MODUL OF CLUSTER ANALYSIS

Marek MACH, Master Degree Programme (5)
Dept. of Information Systems, FIT, VUT
E-mail: xmachm01@stud.fit.vutbr.cz

Supervised by: Dr. Jaroslav Zendulka

ABSTRACT

The Modul of the cluster analysis is a part of a data mining system, which is developed on the Faculty of Information Technologies. This modul communicates with a core of system by the way of a DMSL document, which was created for these puproses. The DMSL language is the XML-based language. This paper presents implementation of the cluster analysis by the method called k-means.

1 ÚVOD

Modul pro shlukovou analýzu je součástí systému pro dolování dat. Základem systému je databáze, která poskytuje data modulu pro předzpracování a transformaci dat. Tento modul vytváří DMSL dokument, který určuje mimo jiné zdroj dat a jejich transformace. S tímto modulem komunikují všechny ostatní moduly, které jsou součástí systému. Jedná se o moduly pro shlukovou analýzu, deskriptivní dolování, klasifikaci a pro dolování asociačních pravidel[4].

Komunikace mezi moduly probíhá prostřednictvím dat uložených ve formě DMSL dokumentu. DMSL dokument je dokument vytvořený v jazyce DMSL, speciálně navrženém pro tento systém. Jedná se o jazyk založený na bázi jazyka XML.

Tento článek je zabývá problematikou implementace shlukové analýzy k-středovou metodou, jejíž výhodou je nízká výpočetní náročnost vzhledem k ostatním užívaným metodám.

2 SHLUKOVÁ ANALÝZA

Shluková analýza je statistická metoda pro analýzu dat. Využívá se v oblasti dobývání znalostí z databází. Jejím cílem je nalezení skupin, tzv. shluků, navzájem si podobných dat[1]. Se shlukem podobných dat se dá následně pracovat kolektivně jako s jednou skupinou.

2.1 VZDÁLENOSTI STŘEDŮ SHLUKŮ

Shluková analýza vychází ze znalosti vzdálenosti mezi jednotlivými objekty. Pokud budeme uvažovat, že všechny objekty jsou charakterizovány *n* numerickými veličinami, pak můžeme vzdálenost mezi dvěma objekty $x_1 = [x_{11,}...,x_{1n}]$ a $x_2 = [x_{21,}...,x_{2n}]$ vyjádřit různými metrikami, mezi které patří [1]:

- euklidovská vzdálenost

$$d_E(x_1, x_2) = \sqrt{\sum_{j=1}^{n} (x_{1j} - x_{2j})^2},$$

- hammingova vzdálenost

$$d_H(x_1, x_2) = \sum_{j=1}^{n} |x_{1j} - x_{2j}|,$$

3 METODA K-STŘEDŮ

K-středová metoda (k-means) se snaží rozdělit objekty do zadaného počtu shluků tak, aby byla minimalizována variabilita uvnitř shluků a maximalizována mezi shluky.

Vzdálenost středů jednotlivých shluků je získávána výpočtem jejich euklidovské vzdálenosti. Tato metrika není vhodná pro všechny typy dat a to je nutné mít na paměti při využívání metody k-středů.

Tato metoda je vhodná v případech, kdy je počet shluků *k* předem znám.

3.1 POPIS ALGORITMU

Algoritmus metody k-středů lze popsat následujícími kroky:

1) Rozdělíme množinu objektů do *k* počátečních shluků. K tomu lze použít náhodné rozdělení nebo jinou shlukovací metodu.
2) Spočteme středy jednotlivých shluků.
3) Pro každý objekt spočteme jeho vzdálenost od středů ostatních shluků. Pokud se nachází některý střed k danému objektu blíže, než střed v jehož shluku se objekt nachází, tak přesuneme objekt do shluku s nejbližším středem.
4) Jestliže došlo v bodě 3 alespoň k jednomu přesunu, tak pokračujeme v algoritmu od bodu 2.

Průběh shlukování a jeho výsledek je závislý na počátečním rozdělení shluků. Algoritmus lze modifikovat přepočítáním středů po každém přesunu objektu do jiného shluku. Následkem této úpravy se stane průběh shlukování závislí i na pořadí, v jakém jsou jednotlivé objekty probírány[3].

3.2 POLOFORMÁLNÍ ZÁPIS ALGORITMU K-STŘEDŮ[1]

1. náhodně se zvolí rozklad do k shluků
2. určí se středy shluků pro všechny shluky v aktuálním rozkladu
3. for každý objekt x do
   - 3.1 urči vzdálenost $d(x,c_k)$, $k=1,\ldots,k$; c_k je střed k-tého shluku
   - 3.2. nastav $d(x,c_k)= \min_k d(x,c_k)$
   - 3.3. if x nepatří do shluku l (k jehož středu c_l má x nejblíže)
     then přesuň x do shluku l
4. if proběhl přesun
   then zpět na krok 2
   else konec

4 ZÁVĚR

Navržený modul shlukové analýzy komunikuje s ostatními částmi systému prostřednictvím DMSL dokumentu. K vytváření shluků podobných objektů využívá algoritmu k-středů.

Při aplikaci tohoto algoritmu vycházíme z předpokladu, že víme do kolika shluků je možné objekty rozdělit. Následkem této úvahy je konstantní počet shluků při výpočtu, během kterého se mění pouze přiřazení objektů k těmto shlukům. Z tohoto důvodu je tato metoda méně výpočetně náročná než jiné, například hierarchické. Velikost výpočetní náročnosti je významná z důvodu aplikace algoritmu na velké datové soubory. Výsledné shluky jsou reprezentovány svými středy, což je výhodné z hlediska zařazování nových objektů. Ty jsou zařazeny ke shluku, k jehož středu mají nejblíže.

LITERATURA

[1] Berka, P.: Dobývání znalostí z databází. Praha: ACADEMIA, 2003, s. 55-59, ISBN 80-200-1062-9.

[2] Han, J., Kamber M.: Data mining: Concepts and Techniques. San Francisco, Morgan Kaufman Publishers, 2001, s. 335-393. ISBN 1-55860-489-8.

[3] Reif, J.: Metody matematické statistiky. Plzeň: Západočeská univerzita, 2000. 286s. ISBN 80-7082-593-6.

[4] Ševčíková, L.: Modul deskriptivního dolování a klasifikace. Brno, 2002. s. 35-38. Diplomová práce na Fakultě informatiky Vysokého učení technického v Brně. Vedoucí diplomové práce Jaroslav Zendulka.