

A Methodological View on Knowledge-Intensive Subgroup Discovery

Martin Atzmueller, Frank Puppe
University of Würzburg, Germany



Subgroup Discovery

2

- **Subgroup Discovery: “discover interesting subgroups of individuals” (in a case base)**
- **Example: Insurance domain - customers with increased car insurance rate (target concept)**
 - "young men owning a sports car" – increased car insurance rate
- **Used for**
 - Exploration
 - Descriptive induction
- **Goal:**
 - Discover not necessarily complete relations
 - Nuggets – partial relations



Subgroup Discovery (cont.)

3

- **“Interesting” subgroups**
 - As large as possible
 - Deviating behavior of designated target variable w.r.t. total population
- **Example - Medical domain: risk groups for target variable coronary heart disease (CHD)**
Family History = + AND Smoker=y => CHD
- **(Conjunctive) description language:**
 $(a_1, V_1) \wedge \dots \wedge (a_n, V_n), V_i \subseteq \text{domain}(A_i)$
Selectors (a_i, V_i)



The Case for Background Knowledge

4

- **Problem: too many (uninteresting) results**
- **Example: Medical domain**
 - Relations with high support already known
 - A lot of background knowledge available



Background knowledge helps SD:

- **Restrict search space**
- **Focus search process**
- **Increase representational expressiveness**
- **Post-processing of sets of subgroups**

