Automatická klasifikace prostorových dat

Petr Kuba, Luboš Popelínský
Fakulta informatiky. Masarykova universita
Botanická 68a, 602 00 Brno
Email: {xkuba,popel}@fi.muni.cz

Klasifikace

- najít rozhodovací strom, který pro každý objekt na základě jeho atributů rozhodne, do které třídy patří.
- používá učící množinu
- strom by měl rozhodovat i příklady, které nejsou v učící množině

Prostorová klasifikace

- při tvorbě stromu neuvažujeme jen atributy objektu, ale i atributy jeho sousedů, tj. objektů na cestě v grafu sousednosti
- **Generalizovaný atribut objektu** - dvojice (jméno_atributu, index), kde
  - index je pozice sousedícího objektu v cestě
  - jméno_atributu je jeho atribut

Příklad klasifikace

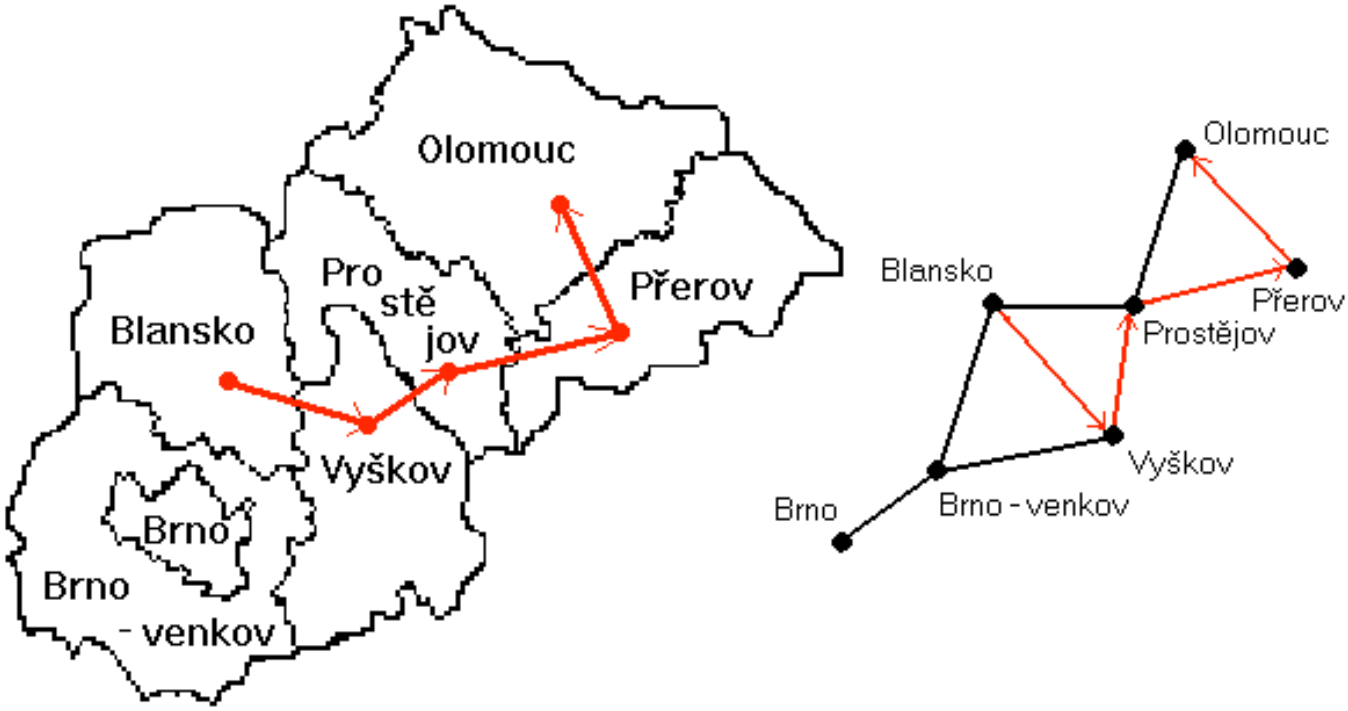
Učící množina:

Jméno	Rozl.	Obcí	Obyvatel	Nezam.
Blansko	942	129	107973	stredni
Brno_m	230	1	384369	nizka
Brno_v	1109	137	158398	nizka
Vyskov	889	80	86752	stredni
Prostejov	770	95	110088	stredni
Prerov	884	103	136845	vysoka
Olomouc	1451	93	225599	stredni

Rozhodovací strom:

obyvatel <= 110088 : stredni (3.0/0.0)
obyvatel > 110088 :
| obyvatel <= 136845 : vysoka (1.0/0.0)
| obyvatel > 136845 :
| | rozloha <= 1109 : nizka (2.0/0.0)
| | rozloha > 1109 : stredni (1.0/0.0)

Příklad prostorové klasifikace



Objekt Blansko má pak tyto generalizované atributy:

- (jmeno, 0), (rozloha, 0), . . . - jeho vlastní atributy
- (jmeno, 1), (rozloha, 1), . . . - atributy přímého souseda, tj. okresu Vyškov
- (jmeno, 2), (rozloha, 2), . . . - atributy dalšího objektu na cestě, tj. okresu Prostějov

⋮

Operace nad prostorovými daty

Graf sousednosti G pro prostorovou relaci $soused/2$ je graf (U, H), kde

- U je množina uzlů reprezentujících objekty
- H je množina hran. Dva uzly N1, N2 jsou spojeny hranou právě tehdy, když platí: $soused(objekt(N1), objekt(N2))$

Relace soused může vyjadřovat např. následující:

- topologickou relaci, např. objekty se dotýkají, překrývají, jsou totožné, jeden obsahuje druhý
- relaci míry, např. vzdálenost objektů ¡ d
- směrovou relaci, např. severně, jižně, východně, západně

Cesta v grafu sousednosti pro graf G je posloupnost [n_0, n_1, ..., n_{k-1}], kde

- n_i je uzel z G
- (n_i, $n_i + 1$) je hrana z G, 0 $\leq$ i ¡ k-1

Základní operace:

- **get_Graph(rel)** — vrací graf sousednosti reprezentující relaci **rel**.
- **get_Neighbourhood(G, o)** — vrací množinu objektů spojených s objektem **o** nějakou hranou z grafu **G**
- **create_Paths(G, i)** — vrací množinu všech cest, které jsou tvořeny uzly a hranami z grafu **G**, jejichž délka je $\leq$ **i**

Implementace

- Postgres - objektově relační databáze
- C4.5 - program pro klasifikaci
- Regression tree - klasifikuje do spojité množiny tříd; není potřeba diskretizovat atribut, podle kterého se klasifikuje; pro spojitá data dává lepší výsledky
- v databázi uložena:
  - popisná data o objektech
  - graf sousednosti
- vytvoříme cesty v grafu pomocí funkce create_Paths
- vytvoříme tabulku obsahující generalizované atributy (operace select)
- klasifikujeme připravená data pomocí C4.5 a regression tree

Aplikace

- statistická data o okresech ČR
- graf sousednosti
- klasifikujeme nezaměstnanost

Výsledný strom:

nezam_1_misto1 $\leq$ 17.05
True (217 of 331)
    zamestnancu1 $\leq$ 22658.00
    True (106 of 217)
        obci1 $\leq$ 42.50
        True (5 of 106)
            LEAF :: Y = 9.71
        False (101 of 106)
            LEAF :: Y = 6.54
    False (111 of 217)
        staveb_prace1 $\leq$ 418.50
        True (87 of 111)

LEAF :: Y = 8.40

False (24 of 111)

LEAF :: Y = 6.38

False (114 of 331)

nezam_1_misto1 $\leq$ 26.70

True (59 of 114)

LEAF :: Y = 9.80

False (55 of 114)

LEAF :: Y = 11.95

Tento strom odhaduje nezaměstnanost s přesností na 2%.