Algoritmy pro shlukování prostorových dat

Marta Žambochová

Katedra matematiky a informatiky
Fakulta sociálně ekonomická
Univerzita J. E. Purkyně v Ústí nad Labem

ROBUST
21.– 26. leden 2018
Rybník - Hostouň

Motivace

- existence velkého množství prostorových databází
- prostorové databáze jsou většinou velmi rozsáhlé
- potřeba analýzy prostorových dat
- potřeba efektivních metod pro shlukovou analýzu dat dodržující prostorovou vazbu objektů

Geografická data (geodata)

- **Prostorová data**
  - data obsahující prostorové určení a prostorové vztahy (topologii) objektu
- **Atributová data**
  - popisují kvalitativní a kvantitativní charakteristiky prostorových dat

Reprezentace prostorových dat

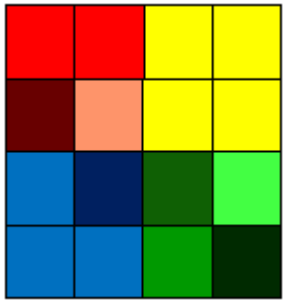
- Vektorový model
- Rastrový model
- Vrstvový přístup
  - jednotlivá data jsou obvykle organizována v tematických vrstvách
- Objektový přístup
  - založen na principech objektově orientovaného programování
  - každý objekt obsahuje geometrii, topologii, tématiku (atributy) a dále i chování (metody)

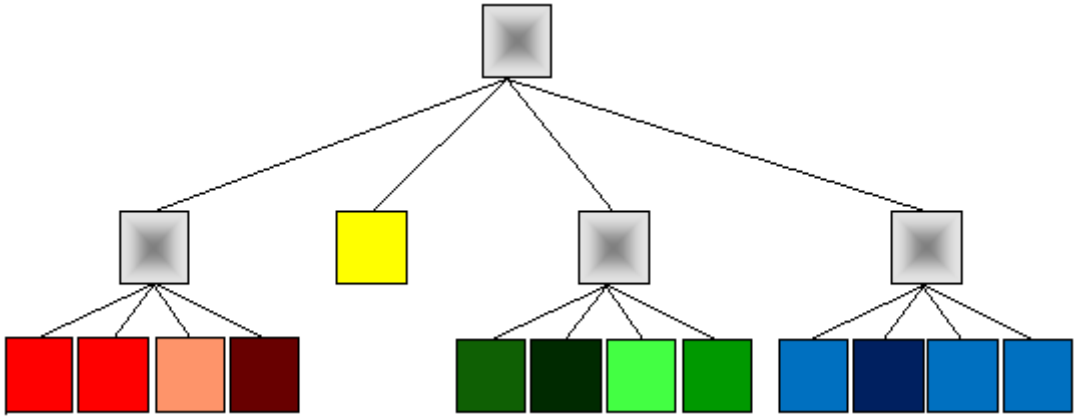
Vektorová reprezentace

- **Bod (Point)**
  - jeho dimenze je 0
- **Linie (Line)**
  - její dimenze je 1
- **Řetězec linií (PolyLine)**
  - jeho dimenze je 1
  - **Plocha (area) =** polygon
  - její dimenze je 2
- **Topologie**
  - prostorové vztahy mezi jednotlivými geometrickými objekty

Rastrová reprezentace

- Nejčastěji se používá **čtvercová mřížka**
- **Topologie** je v rastrovém modelu definována implicitně (je jasné kdo je čí soused), tudíž není nutné ji explicitně ukládat jako pro vektorový model
- Komprimace uložení např. pomocí Quad stromu





PROSTOROVÁ AUTOKORELACE

- měří závislost mezi prostorovými daty
- korelace výskytu jevu v prostoru
  s výskytem tohoto jevu v blízkém okolí
- výsledná hodnota autokorelačních statistik
  (lokální, globální) vypovídá o míře
  prostorového shlukování
- Jedna z nejpoužívanějších charekteristik je
  Moranovo I kritérium a jeho lokální obdoba
  LISA

