Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining

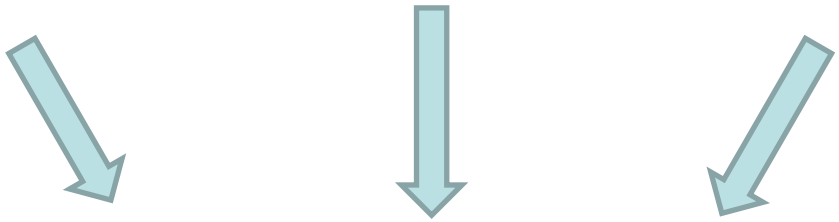
Petra Kralj Novak, Nada Lavrac, Geoffrey I. Webb

TECHNISCHE UNIVERSITÄT DARMSTADT

Contrast
Set
Mining

Emerging
Pattern
Mining

Subgroup
Discovery



Supervised Descriptive Rule Discovery

Inhaltsangabe

TECHNISCHE UNIVERSITÄT DARMSTADT

Einführung

TECHNISCHE UNIVERSITÄT DARMSTADT

Symbolic Data Analysis Techniques

- **Descriptive Induction**
  - Unlabeled Data
  - Erkennen von Mustern

  ➔ Contrast Set Mining

  Emerging Pattern Mining

- **Predictive Induction**
  - Labeled Data
  - Klassifizierung von Beispielen

  ➔ Subgroup Discovery

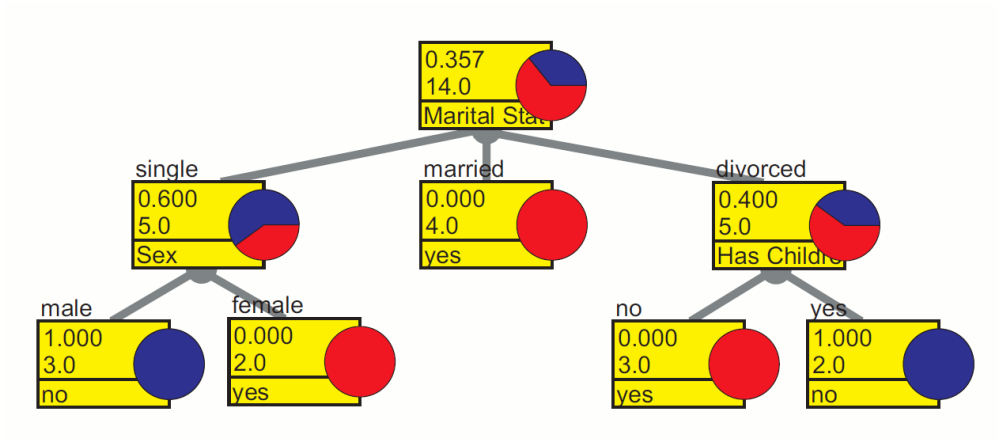
➔ Hauptziel: menschlich interpretierbare Unterschiede zwischen Gruppen finden

Beispiel Datenbank

TECHNISCHE UNIVERSITÄT DARMSTADT

Education	Marital Status	Sex	Has Children	Approved
primary	single	male	no	no
primary	single	male	yes	no
primary	married	male	no	yes
university	divorced	female	no	yes
university	married	female	yes	yes
secondary	single	male	no	no
university	single	female	no	yes
secondary	divorced	female	no	yes
secondary	single	female	yes	yes
secondary	married	male	yes	yes
primary	married	female	no	yes
secondary	divorced	male	yes	no
university	divorced	female	yes	no
secondary	divorced	male	no	yes

Predictive Induction



- C4.5 (1993)
- Hoher Wert auf Vollständigkeit

```
Sex = female → Approved = yes
MaritalStatus = single AND Sex = male → Approved = no
MaritalStatus = married → Approved = yes
MaritalStatus = divorced AND HasChildren = yes → Approved = no
MaritalStatus = divorced AND HasChildren = no → Approved = yes
```

Descriptive Induction

TECHNISCHE UNIVERSITÄT DARMSTADT

```
MaritalStatus = single AND Sex = male  →  Approved = no
Sex = male  →  Approved = no
Sex = female  →  Approved = yes
MaritalStatus = married  →  Approved = yes
MaritalStatus = divorced AND HasChildren = yes  →  Approved = no
MaritalStatus = single  →  Approved = no
```

- Class Attribut entspricht normalem Attribut

- Regeln redundant
- Regeln fehlen

- Aufgabenstellung
- Algorithmus
- Constraints des Benutzers

Inhaltsangabe

TECHNISCHE UNIVERSITÄT DARMSTADT

2.1 Contrast Set Mining

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ziel: Finde ContrastSets mit denen eine Gruppe von anderen unterschieden werden kann

$$ContrastSet_i \rightarrow Gi \qquad ContrastSetj \rightarrow Gj \qquad \ldots\ldots$$

