Constraint-based mining: a major step towards inductive databases

Jean-François Boulicaut
INSA Lyon
LIRIS CNRS FRE 2672
France

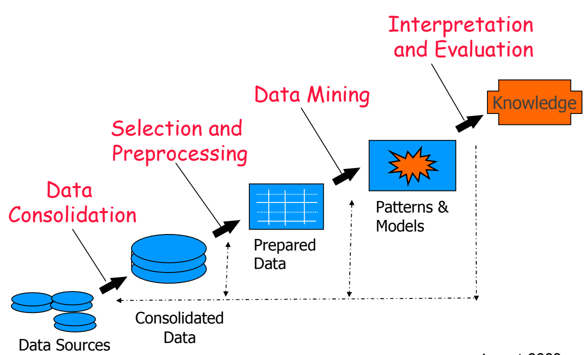
IDA'03 - Berlin (Germany) - August 26, 2003.

Let us motivate the topic

Supporting the iterative and interactive knowledge discovery processes

A database perspective on knowledge discovery

Supporting KDD processes



A prototypical example (1)

- Itemsets in transactional data

	A_1	A_2	A_3		
t1	1	0	0	$\{A_1\}$	baskets - products
t2	1	1	1	$\{A_1, A_2, A_3\}$	documents - keywords
t3	1	0	1	$\{A_1, A_3\}$	sessions - urls
t4	0	1	1	$\{A_2, A_3\}$	cells - genes

A_2 A_3 [2/4, closed, …] A_1 A_2 [1/4, not closed, …]

A prototypical example (2)

- Descriptive rules

A_1	A_2	A_3
1	0	0
1	1	1
1	0	1
0	1	1

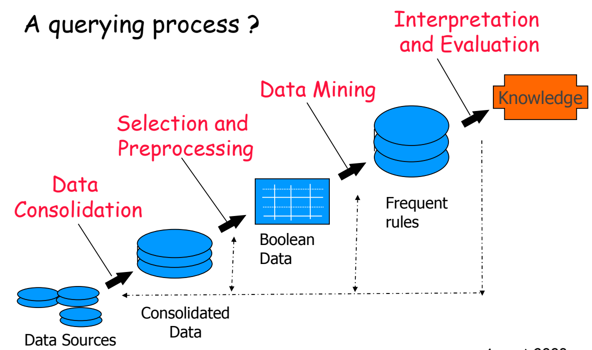
E.g., association rules

Agrawal & al. 93 (sigmod)

$A_1 \Rightarrow A_2$ [1/4, 1/3, ...] $A_1 A_2 \Rightarrow A_3$ [1/4, 1, ...]

KDD processes based on association rules

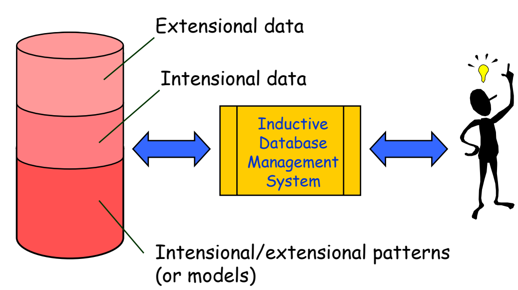
A querying process ?



A prototypical example (3)

- Building the transactional context
  - ✓ Data selection - data preprocessing
- Mining the context
  - ✓ Mining the frequent and « valid » rules
  - ✓ Problems w.r.t. complexity
- Post-processing the itemsets and/or rules
  - ✓ Problems w.r.t. input/output operations
- Using the itemsets and/or rules
  - ✓ Multiple uses of frequent itemsets

The Inductive Database framework



Imielinski & Mannila 96 (cacm)   Boulicaut & al. 99 (dawak)

Objectives of the tutorial

State of the art on the inductive database approach and thus constraint-based data mining

Survey on the cInQ IST-2000-26469 results

An update version of ECML-PKDD '02 tutorial with Luc De Raedt

Project funded by the Future and Emerging Technologies arm of the IST Programme

Overview

1. Introduction to inductive databases
2. Discussing a few query language proposals
3. Query evaluation challenges
   - Constraint-based data mining
   - Optimizing (sequences of) queries
4. Perspectives

What about the bibliography ?

1. Introduction to inductive databases

◆ A vision

«There is no such thing as real discovery, just a matter of the expressive power of the query languages»

Imielinski & Mannila, CACM Nov. 1996

- ✓Make first class citizens out of patterns or models
- ✓ Interesting results for local pattern discovery
- ✓Ongoing research on global pattern discovery (e.g., predictive tasks)

Inductive queries

Tell me something interesting about my data

Give all fragments of molecules that appear in at least 20% of the actives, and in at most 1% of the inactives, and that do not contain a benzene ring.

Give all the maximal sets of genes that are co-regulated in a set of at least 6 tumoral cells and contain exactly one EST.