Moranovo I kriterium

- výpočet i interpretace jsou velmi podobné Pearsonovu korelačnímu koeficientu

$$c_{ij} = (x_i - \bar{x})(x_j - \bar{x})$$

- $I = \frac{\sum_i \sum_j w_{ij} c_{ij}}{s^2 \sum_i \sum_j w_{ij}}$ kde $s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n}$

- kde *n je počet analyzovaných jednotek, x_i je hodnota proměnné v i-té jednotce a $\overline{x}$- aritmetický průměr sledované p*roměnné

LISA
Local Indicators of Spatial Association

- součet všech jednotlivých hodnot je úměrný globální hodnotě Moranovy statistiky
- na základě výpočtu LISA můžeme rozdělit sledované jednotky podle typu prostorové autokorelace do čtyř skupin, viz tabulka
- prostorové shluky vykazující nadprůměrné či podprůměrné hodnoty proměnné v určité jednotce souhlasně s jejím okolím prokazují pozitivní prostorovou autokorelaci

Moranův diagram	hodnota proměnné v prostorové jednotce	
vážená hodnota proměnné v blízkých jednotkách	**nízká – vysoká** negativní prostorová autokorelace	**vysoká - vysoká** pozitivní prostorová autokorelace
	nízká – nízká pozitivní prostorová autokorelace	**vysoká – nízká** negativní prostorová autokorelace

Problém prostorových shluků

- pro prostorový shluk se předpokládá, že je souvislý v prostoru prostorových atributů
- buď se musí řešit otázka sousednosti
- nebo se hledají různé heuristiky, aby se našla forma prostorových shluků, aniž by se musela řešit otázka sousedství
- v prototypových prostorových shlukovacích algoritmech je zaručeno získání sousedních prostorových shluků bez nutnosti vědět, které seskupení jsou sousední
- tyto algoritmy maximalizují funkce odměňování, které podporují sloučení podobných sousedních shluků a rozdělení nehomogenních shluků, pokud vede k významnému nárůstu v celkové odměně

PROSTOROVÉ HIERARCHICKÉ SHLUKOVÁNÍ

- klasické metody hierarchického shlukování nevyužívají prostorové vlastnosti, vytvářejí shluky objektů, které nerespektují prostorové vztahy (zejména požadavek kontinuity jednotlivých shluků)
- odlišnosti PHS od klasického HS v počáteční fázi algoritmu
  - určení sousedství jednotlivých objekty – zpravidla se používají topologická sousedství typu královna
  - matice vzdáleností se vypočítá pouze mezi všemi páry tvořenými sousedními objekty
- zachovává výhody klasického HS (hlavně názornost výstupu v podobě dendogramu či absence potřeby zadání počtu shluků)
- zachovává topologické vlastnosti prostorových dat

CLARANS

Clustering Large Application based on RANdomized Search

- základem algoritmu je metoda *k*-medoidů
- postup algoritmu
  - v 1. kroku náhodně vybere *k* medoidů, každý ze zbývajících objektů přiřadí k jim nejbližšímu medoidu
  - ve 2. kroku zkoumá, zda záměnou náhodně vybraného medoidu a náhodně vybraného objektu, který není medoidem došlo ke zlepšení, v případě zlepšení provede záměnu
  - po provedení daného počtu porovnání a případných záměn algoritmus spočítá a uloží aktuální průměrnou vzdálenost
  - postup se opakuje od prvního kroku, od náhodného výběru *k* medoidů
- lze použít ne jen v případech, kdy je možno definovat průměr, ale i pokud je definována míra podobnosti mezi dvěma objekty
- značná efektivnost při zpracování velkých souborů dat
- robustnost - odlehlé objekty nevytvářejí vždy samostatný shluk.

