Dolování z řešených úloh konstrukčního typu

Karel Vaculík

Cikháj, 2013

Obsah

- Konstrukční úlohy z logiky
- Data mining na grafových datech
- Časté vzory
- Klasifikace

Konstrukční úlohy

- Řešení lze rozložit na podčásti (obvykle kroky) a vztahy mezi nimi (může být i časová následnost)
- Vhodná reprezentace: grafy

Rezoluční důkazy

- IB101 Úvod do logiky a logického programování
- Písemný test: příklad na rezoluci pro výrokovou logiku
  - Úkol: provést rezoluční důkaz, určit typ použité rezoluce (lineární, LI, LD, SLD, ...)

Rezoluční důkazy - příklad

- Zádána formule:

{-A,C}, {-A,E}, {-C,D},

{-C,-F}, {B,C}, {B,-C},

{A,-B,C}, {-B}, {B,E},

{B,D}

Tj.:

$(\neg A \vee C) \wedge (\neg A \vee E) \wedge (\neg C \vee D)$ ...

Rezoluční důkazy - příklad

- Zádána formule:

{-A,C}, {-A,E}, {-C,D},

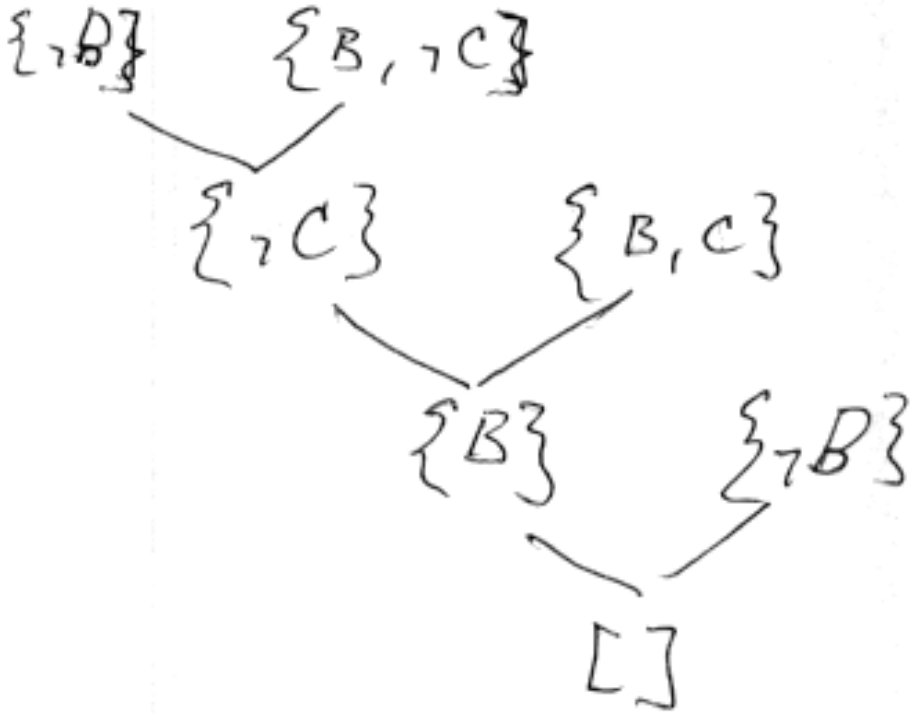
{-C,-F}, {B,C}, {B,-C},

{A,-B,C}, {-B}, {B,E},

{B,D}

Tj.:

$(\neg A \vee C) \wedge (\neg A \vee E) \wedge (\neg C \vee D)$ ...