Give a decision tree that tests upon at most 5 attributes including blood pressure and sex, and that has accuracy at least 90 % on the training data

A long-term DB perspective on data mining

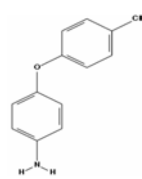
- Why is the relational model so succesful?
  - ✓ A general purpose query language with « nice » properties
    - – simple theoretical foundations and declarative semantics
    - – closure principle
  - The same is needed for KDD applications
  - *The ultimate goal of this approach is to find the equivalent of Codd's relational database model for data mining*

Inspiring examples

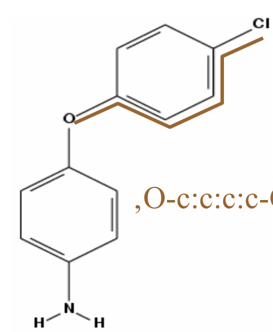
- Molecular fragments
  - ✓ A domain specific IDB
  - Kramer & al. 01 (kdd), De Raedt & Kramer 01 (ijcai)
- Association rules and itemsets
  - ✓ Extremely popular data mining technique for which several "inductive" extensions of SQL have been proposed
- But also, strings, sequential patterns, inclusion and functional dependencies, …, and recently equations, clusters, classifiers ...

Molecular Feature Mining: Molfea

- What ?
  - ✓ Find fragments (substructures) of interest in sets of molecules
- Why ?
  - ✓ Discover new knowledge
  - ✓ Use in predictive models
    - SAR (Structure Activity Relationship)
    - De Raedt & Kramer 01 (ijcai)



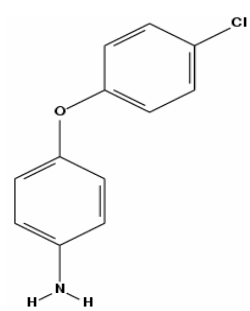
Molecules and fragments



,O-c:c:c:c-Cl'

- 2D-structure
- Fragments
  - ✓ Substructres
  - ✓ Linear fragments
  - ✓ Sequence of atoms and bonds
- Linear fragments
  - ✓ ,o', ,c', ,cl', ,n' ... elements
  - ✓ ,-' ... single bond
    ,=, ... double bond
    ,#' ... triple bond
    ,:' ... aromatic bond
  - ✓ (hydrogens implicit)
- Smarts encoding

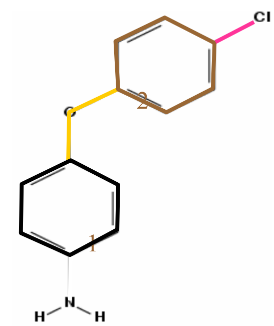
Smiles encoding



- Smiles
  - ✓ Compact encoding of molecular structure
  - ✓ Used by computational chemists
  - ✓ Supported by many tools (e.g., Daylight)
  - ✓ Very compact !
  - ✓ Very efficient matching

$N-c1:c:c:c(-O-c2:c:c:c(-Cl):c:c2):c:c1$

Smiles encoding



$N-$

$N-c1:c:c:c:c:c1$

$N-c1:c:c:c(-O-):c:c1$

$N-c1:c:c:c(-O-$
$c2:c:c:c:c:c2):c:c1$

$N-c1:c:c:c(-O-c2:c:c:c(-Cl):c:c2):c:c1$ 3

Constraint-based mining (1)

- What ?
  - ✓ Use constraints to specify which fragments are interesting
    - The scientist/user "controls" the mining process
  - ✓ Evaluation functions (e.g., generality, frequency)
  - ✓ Primitive constraints (e.g., minimal/maximal frequency)
  - ✓ Queries (e.g., conjunctions of primitive constraints)

Generality relation

- Generality
  - ✓ One fragment is more general than another one if it is a substructure of the other one
  - ✓ Notation : $g \leq s$ *(g is more general than s, i.e., g will match a graph/string whenever s does)*
  - ✓ Graphs : ~ subgraph relationship
  - ✓ Strings : substring / subsequence relationship
    - *aabbcc* $\leq$ *ddaabbccee (substring)*
    - *abc* $\leq$ *aabbcc (subsequence)*
  - ✓ Itemsets : subset relation

Primitives

- MolFea Specific !
  - ✓ g is equivalent to s (*syntactic variants*) only when they are a reversal of one another

  E.g. ‚*C-O-S'* and ‚*S-O-C'* denote the same substructure

  - ✓ g is more general than s iff g is a subsequence of s or g is a subsequence of the reversal of s

  E.g. ‚*Cl-O-S'* $\leq$ ‚*Cl-O-S-c:c:c'*, ‚*O-Cl'* $\leq$ ‚*Cl-O-S'*

- Frequency of a fragment f on a data set D
  - ✓ Percentage of data points in D that f occurs in

Primitive constraints

- $\phi \leq P$, $P \leq \phi$, *not* $(\phi \leq P)$ and *not* $(P \leq \phi)$

  ϕ ... unknown target fragment
  P ... a specific fragment

- $Freq(\phi, D1) \geq t$ — *minimal frequency*

  $Freq(\phi, D2) \leq t$ — *maximal frequency*

  t ... positive real number between 0 and 1
  D1, D2 ... data sets

  E.g. $Freq(\phi, Pos) \geq 0.20$

Examples of Molfea queries

- Assume queries are conjunctions of primitive constraints

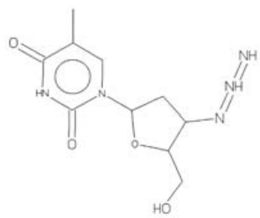
$(\text{`N-O'} \leq \phi)$
$\wedge\ (Freq(\phi, Act) \geq 0.1)$
$\wedge\ (Freq(\phi, Inact) \leq 0.01)$