- Qualitätsbewertung:

$$SuppDiff(X, \mathrm{G_i}, \mathrm{G_j}) = |support(X, \mathrm{G_i}) - support(X, \mathrm{G_j})| \geq \delta$$

2.2 Emerging Pattern Mining

- Ziel: ItemSets deren Support signifikant von einem Datenset zu einem anderen steigt

$$ItemSet_1 \rightarrow D_1 \quad ItemSet_2 \rightarrow D_2$$

- aufkommende Trends in zeitlich gemessenen Datenbanken oder differenzierende Merkmale zwischen Klassen von Daten finden
- Qualitätsbewertung:

$$GrowthRate(ItemSet, D_1, D_2) = \frac{support(ItemSet, D_1)}{support(ItemSet, D_2)}$$

2.3 Subgroup Discovery

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ziel: finde UG, die statistisch gesehen am interessantesten sind bezüglich einer Property of Interest

$$SubgroupDescription \rightarrow Class$$

- Qualitätsbewertung:

$$WRAcc(X, C) = P(X) \cdot \big(P(Y|X) - P(Y)\big) \qquad q_g(X, C) = \frac{p}{n+g}$$

- Zielt auf ein gutes Maß zwischen Rule Coverage und Precision ab

Inhaltsangabe

TECHNISCHE UNIVERSITÄT DARMSTADT

3. Ein einheitliches Framework - Inhalt

TECHNISCHE UNIVERSITÄT DARMSTADT

- Terminologie
- Task Definitions
- Rule Learning Heuristics
- Rule Selection Mechanism

3.1 Vereinheitlichung der Terminologie

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition 1: Kompatibilität von Termen

Terme sind kompatibel, wenn sie …

- *… in äquivalente logische Ausdrücke umgeformt werden können*
- *… die selbe Bedeutung haben*

3.1 Vereinheitlichung der Terminologie

TECHNISCHE UNIVERSITÄT DARMSTADT

Lemma 1: Die Terme in CSM, EPM und SD sind kompatibel

Contrast Set Mining	Emerging Pattern Mining	Subgroup Discovery	Rule Learning
contrast set	itemset	subgroup description	rule condition
groups $G_1, \ldots G_n$	data sets D_1 and D_2	class/property C	class/concept C_i
attribute-value pair	item	logical (binary) feature	condition
examples in groups $G_1, \ldots G_n$	transactions in data sets D_1 and D_2	examples of C and $\overline{C}$	examples of $C_1 \ldots C_n$
examples for which the contrast set is true	transactions containing the itemset	subgroup of instances	covered examples
support of contrast set on G_i support of contrast set on G_j	support of EP in data set D_1 support of EP in data set D_2	true positive rate false positive rate	true positive rate false positive rate

3.2 Vereinheitlichung der Task Definitions

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition 2: Kompatibilität von Task Definitionen

Definitionen sind kompatibel, wenn …

- *… ein Learning Task zu einem anderen umgewandelt werden kann, ohne das Lernziel zu verändern*

3.2 Task Definitions

- CSM
  - Finde ContrastSets mit denen man am besten Gruppen voneinander unterscheiden kann
- EPM
  - Finde ItemSets deren Support signifikant von einem DatenSet zu einem anderen steigt
- SD
  - Finde UG die so groß wie möglich sind und möglichst ausgefallene statistische Charakteristika besitzen bezüglich der Property of Interest

3.2 Vereinheitlichung der Task Definitions

TECHNISCHE UNIVERSITÄT DARMSTADT

Lemma 2:
Die Definitionen von CSM, EPM und SD sind kompatibel

$$EPM(D_1, D_2) \Leftrightarrow CSM(Gi, Gj)$$
$$CSM(G_i, G_j) \Leftrightarrow SD(G_i) \; und \; SD(G_j)$$

Contrast Set Mining	Emerging Pattern Mining	Subgroup Discovery	Rule Learning
Given	**Given**	**Given**	**Given**
examples in G_1 vs. G_j	transactions in D_1 and D_2	in examples C	examples in C_i
from $G_1, \ldots G_i$	from D_1 and D_2	from C and $\overline{C}$	from $C_1 \ldots C_n$
Find	**Find**	**Find**	**Find**
$ContrastSet_{i_k} \rightarrow G_i$	$ItemSet_{1_k} \rightarrow D_1$	$SubgrDescr_k \rightarrow C$	$\{RuleCond_{i_k} \rightarrow C_i\}$
$ContrastSet_{j_l} \rightarrow G_j$	$ItemSet_{2_l} \rightarrow D_2$		

3.3 Vereinheitlichung der Rule Learning Heuristics

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition 3: Kompatibilität von Heuristiken