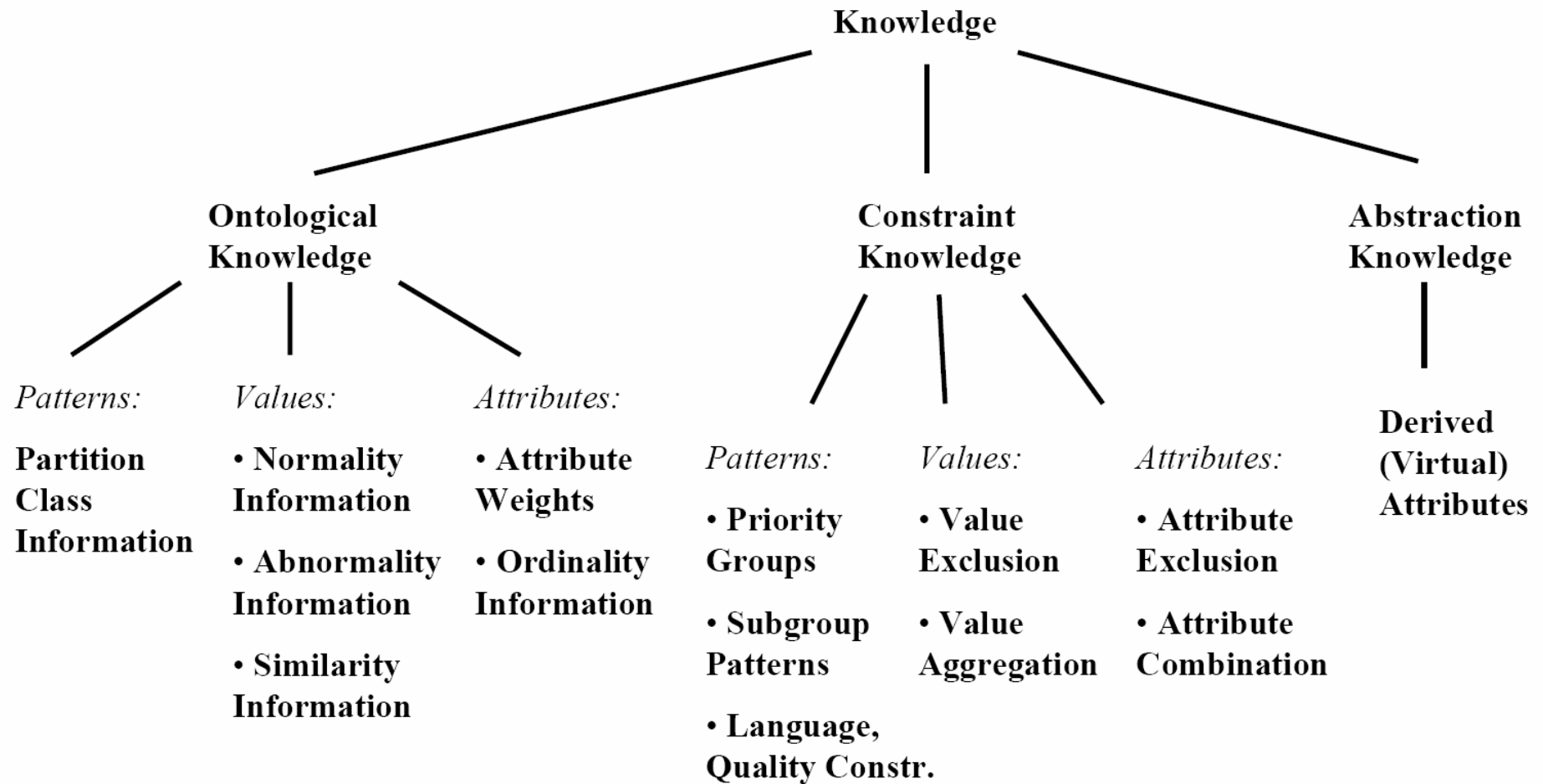


Background Knowledge

5

- **Use common knowledge-concepts**
 - **Reduce KA costs**
 - **“Knowledge-Acquisition Bottleneck”**
=> **Make knowledge formalization easy for the user/domain specialist**
- **Classes of background knowledge:**
 - **Constraints**
 - **Ontological knowledge**
 - **Abstraction knowledge**

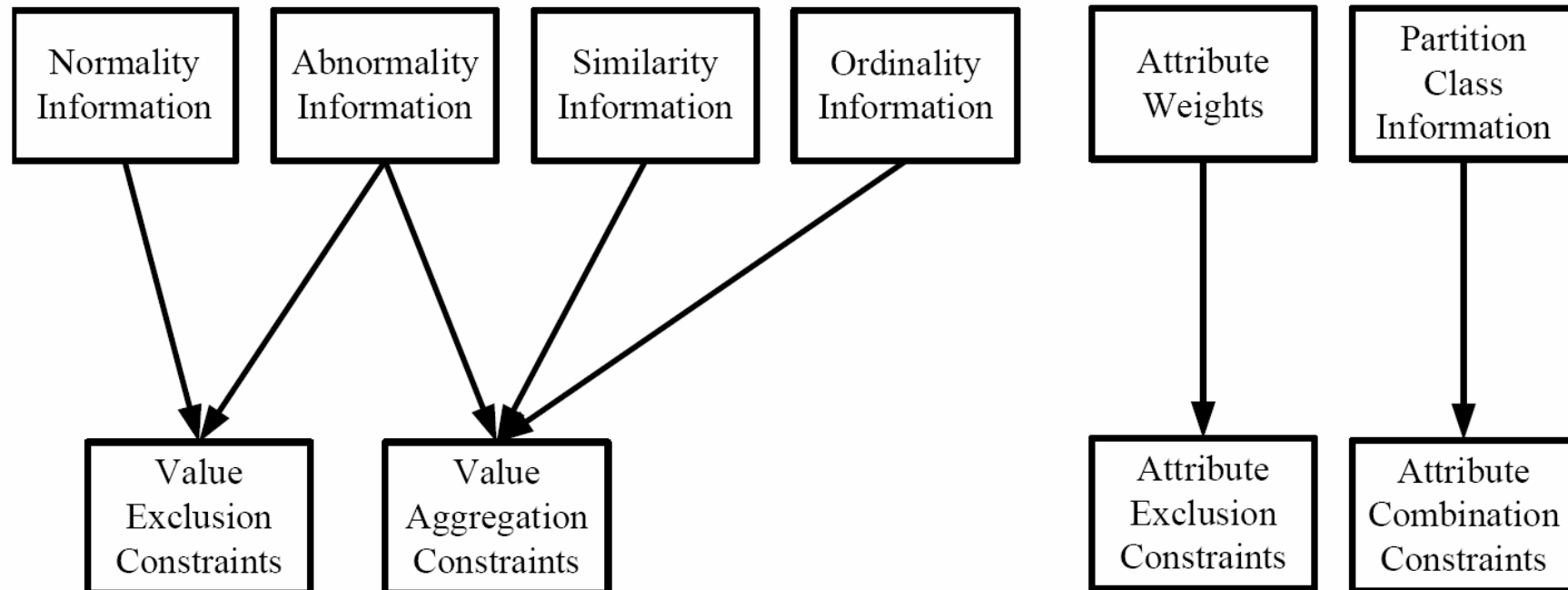




Constraint Derivation

7

Use ontological knowledge to derive constraints



- **Abstraction knowledge**
 - Virtual, attributes derivable by rules
 - Often correspond to known dependencies
 - Can be refined incrementally
- **Main purpose:**
 - Focus SD on relevant concepts
 - Reduce known correlations by data abstraction
 - Improve handling of missing values (especially helpful in medical domains)



How to Use Background Knowledge

9

Tasks:

- Restrict search space
 - Focus SD method
- } Direct integration into subgroup discovery step
- Increase expressiveness
 - Post-processing: increase interestingness of result set of subgroups