$not(\text{‚F'} \leq \phi) \wedge not\ (\text{‚Cl'} \leq \phi)$
$\wedge\ not\ (\text{‚Br'} \leq \phi) \wedge not\ (\text{‚I'} \leq \phi)$
$\wedge\ (Freq(\phi, Act) \geq 0.05)$
$\wedge\ (Freq(\phi, Inact) \leq 0.02)$

The HIV data set De Raedt & al 01 (sigkdd)

- Developmental Therapeutics Program's AIDS Antiviral Screen Database (http://dtp.nci.nhi.gov)
  - One of the largest public domain databases of this type
- Measures protection of human CEM cells from HIV-1 infection using a soluble formazan assay
- 41768 compounds have been selected among the 43382 ones
  - 40282 Confirmed Inactive
  - 1069 Confirmed Moderately Active
  - 417 Confirmed Active

AZT (Azidothymidine)



$N=N=N-C-C-C-n:c:c:c=O$

$N=N=N-C-C-C-n:c:n:c=O$

The majority of these fragments are derivatives of AZT.

Gives insight into the structural requirements for anti-HIV activity.

A rediscovery that proves the principle

Back to itemsets

Interesting « new » evaluation functions and primitive constraints

… thanks to Galois connection

Evaluation functions for itemsets (1)

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0

f (T, r) set of attributes shared by transactions in T

g (S, r) set of transactions that contain each attribute in S

f ({1,2}) = {A,C}

g ({A,B}) = {1,4,6}

Freq (S,r) is the size of g(S,r)

Evaluation functions for itemsets (2)

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0

h (S, r) = f(g(S,r),r) closure for a set of attributes

h' (T, r) = g(f(T,r),r) closure for a set of transactions

h ({A,B}) = f ({1,4,6}) = {A,B,C}
h ({A,B,C}) = f ({1,4,6}) = {A,B,C}

h' ({1,2}) = g ({A,C}) = {1,2,3,4,6}
h' ({1,4,6}) = g ({A,B,C}) = {1,4,6}

Primitive constraints based on closures

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0

$C_{close}(S,r)$	$C_{free}(S,r)$
$h(S,r) = S$	closed set generators
{A,B,C} is closed	{A,B} is free
{C,D} is not closed	{C,D} is not free

$h(\{A,B\}) = \{A,B,C\} = h(\{A,B,C\})$

Condensed representations w.r.t. frequency queries

Free sets Boulicaut & al. 00 (pkdd)

Assume r

ABCDE
ABCD
ACD
ABE
CD
CE

BD BC DE

ABC ABD BCD ACE BCE ADE BDE CDE

ABCD ABCE ABDE ACDE BCDE

ABCDE

See also key patterns
Bastide & al. 00 (sigkdd explorations)

Using free sets: an example

- Prototype ac-miner Boulicaut et al. 03 (dmkd journal)

X, Y : n — X is a free set of attributes

$X \cup Y = h(X,r) = h(X \cup Y,r)$

$Freq(X,r) = Freq(X \cup Y,r) = n$

A B , C : 3 tells that {A,B,C} is closed and frequent

A subset of the (frequent) association rules with confidence 1

$A\ B \Rightarrow C$ [3/6, 1]

See Becquet & al. 02 (genome biology) for an application

Primitive constraints on itemsets

$C_{minfreq}(S,r)$	$C_{maxfreq}(S,r)$
$A \notin S$	$A \in S$
$\{A,B,C,D\} \supset S$	$\{A,B,C,D\} \subseteq S$
$S \cap \{A,B,C\} = \emptyset$	$S \cap \{A,B,C\} \neq \emptyset$
$sum(S.val) \leq v$	$sum(S.val) > v$

$\nexists A \in 2^n$, Interest(A) > Interest (S)

Primitive constraints based on closures

e.g., $C_{close}(S,r)$, $C_{free}(S,r)$, etc.

Examples of queries

- On itemsets

$S \cap \{A,B,C\} = \emptyset \wedge Freq(S, r) \geq 0.1$

$C_{close}(S,r) \wedge C_{freq}(S,r)$ $(C_{free}(S,r) \wedge C_{freq}(S,r) \wedge h(S,r))$

$C_{close}(S,r) \wedge g(S,r)$

- On association rules

$C_{freq}(X \Rightarrow Y,r) \wedge C_{conf}(X \Rightarrow Y,r) \wedge \text{sum}(X \cup Y.\text{val}) \leq v$

$C_{freq}(X \Rightarrow Y,r) \wedge C_{free}(X,r) \wedge C_{close}(X \cup Y,r)$

See Bastide & al. 00 (cl) and the various contributions to cover computations (non redundant association rule mining by, e.g., Bykowski, Zaki, Phan Luong, Kryskiewicz)

33 August 2003

Back to simple abstractions

- What is an inductive database ?
  - ✓ A set of data sets
  - ✓ A set of pattern sets
- IDB languages
  - ✓ A query language that generates data sets
  - ✓ A query language that generates pattern sets
- Closure principle
  - ✓ The result of a query should be a pattern set, a data set or a combination thereof

34 August 2003

Manipulation

create data set D
create view data set D
create pattern set P
create pattern view P

- Insert / Delete / Update statements
- Data and pattern sets can be extensional or intensional

35 August 2003

Illustration

create data set D1 with q1
create pattern view P1 as q2 (D1)
At this point assume P1 = PSet1
update data set D1 with q2
Update P1 too : P1 = update(PSet1)