DBSCAN

Density-Based Spatial Clustering of Aplications with Noise

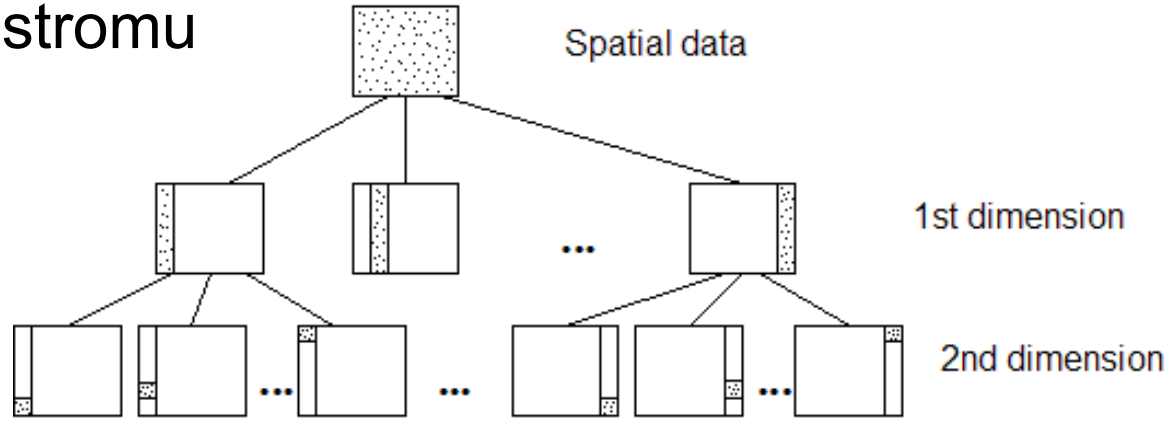
- vychází z hustoty definované pro blízké okolí každého objektu, z dosažitelnosti objektů zjištěné na základě této hustoty a propojenosti dvou objektů ověřené pomocí dosažitelnosti vybraných objektů
- shlukem je neprázdná množina objektů, pro kterou platí následující dvě vlastnosti
  - shluk vždy obsahuje všechny body, jenž jsou dosažitelné z libovolného jeho bodu
  - libovolné dva body ležící v tomtéž shluku musí být propojené
- není založen na vzdálenostech mezi objekty, a tím umožňuje nacházet shluky obecně libovolného tvaru
- nevýhodou je nutnost zadat parametry hustoty.

STING
STatistical INformation Grid

- algoritmus pracuje ve dvou fázích
  - v první fázi rekurzivním způsobem vytvoří mřížkovou strukturu tvořenou několika úrovněmi pravoúhlých buněk. Pro každou buňku jsou vypočítány a uchovány statistické charakteristiky potřebné pro další zpracování, jako je počet objektů, minimální a maximální hodnota obsažená v buňce, aritmetický průměr, směrodatná odchylka a typ pravděpodobnostního rozdělení
  - ve druhé fázi algoritmu probíhá vlastní shlukování, které se provádí nad mřížkou vytvořenou v první fázi
- výhodou algoritmu je především rychlost zpracování, která není závislá na počtu objektů ale pouze na počtu buněk mřížkové struktury, tím je algoritmus vhodný i pro shlukování dat ve velkých souborech
- další výhodou je možnost nalezení shluků různých tvarů

SCAHIPAT
Spatial Clustering Algorithm Based on Hierarchical-Partition Tree

- vychází z technologií prostorového indexování (quad tree)
- 1.fáze = vytvoření H-P stromu



- 2.fáze = výpočet statistik (hustota, hranice) pro každou podmnožinu
- 3.fáze = spojování vhodných podmnožin
- 4.fáze = vyřazení objektů, které jsou podezřelé z odlehlosti

Shrnutí

- Metody prostorového shlukování se snaží klasifikovat objekty podle míry podobnosti jejich znaků a současně respektovat požadovaná prostorová omezení.
- Nové algoritmy jsou převážně obměnou, resp. kombinací stávajících algoritmů.
- Autoři často „šijí na míru“ nový algoritmus na svá data.

Děkuji za pozornost