Problem-Statements Monotonicity-Property of

Frequency The Apriori Algorithm Example

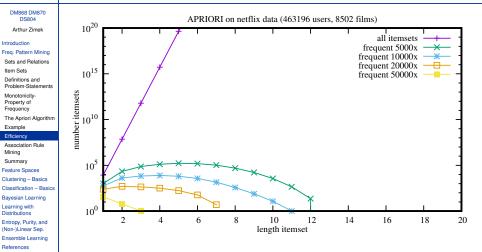
Efficiency

Association Rule Mining Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

References





Outline

DM868 DM870 DSR04

Arthur Zimek

Introduction Freg. Pattern Mining

Sets and Relations

Item Sets

Definitions and

Problem-Statements

Monotonicity-Property of

Frequency

The Apriori Algorithm

Example

Efficiency

Association Rule Mining

Summary

Feature Spaces

Clustering - Basics Classification - Basics

Bayesian Learning Learning with

Dietributione Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Frequent Pattern Mining

Association Rule Mining



Definition: Association Rule

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining Sets and Relations Item Sets Definitions and

Problem-Statements

MonotonicityProperty of
Frequency

The Apriori Algorithm

Example Efficiency

Association Rule

Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

association rule: expresses an implication of the form $X \Rightarrow Y$, where X and Y are itemsets, $X \cap Y = \emptyset$

implication: describes a co-occurrence, not a causality

An association rule does not necessarily need to hold in all cases. We can describe its strength (or weakness), based on the observed cases:

support: The support of an association rule in \mathcal{D} is the support of the union of its components:

$$s(X \Rightarrow Y) = s(X \cup Y)$$

frequency: Analogously, $f(X \Rightarrow Y) = f(X \cup Y)$

confidence:
$$conf(X \Rightarrow Y) = \frac{s(X \cup Y)}{s(X)}$$



Problem 2: Association Rule Mining

DM868 DM870 DS804 Arthur 7imek

Introduction

Freg. Pattern Mining Sets and Relations

Item Sets Definitions and

Problem-Statements

Monotonicity-Property of Frequency

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary

Feature Spaces Clustering - Basics

Classification - Basics Bayesian Learning Learning with

Dietributione Entropy, Purity, and

(Non-)Linear Sep. Ensemble Learning

References

Given:

- a set of items I
- ▶ a transaction database D over I
- ightharpoonup a support threshold σ and a confidence threshold c

Find all association rules $X \Rightarrow Y$ in \mathcal{D} with a support of at least σ and a confidence of at least c, i.e.:

$$\{X \Rightarrow Y | s(X \Rightarrow Y) \ge \sigma \wedge \operatorname{conf}(X \Rightarrow Y) \ge c\}$$

T1: {bread, butter, milk, sugar}

T2: {butter, flour, milk, sugar}

T3: {butter, eggs, milk, salt}

Т4: {eqqs}

T5: {butter, flour, milk, salt, sugar}



Find Association Rules

DM868 DM870 DS804

Arthur Zimek
Introduction
Freg. Pattern Mining

Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-

Property of Frequency 1996].

for frequent itemset *X*:

▶ for each $Y \subset X$, $Y \neq \emptyset$, build the rule $Y \Rightarrow (X \setminus Y)$

This is part of the Apriori algorithm [Srikant and Agrawal,

- delete rules with confidence below a given threshold c

The Apriori Algorithm Example

Efficiency Association Rule

Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning Note that:

For all involved itemsets $(X, Y, (X \setminus Y))$, we have the support from the solution of Problem 1 (stored or reconstructable from closed frequent itemsets). Thus we don't need a single database scan here.



Monotonicity of Confidence

DM868 DM870 DS804

Mining Given:

▶ itemset X

Theorem 2.1

 $ightharpoonup Y \subset X, Y \neq \emptyset$

If $conf(Y \Rightarrow (X \setminus Y)) < c$, then $\forall Y' \subset Y$:

$$conf(Y' \Rightarrow (X \setminus Y')) < c.$$

This property allows the construction of all association rules satisfying some confidence threshold from all frequent itemsets with a procedure similar to the Apriori construction of frequent itemsets, but without database scan. [Srikant and Agrawal, 1996]