- Incremental data mining !

insert P2 into pattern view P1

- Pattern view update problem

(D,P)

$\Downarrow$

(D',P')

36 August 2003

Pattern domains

- Pattern domains
  - ✓ Language (e.g., itemsets, sequences, graphs, dependencies, decision trees, clusters)
  - ✓ Evaluation functions (e.g., frequency, closures, generality, validity, accuracy)
  - ✓ Primitive constraints (e.g., minimal and maximal frequency, freeness, syntactical constraints, minimal accuracy)
  - ✓ Combining « primitives » leads to inductive queries … yet a linguistic component has to be designed

37 August 2003

Solvers

- Solvers (see Part 3 of this tutorial)
  - ✓ Computing solution spaces for (more or less) primitive constraints
    - Solving some primitive constraints can be extremely hard
- Complex inductive query evaluation
  - ✓ What ?
    - Arbitrary boolean combination of primitive constraints
    - Combination of pattern domains

38 August 2003

2. Discussing a few query language proposals

- ✓ MINE RULE Meo & al. 96 (vldb), 98 (icde,dmkd)
- ✓ MSQL Imielinski & Virmani 96 (kdd), 99 (dmkd)
- ✓ LDL++ Giannotti & Manco 99 (pkdd)
- ✓ RDM De Raedt 00 (ilp)
- ✓ DMQL Han & al. 96 (kdd), Han & Kamber 01 (mk)

A critical evaluation of several proposals

Deliverable D0 cInQ (01)
Botta & al. 2002 (dawak) 2003 (book dbdm)

Comment: query language vs. software librairies

39 August 2003

Supporting association rule mining (1)

- Pre-processing : manipulating data sets
  - ✓ E.g., compute a transactional context
    - Selections of relevant sources, agregations, sampling, discretizations, etc
- Data Mining : generating pattern sets
  - ✓ E.g., compute 5%-frequent association rules
    - A query as some « syntactic sugar » on top of an algorithm

… can we do better?

40 August 2003

Supporting association rule mining (2)

- Post-processing : manipulating pattern sets
  - ✓ E.g., identify interesting rules among the (tens of thousands) of frequent ones
    - – Selections of relevant patterns, redundancy elimination, grouping, etc
  - ✓ Querying materialized collections of patterns
  - ✓ Crossing over the patterns and the data
- What about standard query languages?

41 August 2003

MINE RULE (1)

- A SQL-like operator on transactional DB

Table Purchase

Tid	Customer	Item	Date	Price	Qty
1	c1	ski-pants	12/1	55	1
1	c1	beer	12/1	4	2
2	c2	shirts	12/1	21	1
2	c2	jackets	12/1	115	1
3	c1	diapers	12/1	18	1
…	…	…	…	…	…

42 August 2003

MINE RULE (1)

```
MINE RULE exemple1 as
SELECT DISTINCT   1..n Item as BODY, 1..1 Item as HEAD,
                  SUPPORT, CONFIDENCE
FROM Purchase
GROUP BY Tid
EXTRACTING RULES WITH SUPPORT: 0.01,
                      CONFIDENCE: 0.7
```

E.g., shirt socks jacket ⇒ boots (0.01,0.73)

43 August 2003

MINE RULE (2)

```
MINE RULE exemple2 as
SELECT DISTINCT   1..n Item as BODY, 1..1 Item as HEAD,
                  SUPPORT, CONFIDENCE
WHERE HEAD.Item=« umbrellas »
FROM Purchase
GROUP BY Tid
HAVING COUNT(*)<6
EXTRACTING RULES WITH SUPPORT: 0.001,
                      CONFIDENCE: 0.7
```

E.g., jacket flight_Dublin ⇒ umbrellas (0.01,0.79)

44 August 2003

MINE RULE (3)

MINE RULE exemple3 as
SELECT DISTINCT 1..n Item as BODY, 1..n Item as HEAD, SUPPORT, CONFIDENCE
FROM Purchase
GROUP BY Customer
CLUSTER BY Date
HAVING BODY.Date < HEAD.Date
EXTRACTING RULES WITH SUPPORT: 0.01, CONFIDENCE: 0.9

E.g., ski_pant ⇒ jacket (0.02,0.92)

MINE RULE (4)

MINE RULE WordOfMouth as
SELECT DISTINCT 1..1 Customer as BODY,
1..n Customer as HEAD,
SUPPORT, CONFIDENCE
WHERE BODY.Date <= HEAD.Date
FROM Purchase
GROUP BY Item
EXTRACTING RULES WITH SUPPORT: 0.01,
CONFIDENCE: 0.9

E.g., c7 ⇒ c3 c12 (0.02,0.93)

MINE RULE (4)

++

- ✓ Data selection by means of « full » SQL
- ✓ Query evaluation can be effective because of ad-hoc strategies

--

- ✓ Dedicated to association rules
- ✓ Poor possibilities for expressing background knowledge
- ✓ No specific mechanism for rule post-processing (results are stored in relational tables)

MSQL (1)

- Further integration within SQL

job=research ∧ age = [26,38] ⇒ position=AssProf (0.31,0.95)

Emp(Id, Age, Job, Salary, Position)

GET_RULES (Emp)
INTO Rules
WHERE … and support > 0.1 and confidence > 0.8

SELECT_RULES (Rules)
WHERE body has { (Age=*) (Job=*)}
and head is { (Position=*)}