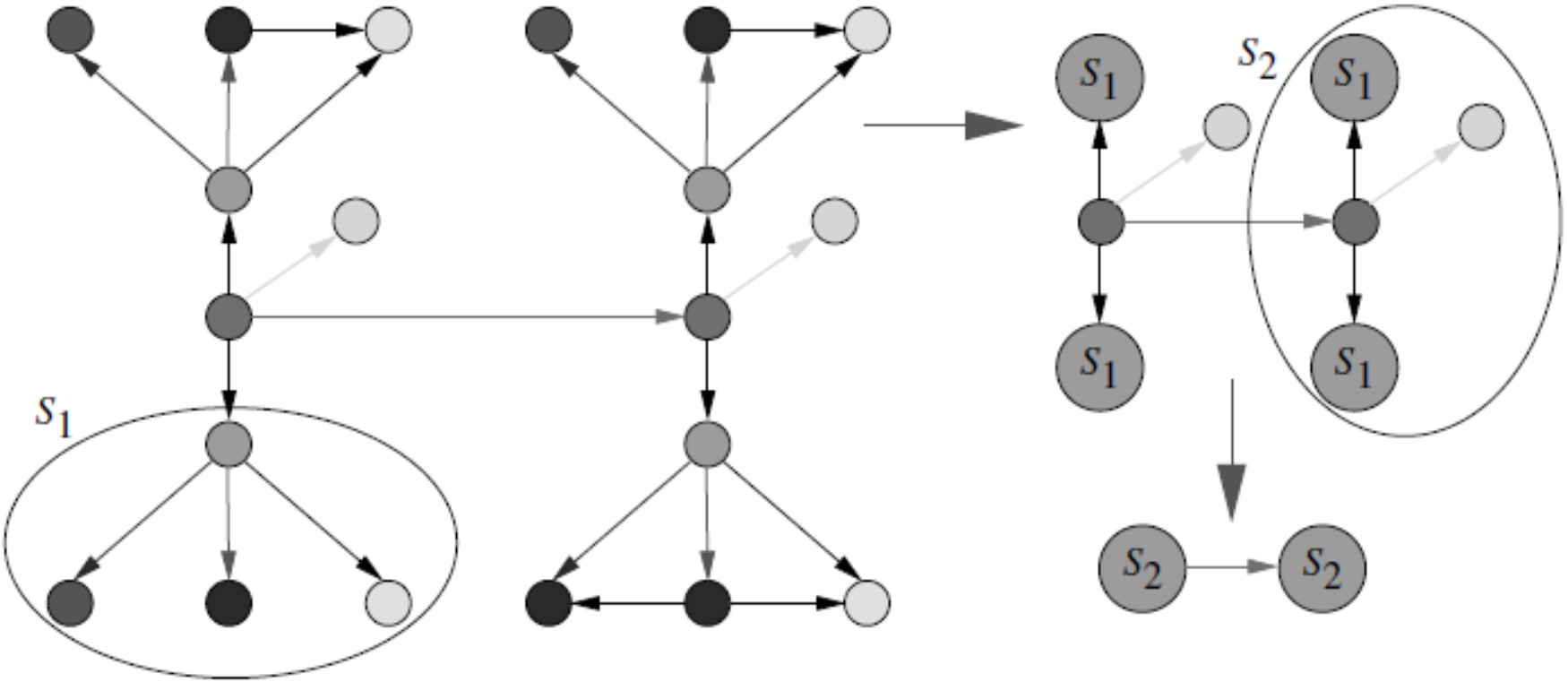
Rezoluční důkazy – data

- 393 řešení => 393 rezolučních stromů
- 2 zadání (183 + 210 stromů)
- Rezoluce provedena:
  - Správně (třída *positive*): 322 případů
  - Špatně (třída *negative*): 71 případů
- Další parametry: počet získaných bodů, typ rezoluce podle studenta, typ skutečně provedené rezoluce, ...
- Cíl: získat nějaké zajímavé informace, které by pomohly zlepšit výuku