Arthur Zimek

Freq. Pattern Mining Sets and Relations Item Sets

Definitions and Problem-Statements Monotonicity-

Property of Frequency The Apriori Algorithm

Example

Efficiency

Association Rule

Mining

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with Distributions

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References



Pruning of Association Rules

DM868 DM870 DS804

Arthur Zimek

Introduction
Freq. Pattern Mining
Sets and Relations
Item Sets

Item Sets
Definitions and
Problem-Statements

Monotonicity-Property of Frequency
The Apriori Algorithm

Example

Efficiency

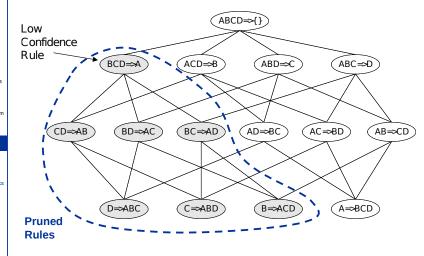
Association Rule Mining

Summary

Feature Spaces Clustering – Basics Classification – Basics

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep.

Ensemble Learning



Adapted from Tan et al. [2006], Fig. 6.15.



Example: Association Rules

DM868 DM870 DS804

Arthur Zimek

Association rules for BEGH (see slide 88), confidence > 60%, support 4:

Introduction Freg. Pattern Mining Sets and Relations Item Sets Definitions and Problem-Statements

Monotonicity-Property of Frequency The Apriori Algorithm

Example Efficiency Association Rule

Mining Summary

Feature Spaces Clustering - Basics Classification - Basics Bayesian Learning

Learning with Dietributione Entropy, Purity, and ng

Non-)Linear Sep.
Insemble Learnin
References

	S(Antecedent)	Confidence		Rule
Ø	4		1.000	
Н	5	4/5 =	0.800	$BEG \Rightarrow H$
GH	7	4/7 ≈	0.571	-
EH	8	4/8 =	0.500	-
BH	5	4/5 =	0.800	$EG \Rightarrow BH$
G	6	4/6 ≈	0.667	$BEH \Rightarrow G$
EG	7	4/7 ≈	0.571	-
BG	7	4/7 ≈	0.571	-
E	5	4/5 =	0.800	$BGH \Rightarrow E$
BE	5	4/5 =	0.800	$GH \Rightarrow BE$
В	4	4/4 =	1.000	$EGH \Rightarrow B$
	H GH EH BH G EG BG E	H 5 GH 7 EH 8 BH 5 G 6 EG 7 BG 7 EB 5 BE 5	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$

$BEG \Rightarrow H$
-
-
$EG \Rightarrow BH$
$BEH \Rightarrow G$
-
-



Interpretation of Support and Confidence

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperty of
Frequency
The Apriori Algorithm
Example

Efficiency
Association Rule
Mining

Summary Feature Spaces

Clustering – Basics Classification – Basics Bayesian Learning

Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

support: measures the frequency of the item set

- rules with very low support may occur simply by chance
- rules with low support are uninteresting from a business perspective

confidence: measures the reliability of the rule

- ▶ $X \Rightarrow Y$ the higher the confidence, the more likely Y is present in transactions that contain X
- estimate of the conditional probability of Y given X



Limitations of Support and Confidence as Measures

DM86	8 DM	870			
DS804					
	_				

Arthur Zimek

Introduction
Freq. Pattern Mining

Sets and Relations Item Sets

Definitions and Problem-Statements

Monotonicity-

Property of Frequency

The Apriori Algorithm

Example

Efficiency Association Rule

Mining

Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning Learning with

Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

References

	coffee	no coffee	
tea	150	50	200
no tea	650	150	800
	800	200	1000

 $\{tea\} \Rightarrow \{coffee\}$

support?

confidence?

 $\{\} \Rightarrow \{\textit{coffee}\}$

support?

confidence?

Conclusion?

(Discussed by Tan et al. [2006], page 372f., example 6.3.)