MSQL (2)

Emp(Id, Age, Job, Salary, Position)

SELECT ∗
FROM Emp
WHERE violates all (GET_RULES (Emp)
WHERE body is {(Age=∗)}
and head is {(Salary=∗)}
and confidence > 0.3)

Connecting patterns to data

49 August 2003

MSQL (3)

GET_RULES (Source) INTO R1
WHERE body has {(Age=∗)}
and head has {(Salary=∗)}
and support > 0.1
and confidence > 0.9
and not exists (GET_RULES (Source) INTO R2
WHERE body has {(Age=∗)}
and head has {(Salary=∗)}
and support > 0.1
and confidence > 0.9
and R2.body has R1.body)

A correlated query for mining rules with minimal body

50 August 2003

MSQL (4)

++

- ✓ Query evaluation can be effective on data and persistently stored rules
- ✓ Useful operators for association rule mining (discretization, crossing over data and patterns)

--

- ✓ Dedicated to (propositional) association rules
- ✓ Limits of the underlying relational framework (e.g., for the definition of background knowledge)

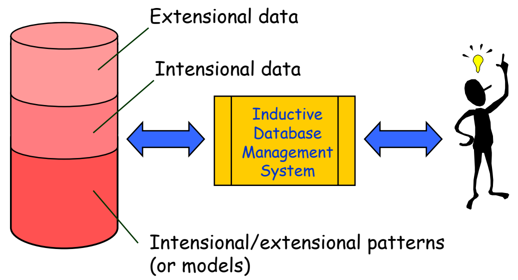
51 August 2003

A « synthesis »

- ◆ DMQL Han & al. 1996 (kdd) Han & Kamber 2001 (m-k)
  - ✓ A typical example of « syntactic sugar » for using many different (efficient) data mining algorithms
- ◆ Research challenges
  - What are the fundamental primitives ?
  - Pre and post-processing are so poorly supported !
    - Querying relational databases that contain itemset or rule collections is not a solution
  - Look for primitives and expressivity for practical data mining problems (at the specification level)
  - Linguistic issues

52 August 2003

Query evaluation in inductive databases



Inductive query evaluation

A crucial need for optimizations

Computing the solutions can be impossible …

… when possible, optimization is crucial to support interactivity and the dynamic aspect of knowledge discovery processes

3. Query evaluation challenges

- Single inductive query evaluation
- How to compute ?

  Th(L⊗E,r,q) = {(ϕ,e)∈ L⊗E | q(r,ϕ) is true}

  – q is an inductive query (say a combination of primitive constraints) on the inductive database r
  – L a language of patterns
  – e is a property of pattern ϕ (e.g. frequency)

  ✓ « Generate and test » is generally impossible

  ✓ « Pushing constraints » can be difficult

Properties of constraints

- Anti-monotonicity of q w.r.t. ≤
  - ✓ q is anti-monotone w.r.t. ≤ if and only if
    - – For all g,s : g ≤s and s satisfies q implies g satisfies q
    - – E.g., the minimal frequency is anti-monotonic w.r.t. generality (strings, itemsets, etc)

    The famous Apriori algorithm Agrawal & al. 94 (vldb) or its generalization: the levelwise algorithm Mannila & Toivonen 97 (dmkd)
  - ✓ Many other constraints are anti-monotonic w.r.t. ≤ (See, e.g., Ng & al. 98 (sigmod))

Anti-monotonic constraints on itemsets

$C_{minfreq}$ (S,r)

$A \notin S$

$\{A,B,C,D\} \supset S$

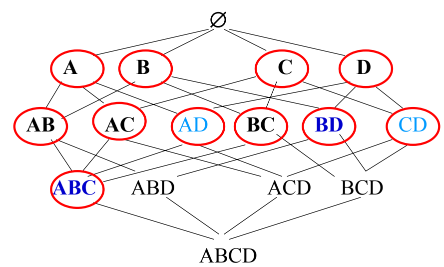
$S \cap \{A,B,C\} = \emptyset$

$sum(S.val) \leq v$

C_{free} (S,r) ...

Application to frequent itemset mining
(Apriori with $Freq(S,r) \geq 2$)

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0



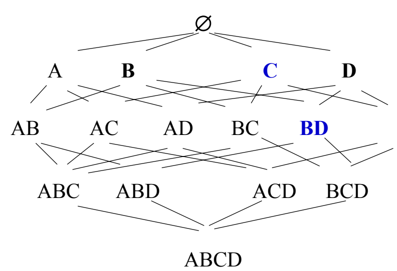
Solution space characterized by {{A,B,C},{B,D}}

Borders of theories Mannila & Toivonen 97 (dmkd)

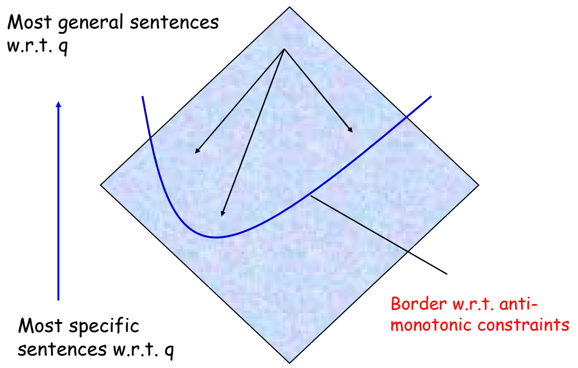
- Positive border
  - ✓ The most specific (interesting) sentences
    - E.g., the maximal frequent sets
  - ✓ In Machine Learning terminology : the S-set of the version space (see papers by Mitchell, Hirsh, Mellish)
- Negative border
  - ✓ The most general sentences that are not interesting
    - E.g., the minimal infrequent sets
- Borders: a tool for complexity analysis