Data mining na grafových datech

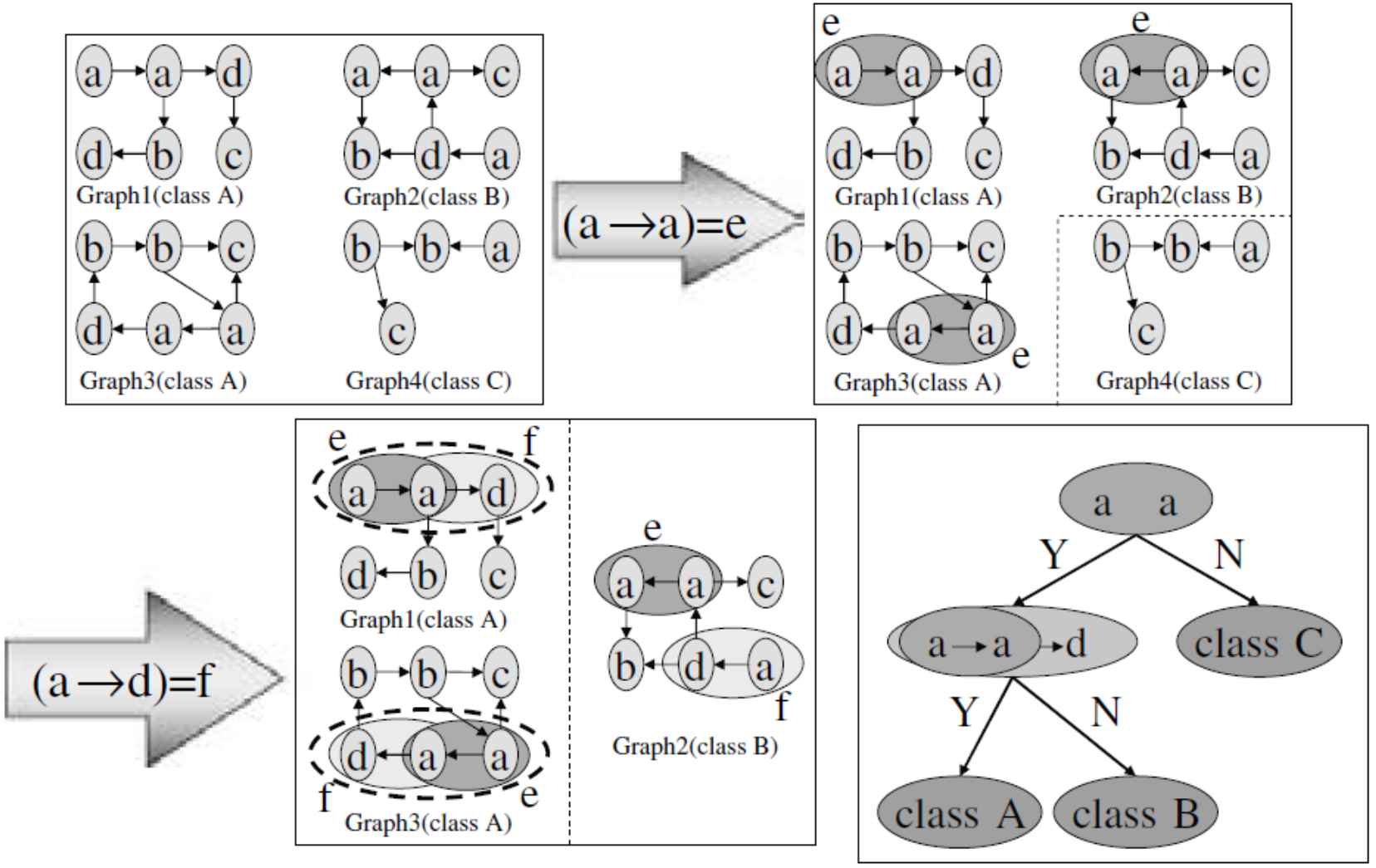
Data mining na grafových datech

- Hledání častých vzorů



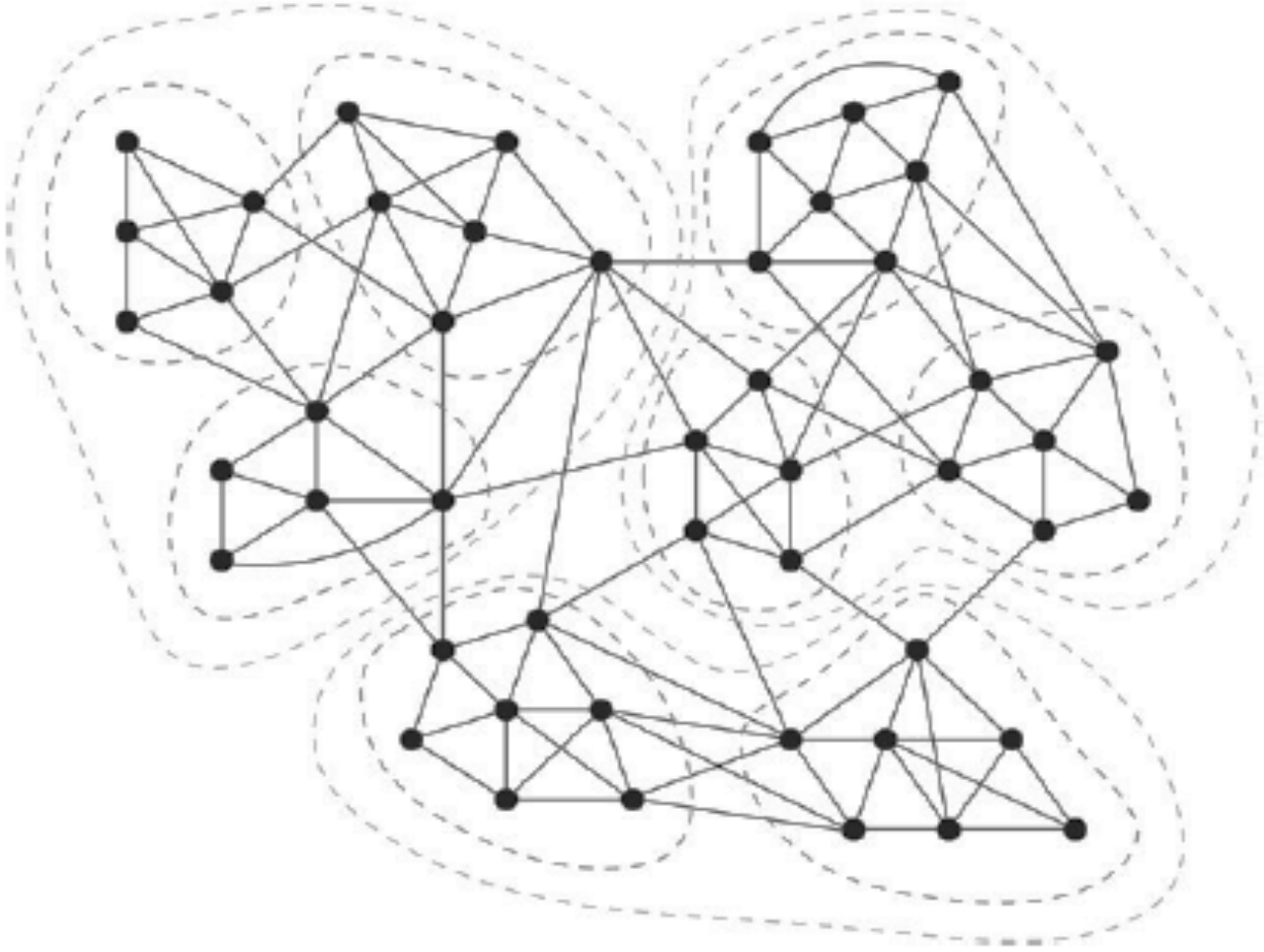
Data mining na grafových datech

- Klasifikace



Data mining na grafových datech

- Shlukování



Data mining na grafových datech

Algoritmy pro práci s:

- Grafy obecně
  - Neorientované grafy
  - Orientované grafy

Data mining na grafových datech

Algoritmy pro práci s:

- Grafy obecně
  - Neorientované grafy – nevhodné pro rezoluční stromy
  - Orientované grafy – publikované algoritmy nejsou k dispozici, nenaleznou všechny časté podgrafy, ...

Data mining na grafových datech

Algoritmy pro práci s:

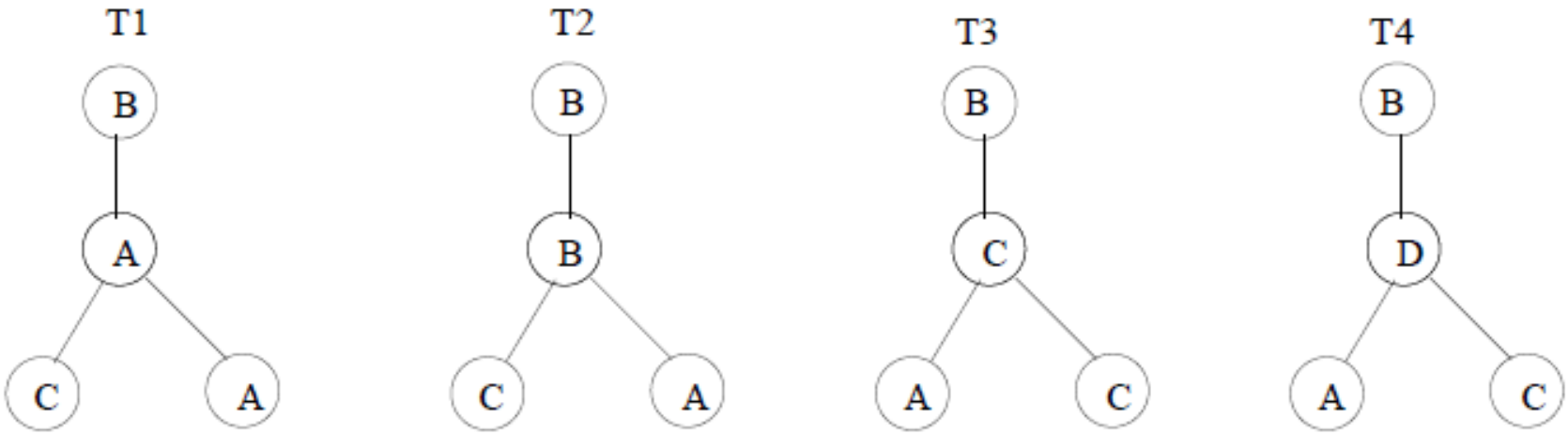
- Grafy obecně
  - Neorientované grafy – nevhodné pro rezoluční stromy
  - Orientované grafy – publikované algoritmy nejsou k dispozici, nenaleznou všechny časté podgrafy, ...
- Stromy