Interestingness

DM868 DM870 DS804

Arthur Zimek
Introduction
Freg. Pattern Mining

Sets and Relations Item Sets Definitions and Problem-Statements Other measures to assess the interestingness of a rule include:

Lift:
$$Lift(A \Rightarrow B) = \frac{\operatorname{conf}(A \Rightarrow B)}{f(B)}$$

Jaccard:
$$Jaccard(A \Rightarrow B) = \frac{s(A \cup B)}{s(A) + s(B) - s(A \cup B)}$$

conviction:
$$conviction(A \Rightarrow B) = \frac{1 - f(B)}{1 - conf(A \Rightarrow B)}$$

The Apriori Algorithm
Example
Efficiency
Association Rule

Monotonicity-Property of Frequency

Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Bayesian Learning
Learning with
Distributions

Entropy, Purity, and (Non-)Linear Sep.
Ensemble Learning

Recommended Reading:

Advanced reading:

Vreeken and Tatti [2014]

Example of Lift:

- Lift:
 - Lift(I \rightarrow Z) = Conf(I \rightarrow Z) / (Sup(Z) / N)
- Example:
 - Lift(q \rightarrow p) = Conf(q \rightarrow p) / (Sup(p) / N) = = 1 / (25/30) = **1.2**
 - Lift(q \rightarrow r) = Conf(q \rightarrow r) / (Sup(r) / N) = = 1 / (5/30) = **6**

р	q	r	•••
1	1	1	
1	1	1	
1	1	1	
1	1	1	
1	1	1	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
1	0	0	
0	0	0	
0	0	0	
0	0	0	
0	0	0	
0	0	0	

Frequent Itemsets and Association Rules in R

- There are useful packages to perform frequent itemset and association rule mining in R
- These packages contain a variety of off-the-shelf algorithms and sophisticated analysis tools, including visualization-based tools
- One such package is the arules package
- You can learn about this package, e.g., at:
 - CRAN Package Documentation
 - Arules Tutorial by Michael Hahsler

Frequent Itemsets and Association Rules in R

Optional Exercise: Consider the following data set with 10 transactions, represented as a binary matrix where each column (except for the 1st, which contains the transaction ID) stands for an item and the presence or absence of that item in each transaction is denoted by a value 1 or 0, respectively.

ID	Milk	Coffee	Tea	Bread	Butter	Rice	Beans
1	0	1	0	1	1	0	0
2	1	0	1	1	1	0	0
3	0	1	0	1	1	0	0
4	1	1	0	1	1	0	0
5	0	0	1	0	0	0	0
6	0	0	0	0	1	0	0
7	0	0	0	1	0	0	0
8	0	0	0	0	0	0	1
9	0	0	0	0	0	1	1
10	0	0	0	0	0	1	0

Frequent Itemsets and Association Rules in R

Optional Exercise:

This data set is available in a CSV file 10_Groceries_Transactions.csv. This file has a header with the variable names in the 1st line, the other 10 lines represent the transactions. Each line contains 8 values (columns) separated by comma (","). In this exercise you are asked to:

- 1. Read the file into a data.frame called TR_10_Frame using read.table();
- Remove the 1st column (Transaction IDs) and represent the remaining 7 columns as a binary matrix called TR_Matrix;
- Convert this matrix into an object of class transactions using function as(), and name this object TR_obj;
- 4. Inspect and visualise this object using functions summary(), inspect() and image();
- 5. Use functions apriori() and inspect() to generate the rules with minimum relative support of 3/10 and minimum confidence of 9/10.



Outlook: Other Algorithms

DM868 DM870 DS804

Arthur Zimek

Freq. Pattern Mining
Sets and Relations
Item Sets
Definitions and
Problem-Statements
MonotonicityProperly of
Frequency
The Apriori Algorithm
Example
Efficiency

Mining Summary

Feature Spaces
Clustering – Basics
Classification – Basics

Association Rule

Bayesian Learning Learning with Distributions Entropy, Purity, and (Non-)Linear Sep. Ensemble Learning

- Frequent pattern mining is a large research field, for an overview see the collection of topics edited by Aggarwal and Han [2014].
- Another important algorithm for frequent pattern mining is FP-Growth [Han et al., 2000].
- ► Frequent patterns have been defined in other application scenarios such as, e.g., graphs, spatiotemporal data, sequential data (for example protein sequences, as in the work of Birzele and Kramer [2006]).
- ► The principle of anti-monotonicity for pruning has been applied in many other application areas (e.g., in subspace clustering [Zimek et al., 2014]).

References

- M. J. Zaki and W. Meira Jr., "Data Mining and Machine Learning. Fundamental Concepts and Algorithms". Cambridge University Press, 2nd edition, 2020
- P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining", Addison-Wesley, 2006