Heuristik h_1 ist kompatibel mit Heuristik h_2, wenn …

- *… h_2 von h_1 abgeleitet werden kann*
- *… für zwei Regeln R und R‘ gilt: $h_1(R) > h_1(R') \Leftrightarrow h_2(R) > h_2(R')$*

3.3 Rule Quality Measures

TECHNISCHE UNIVERSITÄT DARMSTADT

- CSM

$$SuppDiff(X, G_i, G_j) = |support(X, G_i) - support(X, G_j)| \geq \delta$$

- EPM

$$GrowthRate(X, D_1, D_2) = \frac{support(X, D_1)}{support(X, D_2)}$$

- SD

$$WRAcc(X, C) = P(X) \cdot \big(P(Y|X) - P(Y)\big) \qquad q_g(X, C) = \frac{p}{n+g}$$

$$support(X, Y) = \frac{count(X,Y)}{|Y|} = TPr(X, Y)$$

3.3 Vereinheitlichung der Rule Learning Heuristiken

TECHNISCHE UNIVERSITÄT DARMSTADT

Lemma 3a: Support Difference *Heuristik von CSM und* Weighted Relative Accuracy *von SD sind kompatibel*

$$
\begin{aligned}
&WRAcc(X,Y) = \\
&\quad = P(X)\cdot[P(Y|X) - P(Y)] = P(Y\cdot X) - P(Y)\cdot P(X) \\
&\quad = P(Y\cdot X) - P(Y)\cdot[P(Y\cdot X) + P(\overline{Y}\cdot X)] \\
&\quad = (1 - P(Y))\cdot P(Y\cdot X) - P(Y)\cdot P(\overline{Y}\cdot X) \\
&\quad = P(\overline{Y})\cdot P(Y)\cdot P(X|Y) - P(Y)\cdot P(\overline{Y})\cdot P(X|\overline{Y}) \\
&\quad = P(\overline{Y})\cdot P(Y)\cdot[P(X|Y) - P(X|\overline{Y})] \\
&\quad = P(Y)\cdot P(\overline{Y})\cdot[TPr(X,Y) - FPr(X,Y)]
\end{aligned}
$$

konstant

$$WRAcc(X,C) = WRAcc(X,G_1) = P(G_1)\cdot P(G_2)\cdot[support(X,G_1) - support(X,G_2)].$$

3.3 Vereinheitlichung der Rule Learning Heuristiken

TECHNISCHE UNIVERSITÄT DARMSTADT

Lemma 3b: Die Growth Rate *Heuristik vom EPM und die* Generalization Quotient *Heuristik sind kompatibel*

$$GrowthRate(X, D_1, D_2) = \frac{support(X, D_1)}{support(X, D_2)}$$

$$= \frac{count(X, D_1)}{count(X, D_2)} \cdot \frac{|D_2|}{|D_1|} = \frac{p}{n} \cdot \frac{N}{P}$$

konstant

$$\Rightarrow GrowthRate(X, C, \overline{C}) = q_0(X, C) \cdot \frac{N}{P}$$

3.3 Vereinheitlichung der Rule Learning Heuristiken

TECHNISCHE UNIVERSITÄT DARMSTADT

Lemma 3: Die Definitionen der CSM, EPM und SD Heuristiken sind paarweise kompatibel

Contrast Set Mining	Emerging Pattern Mining	Subgroup Discovery	Rule Learning
$SuppDiff(X, G_i, G_j)$		$WRAcc(X, C)$	Piatetski-Shapiro heuristic leverage
	$GrowthRate(X, D_1, D_2)$	$q_g(X, C)$	odds ratio for $g = 0$ accuracy/precision, for $g = p$

Zielen alle auf ein gutes Maß zwischen Rule Coverage und Precision ab

3.4 Vergleich der Rule Selection Mechanismen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Statistic Test
  - Contrast Set Mining
  - Statistical significance pruning
  - Entfernt alle ContrastSets, die Spezialisierungen von generelleren CS sind, wenn sie einen ähnlichen Support haben

- Weighted Covering Approach
  - Subgroup Discovery
  - Benutzt gewichtete Version von q_g und WRAcc
  - Nach jeder Iteration wird die Gewichtung von pos. Bsp. Gesenkt

$$q'_g(X,Y) = \frac{p'}{n+g}$$

Inhaltsangabe

Fazit

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ziel: Vereinheitlichung von Bereichen
- nur Terminologie, Task Definition und Heuristiken werden verglichen
- Rule Selection Mechanism unterscheiden sich!

  ➔ andere Schwerpunkte

Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining

Petra Kralj Novak, Nada Lavrac, Geoffrey I. Webb

TECHNISCHE UNIVERSITÄT DARMSTADT

Vielen Dank für Eure Aufmerksamkeit!