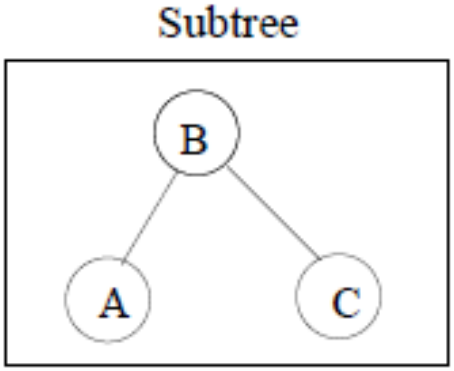
Sleuth

- Algoritmus pro hledání častých vzorů na stromech
- Pro kořenové stromy, uspořádané i neuspořádané
- Hledá časté podstromy *induced* i *embedded*

Sleuth

- Induced vs. embedded:





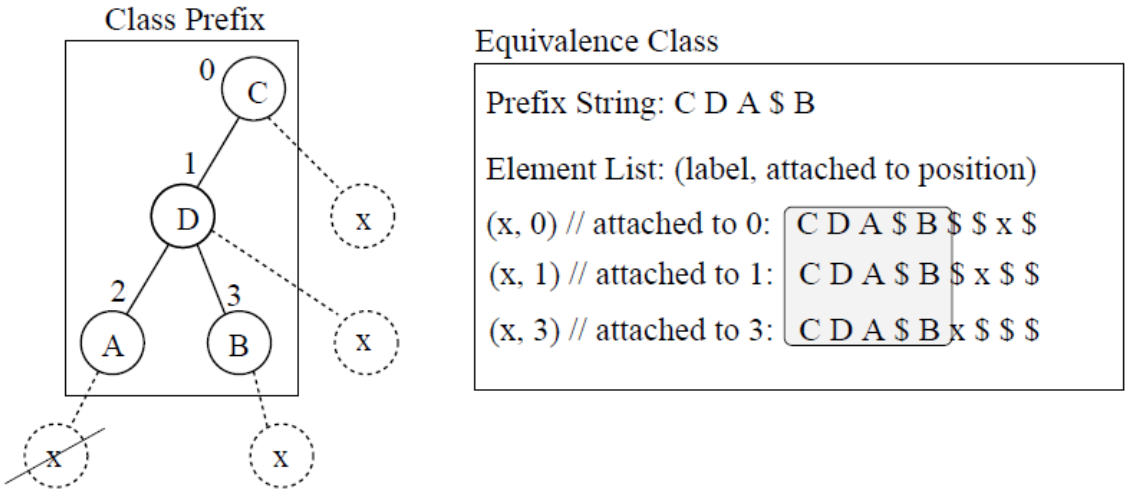
Induced: T2

Embedded: T1, T2, T3, T4

Sleuth

Strategie:

1) Generování stromů:



   – Pro neuspořádané stromy kanonická forma

2) Odfiltrování vzorů s nízkou hodnotou support

Časté vzory v rezol. důkazech

- Použito všech 393 příkladů – 2 zadání nebyla tak odlišná, aby to výrazně ovlivnilo výsledky
- Neuvažováno pořadí literálů v klauzulích (tzn. {C, -B} = {-B, C})