$Freq(S,r) \geq 2 \wedge A \notin S$

- Notice that the conjunction or the disjunction of anti-monotonic constraints is anti-monotonic



Solution space characterized by {{C},{B,D}}



61 August 2003

Use of borders

- Single border can represent the whole theory
- Borders are a *condensed representation* of the solution space
  - ✓ They store only a selection of the relevant solutions
  - ✓ Computing borders or theories ?
    - E.g., feature construction
      vs.
      association rule mining
- Using borders in the « Guess and Correct » approach

62 August 2003

« Guess and Correct » Mannila & Toivonen 97 (dmkd)

```
C := Bd⁺(O)                         Clean the guess O
E := ∅
While C is not empty
    do  E := E ∪ C
        O := O \ {φ ∈ C | q(r,φ) is false}
        C := Bd⁺(O) \ E
    od
C := Bd⁻(O) \ E                     Expand the corrected O
While C is not empty
    do  O := O ∪ {φ ∈ C | q(r,φ) is true}
        C := Bd⁻(O) \ E
    od
Output O                            O = Th(L,r,q)
```

63 August 2003

Representing solutions w.r.t. monotonic constraints

- Many useful constraints are monotonic
  - ✓ E.g., the maximal frequency constraint

    $Freq(\phi, Act) < x$ is monotonic w.r.t. $\leq$

    - If we have a fragment $g \leq s$, then if g is a solution then s is a solution as well
- Monotonic constraints impose a border G on the space of solutions
  - ✓ q is monotonic w.r.t. $\leq$ if and only if not(q) is anti-monotonic w.r.t. $\leq$

64 August 2003

Monotonic constraints on itemsets

$C_{maxfreq}(S,r)$

$A \in S$

$\{A,B,C,D\} \subseteq S$

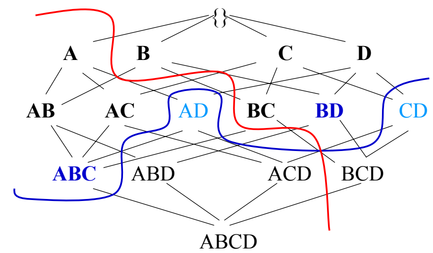
$S \cap \{A,B,C\} \neq \emptyset$

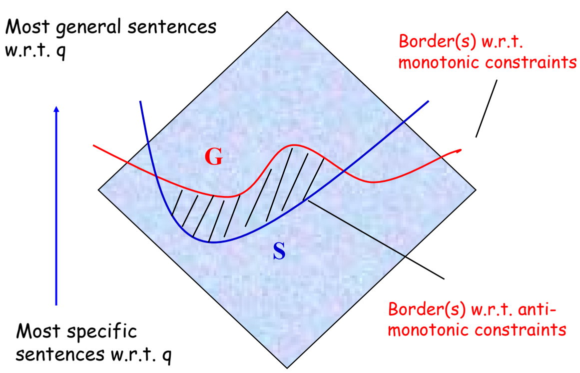
sum(S.val) > v

etc

$Freq(S,r) \geq 2 \wedge (A \in S)$

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0





Mitchell's Version Spaces (1)

- Consider now two constraints :

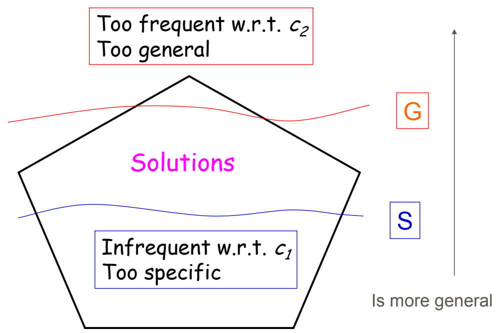
$$c_1 = freq(f, D) \geq x$$
$$c_2 = freq(f, E) \leq y$$

- We want to compute

$$sol(c_1 \wedge c_2) = \{f \mid \exists s \in S, g \in G : g \leq f \leq s\}$$

where S and G are defined w.r.t. $c_1 \wedge c_2$

Mitchell's Version Spaces (2)



Constraint-based mining and VS

Anti-monotonic	Monotonic
$freq(f, D) \geq x$	$freq(f, D) \leq x$
$f \leq P$	$f \geq P$
$not(P \leq f)$	$not(P \geq f)$
In ML	**In ML**
$f \leq P$	$not(f \leq P)$
~	~
P is a positive example	P is a negative example

Computing borders

- Borders S and G characterize the set of solutions for inductive queries that are conjunctions of monotonic and anti-monotonic constraints
- Combination of (well-known) algorithms
  - ✓ Levelwise algorithm
  - ✓ Mitchell's and Mellish's version space algorithms
  - ✓ Max-Miner Bayardo 97 (sigmod), etc.