- **Embedded in subgroup discovery step**
 - **Quality constraints (thresholds) considering the quality function**
 - **Exclusion, aggregation/combination constraints; ontological knowledge used for its derivation**
- **Subgroup pattern constraints**
 - **Starting points for search**
 - **Inhibit rediscovery of subgroups**
- **Focus only on the most interesting concepts utilizing abstraction knowledge**



Increasing the Expressiveness

11

- **Abstraction Knowledge – Derived Attributes**
 - **Abstract (known) associations into new attributes**
 - **New attributes: more meaningful, reasonable, more interesting**
 - **Minimize missing values**
- **Appropriate aggregated values – value aggregation constraints**
 - **Ordinal groups**
 - **Similarity information**



- **Subgroup Patterns:**
 - **Identify conforming, deviating, contradicting patterns**
 - **Mark patterns conforming to user-defined constraints**
- **Partition class knowledge: e.g., mark subgroups according to inclusion of contained attributes in partition classes**





Case Management

Attribute Navigator Navigator

Attribute Navigator

- Leber, detailliert
- *Pankreas, detailliert
- *Nieren*
- *Nebennieren*
- *Sono Aorta*
- *Sono V. cava*
- *Große Arterien de
- *Nierenarterien*
- *Solide Raumforder
- *Liquide Raumforde
- *Bauchdecke*
- Sonographie *Abdc
- *Leber, Befund
- *DHC*
- *Gallenblase*
- Sonograph
- Gallen
- Gallen
- SI-Gall
- SI-Indiz
- SI-Z. n
- *Pfortadersyste
- *Milz*
- *Pankreas, Bef
- *Prostata, Norm
- *Flüssigkeit im Abdc
- *Pleuraerguss*

Population Info: 7096 from 7096 (total)

Attribute	Population

Current Subgroup

((SI-Leberzirrhose, sonographisch = wahrscheinlich) | Gefäße der Leber, sonographisch=Rarefizierung d SI-Gallenblasenkrankheit, sonographisch=möglich

Quality: 2.99, Size: 304 (8.44%)

Add Clear

649 2953

189 TP, 115 FP

Overview Sorted

Overview (Size: 304, True positive: 189, Analyzed Instances: 3602.0, Missing targ...

Attributes	Sel...	Values
Leberverformbarkeit, sonographisch	<input checked="" type="checkbox"/>	nicht oder kaum vermindert mäßig vermi... deutlich vermindert -?
SI-Altersgruppen	<input checked="" type="checkbox"/>	30-49 50-69 >70
SI-Aortensklerose, sonographisch	<input checked="" type="checkbox"/>	nic... verkalkend *Missing*
SI-Aszites, sonographisch	<input checked="" type="checkbox"/>	vorhanden nicht nachweisbar *Missing*
SI-BMI-Bewertung	<input checked="" type="checkbox"/>	Adi... Adipos... Übergewicht Normalgewicht *Missing*
SI-Cholangitis, sonographisch	<input checked="" type="checkbox"/>	nicht nachweisbar
SI-Cholezystolithiasis, sonographisch	<input checked="" type="checkbox"/>	liegt vor nicht nachweisbar *Missing*
SI-Chronisch degenerative Nierenerkrankung lir	<input checked="" type="checkbox"/>	... nicht ableitbar *
SI-CDN, sonographisch	<input checked="" type="checkbox"/>	Hin... keine Hinweise
SI-Chronisch degenerative Nierenerkrankung r	<input checked="" type="checkbox"/>	mö... nicht ableitbar
SI-Chronische Nephritis, sonographisch	<input checked="" type="checkbox"/>	keine Hinweise *Missi...

Breakpoints Variables Debugger

Statistics for current Subgroup

Property	Value
SG Size	304 (8.44%)
Population	3602
p	0.622
p0	0.180
Rel. Gain	2.989
TP (True Positive)/FP ...	189 (62.17%) / 115 (...)
Significance (P)	<=0.0001

Four-Fields-View

	TV+	TV-
SG+	189	115
SG-	460	2838

Pos. Pred. Value: 0.622
Sensitivity: 0.291
Specificity: 0.961
Significance Level: <=0.0001



- **Presented methodological view on knowledge-intensive SD**
 - **Types/classes of background knowledge**
 - **Benefits/Application**
- **Approach has already been successfully applied in the medical domain**
- **Future work:**
 - **Consider methods for constructive induction of abstractions**
 - **Fine-tuning of aggregations of attribute values**



Questions?



A Methodological View on Knowledge-Intensive Subgroup Discovery

Martin Atzmueller, Frank Puppe
University of Würzburg, Germany