Typ	**počet**
<	1
F	5
LD	12
LI	32
Linear	261
S	2
SLD	10
none	65
unknown	5

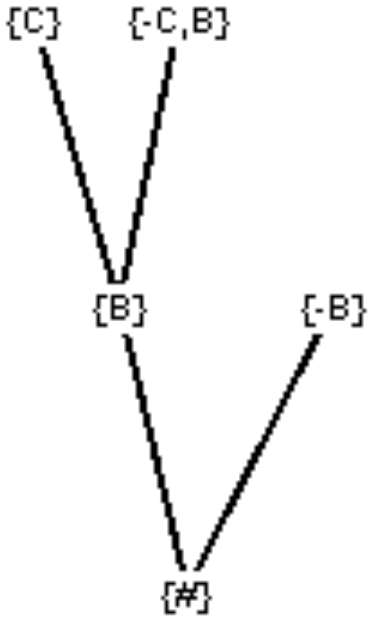
Časté vzory v rezol. důkazech

- S menším počtem uzlů nejčastěji (~ 90 %)

  {#}; {-C,B}; ... ne příliš zajímavé

- S vyšším počtem uzlů:

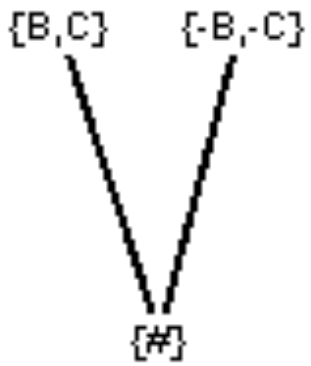
  ~ 28 %



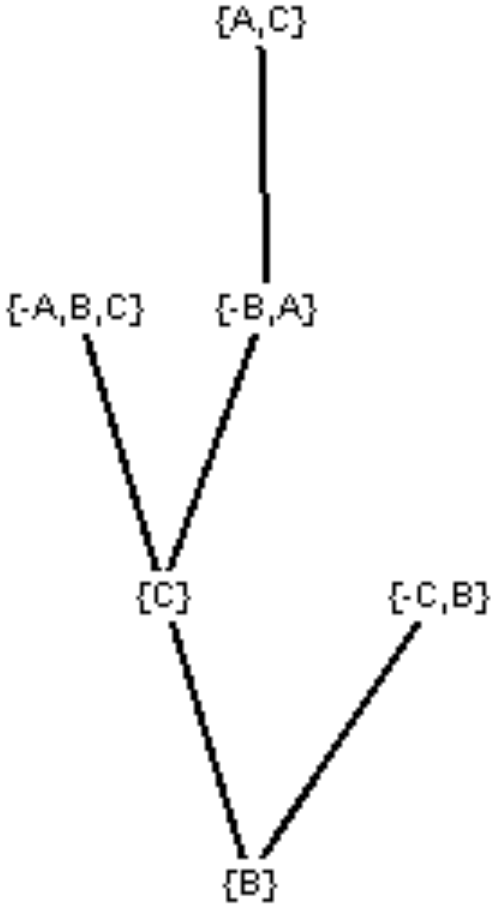
Časté vzory v rezol. důkazech

- Časté vzory pouze u negativních příkladů:

~ 5 %



~ 2 %

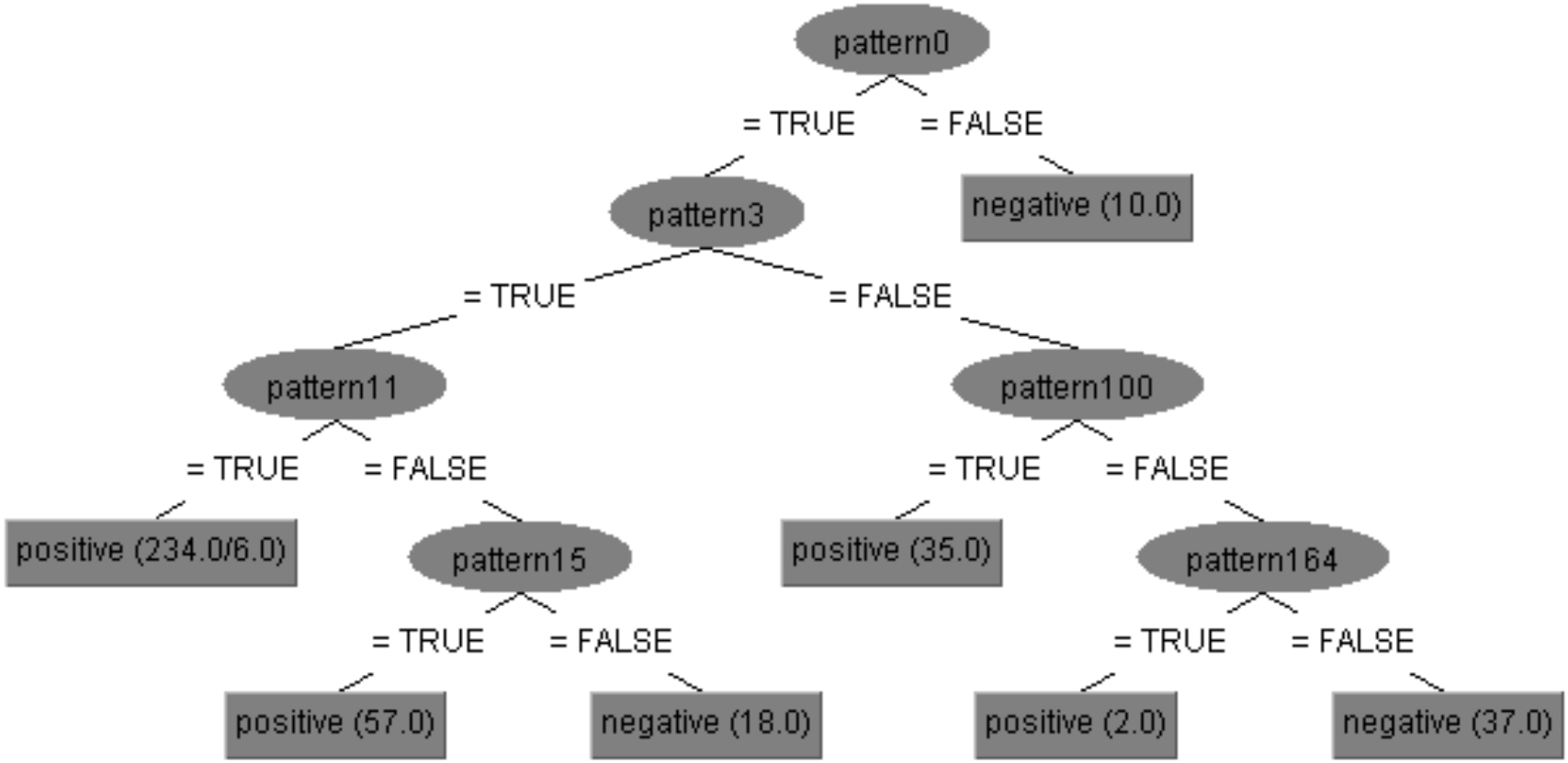


Klasifikace

- Na všech datech (393 př.), atributy = vzory
- Min. support 10 % => 171 vzorů
- Třídy: *Positive* (322 př.), negative (71 př.)
  => *Baseline*: 322/(322+71) * 100 % = 81,93 %
- Reprezentace dat:

	Vzor 1	Vzor 2	...	Vzor 171	Třída
Strom 1	1	1	...	0	positive
Strom 2	1	0	...	0	positive
...	...	...	...	...	...
Strom 393	1	1	...	0	positive

Klasifikace – J48 (Weka)



Správnost: 97,96 %

Další kroky

- Zvolit vhodnou metriku pro posouzení relevance (support vs. velikost vzoru vs. ... ?)
- Použití dalších informací (chyby nalezené opravujícím, typ rezoluce, ...)
- Odfiltrovat neužitečné vzory (pro danou hodnotu support vzít ten největší, jeho podvzory neuvažovat, ...)
- V uzlech vyzkoušet jiný typ informace
- ...

Děkuji za pozornost

Použité zdroje

- M. J. Zaki. Efficiently mining frequent embedded unordered trees. *Fundamenta Informaticae*, 66(1-2):33–52, 2005.
- Diane J. Cook, Lawrence B. Holder. Mining graph data. John Wiley and Sons, 2007.
- Charu C. Aggarwal, Haixun Wang. Managing and Mining Graph Data. Springer, 2010
- Satu Elisa Schaeffer. Graph clustering. Journal Computer Science Review, 2007