Generic algorithms for solving conjunctive constraints

- Condensed representation of the solution
  - ✓ Level wise version space algorithm

    De Raedt & al. 01 (ijcai)
- Theory level
  - ✓ A generic levelwise algorithm for pushing conjunctions of anti-monotonic and monotonic constraints Boulicaut & Jeudy 02 (ida)
  - ✓ Dual Miner Gehrke & al. 02 (kdd)

Other trends in constraint-based mining for local patterns (1)

- Pushing constraints into recent efficient frequent pattern mining algorithms, typically FP-Growth
  - ✓ Further studies on constraint properties like succinctness, convertibility, etc
  - ✓ Impressive results by SFU group (Han & al.)
- « Pushing » non anti-monotonic nor monotonic constraints
  - ✓ Regular expression constraints (e.g., Garofalakis & al.)
  - ✓ Optimization constraints (e.g., Morishita & Seke)

Other trends in constraint-based mining for local patterns (2)

- Towards adaptative strategies
  - ✓ Mining frequent sequences under regular expressions: the RE-Hackle framework
    Albert-Lorincz & Boulicaut 03 (sdm)
  - ✓ The preprocessing framework Exante for itemset mining under conjunction of monotonic and anti-monotonic constraints
    Bonchi & al. 03 (pkdd)
  - ✓ ExaMiner
    Bonchi & al. 03 (icdm)

Other trends in constraint-based mining for local patterns (3)

- Studies of ε-adequate representations
  Mannila & Toivonen 96 (kdd)
- Assume the class of queries that returns the frequency of an itemset, look for alternative representations on which we can provide its frequency with a precision of at most ε
  - ✓ E.g., the collection of γ-frequent sets is $\gamma/2$-adequate
    Is it possible to find smaller representations, i.e., condensed representations

Consended representations for frequency queries on itemsets

$Th(L\otimes E,r,q) = \{(\phi,e)\in L\otimes E \mid q(r,\varphi) \text{ is true}\}$

- Problems with borders
  - ✓ For some application, evaluation functions have to be known or approximated
  - ✓ Interesting results since the seminal work on close
    Pasquier et al. 99 (icdt)
  - ✓ Exact and approximate representations have been studied

Frequent closed (or free) sets as a condensed representation

Assume r

r
ABCDE
ABCD
ACD
ABE
CD
CE

<u>BD</u> <u>BC</u> <u>DE</u>

ABC ABD BCD <u>ACE</u> <u>BCE</u> ADE BDE CDE

ABCD ABCE ABDE ACDE BCDE

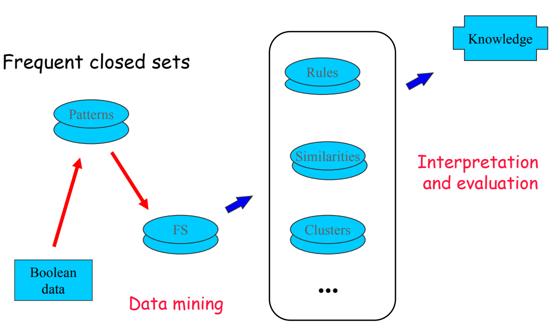
ABCDE

Apriori vs. Close

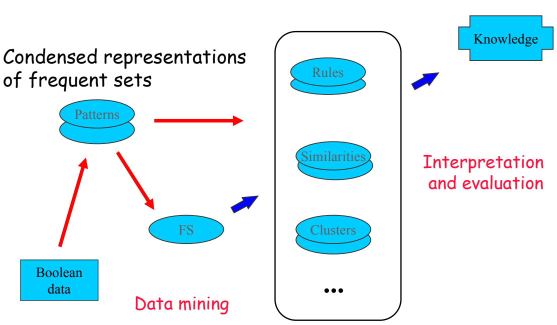
Dataset/ Frequency threshold	Time in sec.	$\|\|FS_\sigma\|\|$	Scans	Time in sec. ($1^{st}/2^{nd}$ step)	$\|\|FC_\sigma\|\|$	Scans
ANPE/σ=0.05	1 463.9	25 781	11	69.2 / 6.2	11 125	9
Census/σ=0.05	7 377.6	90 755	13	61.7 / 25.8	10 513	9
ANPE/σ=0.1	254.5	6 370	10	25.5 / 1.1	2 798	8
Census/σ=0.1	2 316.9	26 307	12	34.6 / 6.0	4 041	9
ANPE/σ=0.2	108.4	1 516	9	11.8 / 0.2	638	7
Census/σ=0.2	565.5	5 771	11	18.0 / 1.1	1 064	9

Back to the concept of closed sets

Frequent closed sets



Condensed representations of frequent sets



Computing (frequent) closed sets

	A	B	C	D
1	1	1	1	1
2	1	0	1	0
3	1	0	1	0
4	1	1	1	1
5	0	1	1	0
6	1	1	1	0

An efficient approach for computing (frequent) closed sets

i) Compute the free sets
ii) Ouput their closures h

Other algorithms exist (e.g., charm, closet)

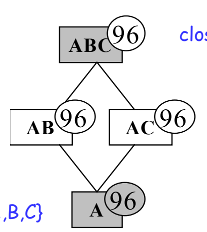
New developments
See closet+ (sigkdd 2003)

