Databázové systémy 2

19 Metody dolování znalostí (update)

Dolováním znalostí nazýváme proces netriviálního získávání implicitní, dříve neznámé a potenciálně užitečné informace z dat.

Znalosti získáváme z numerických dat a z dat nenumerických – multimediálních (obrázky, hudba, video). Metody dolování znalostí dat jsou založena na metodách numerických, proto musejí být data nenumerická nejdříve na data numerická převedena.

Předzpracování dat před dolováním znalostí

- **Filtrace a integrace dat** – výběr atributů vhodných pro analýzu, ošetření nebo vyloučení dat chybných, chybějících, redundandních, irelevantních, konstantních; sjednocení formátů, měrných jednotek; numerické zakódování některých dat, sjednocení kódování; kategorizace a dichotomizace dat
- **Transformace -** standardizace atributů (odstranění závislosti reálných atributů na jednotkách měření), normalizace objektů (odstranění závislosti na velikosti objektu), hlavní komponenty
- **Odvozování** - odvozené atributy • agregované údaje

Metody dolování znalostí

- **Asociace**
- **Shlukování**
- **Rozhodovací stromy**

Asociace

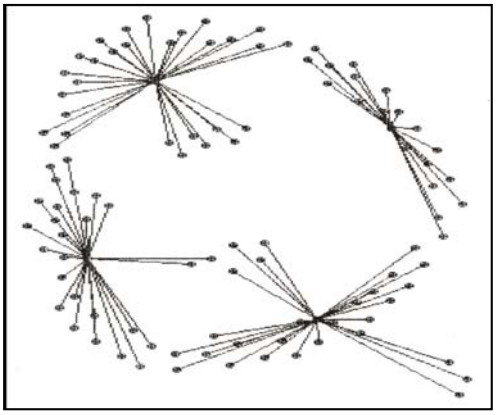
- Klasické – mezi dvěma podmnožinami atributů
- Transakční – v rámci množiny atributů
- Agregované – mezi podmnožinou atributů a jejich charakteristikami

Algoritmy generující asociace

- Triviální
- Uspořádané generování pravidel
- Vzorkování
- K-množiny
- …

Transakční asociace jsou zpracovány podrobně v předcházející přednášce – analýza nákupního košíku.

Shluková analýza



- analýzuje, zda se množina objektů přirozeně rozpadá na výrazné podmnožiny (shluky) objektů vzájemně si podobných a přitom nepodobných objektům podmnožin ostatních
- případně dále analyzuje, zda existuje celá hierarchie takových rozkladů
- pokud shluky existují, čím jsou charakteristické
- jak se případné další objekty zařadí do již definovaných shluků

Shluková analýza netvoří ucelenou teorii, ale je to řada metod založených na různých principech (různorodost řešených problémů, požadovaných typů výsledků, velká data, neurčitost definice shluku)

Metody dle cíle shlukování

- hierarchické – produkující hierarchii rozkladů, kde každý rozklad je zjemněním předcházejícího
- nehierarchické – produkující prostý rozklad objektů na podmnožiny

Metody dle typu výsledných shluků

- shluky kulové, body soustředěny kolem svého těžiště
- shluky obecné tvoří souvislé husté oblasti nejrůznějších tvarů

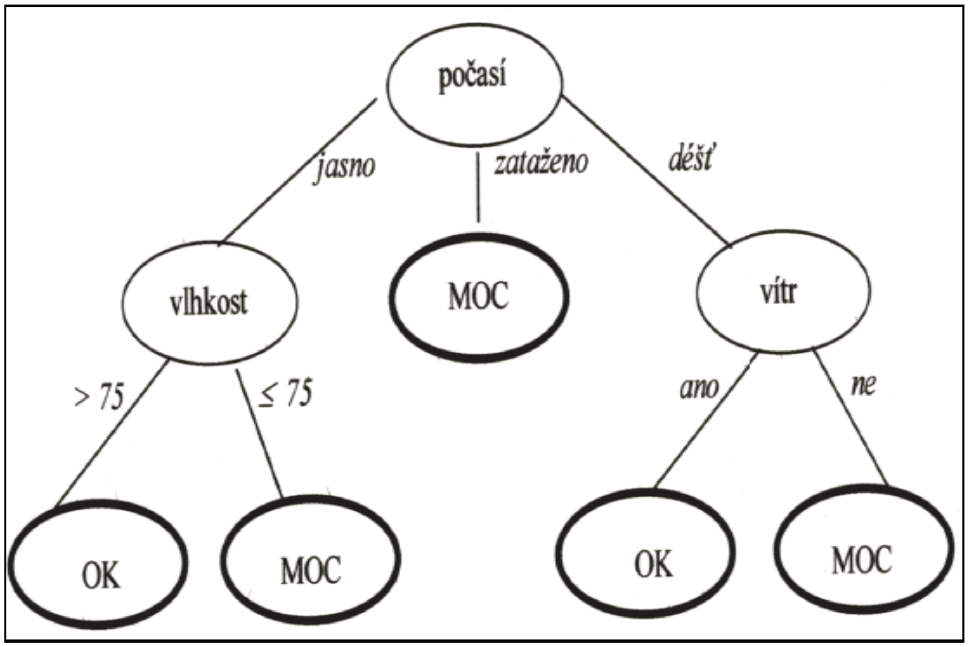
Metody dle typu rozkladu

- shluky disjunktní
- shluky překrývající se

Algoritmy

- nehierarchické (optimalizační k-středové, analýzy módů, fuzzy k-středové, neuronové sítě)
- hierarchické (aglomerativní, divizivní)
- vzorkování

Rozhodovací stromy



Poznámky :

metoda k-středové = k-means

další metody uvedené v původní verzi je možné zmínit jen informačně