Condensed representations of frequent itemsets

- Maximal itemsets e.g., Bayardo 97 (sigmod) Max-Miner
- Version spaces e.g. De Raedt 01 (ijcai)
- Closed sets Pasquier & al. 99 (icdt) - Boulicaut & Bykowski 00 (pakdd) - Han & Pei 00 (wdmkd) - Zaki 00 (sigkdd) - Close - Closet - Charm
- Free sets
- δ-free sets

  Boulicaut & al. 00 (pkdd) 03 (dmkd) - Bastide & al. 00 (sigkdd explorations) Min-Ex - Pascal
- $\vee$-free sets Bykowski & Rigotti 01 (pods) 03 (is) Kryskiewicz 01 (icdm)
- NDI Calders & Goethals 02 (pkdd)

Free and closed sets



closure({ABC})={ABC}

closure({A})={A,B,C}

δ-freeness

- A δ-free-set is such that there is no δ-strong rules that holds between its subsets

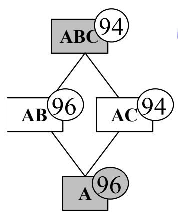
  $X \Rightarrow_\delta Y$ is δ-strong if it has at most δ exceptions

A	B	C	D
1	0	1	0
1	1	1	0
0	1	1	1
0	1	0	1
1	1	1	0

{A,B} was free but is not 1-free

$C_{\delta\text{-Free}}(S)$ checking δ-freeness

An example of a 2-free sets



B,C $\in$ closure$_2$({A})

Examples of condensed representations

1	ABCD
2	AC
3	AC
4	ABCD
5	BC
6	ABC

Threshold 2

16 frequent sets

1 maximal frequent set

Frequent closed sets

C, AC, BC, ABC, ABCD

Frequent free sets

$\emptyset$, A, B, D, AB

Frequent 1-free sets

$\emptyset$, B, D

« Approximation » from closed sets

- ε-adequate representation

If S is not included in a γ-frequent closed set

Then S is not frequent (return Freq(S,r) =0)

Else S is frequent

Let choose the frequent closed set X s.t. S $\subseteq$ X that has the maximal support and return Freq(S,r) = Freq(X,r)

Approximation from δ-free sets

- ε-adequate representation

If S is a superset of an element from FreeBd$^-$

Then S is not frequent (return Freq(S,r) =0)

Else S is frequent

Let choose the frequent δ-free set Y $\subseteq$ X that has the minimal support and Freq(Y,r) - Freq(X,r) $\leq$ |X \ Y|δ

Computing frequent δ-free-sets

- Min-Ex is an effective levelwise algorithm that computes every frequent δ-free set in r
  - ✓ Thanks to freeness anti-monotonicity and an effective freeness test Bykowski 02 (Ph.D)
- Promising experimental validation on dense datasets
  - High condensation and pruning even for low δ
  - Low error in practice even for « large » δ values
- Recent proposal of new approximate condensed representations Han & al. 02 (icdm)

Other trends in constraint-based mining for local patterns (4)

- Boolean inductive queries

  De Raedt 02 (dtdm) - De Raedt & al. 02 (icdm) - De Raedt 03 (sigkdd explorations) - Dan Lee & De Raedt 03 (icdm)
- Query optimization
  - ✓ Single query vs. Sequence of query
  - ✓ Interactive mining and optimization of sequences of queries
    - Containment - Equivalence - Dominance

      Baralis and Psaila 99 (dawak)
- Operations on solution sets (VS, VS trees)

A flavor about the potential for optimization

Claim

Let q_1 and q_2 be two queries that are logically equivalent.

Then $sol(q_1) = sol(q_2)$

Using logical rewrites to optimize the mining process.

E.g. $(a_1 \vee a_2) \wedge (m_1 \vee m_2)$ is logically equivalent to

$(a_1 \wedge m_1) \vee (a_2 \vee m_1) \vee (a_1 \wedge m_2) \vee (a_2 \vee m_2)$

One versionspace versus the disjunction of four

What is best ?

4. Perspectives

- Other forms of primitives?
  - ✓ E.g. accuracy of rule/hypotheses is larger than x
  - ✓ Neither monotonic nor anti-monotonic
- Optimization primitives?
  - ✓ Find n best patterns according to some objective criterion
- Study advanced strategies (including adaptative ones) as the core technology for inductive database management systems

Considering global pattern mining

- Special issue of SIGKDD Explorations on Constraint-based mining, june 2002
- Other forms of tasks?
  - ✓ Clustering, e.g., Cardie et al. 01 (icml)
    - Formulate constraints on number of desired clusters, and cluster membership
  - ✓ Prediction, e.g., Garofalakis & al. 01 (sigkdd explorations)
    - Some approaches to decision tree learning exist
  - ✓ Equation discovery, e.g., Dzeroski & al. 03 (kdid)
    - Discovery of polynomial equations under constraints (heuristic solver)

To get some up-to-date information

- Proceedings of the two first International Workshops on Knowledge Discovery in Inductive Databases
  - ✓ KDID 2002 co-located with ECML-PKDD 2002, Helsinki (August 2002)
  - ✓ KDID 2003 co-located with ECML-PKDD 2003, Catvat-Dubrovnik (September 2003)
    - Invited talk by Minos Garofalakis (Bell Labs)
- http://www.cinq-project.org

Acknowledgments

cInQ consortium

- INSA Lyon
- University of Torino
- Politecnico di Milano
- Albert-Ludwigs University Freiburg
- Nokia Research Center Helsinki (and HIIT-BRU)
- Institute Jozef Stefan