MINING QUANTITATIVE ASSOCIATION RULES IN LARGE RELATION TABLES

Pavel JURKA, Master Degree Programme (5)
Dept. of Information Systems, FIT, BUT
E-mail: xjurka01@stud.fit.vutbr.cz

Supervised by: Ing. Vladimír Bartík

ABSTRACT

This work is engaged in a technique for association rule mining from relational databases which are most common for data storage nowadays. The described method works with quantitative attributes as well as with categorical data.

1 ÚVOD

V dnešní době je získáváno čím dál více dat, která jsou ukládána v relačních nebo novějších objektových databázích. Tato data často nemají sama o sobě žádnou vypovídající hodnotu. To je důvod, proč se používají metody, které objevují závislosti mezi daty. Ty se poté nejčastěji využívají v obchodním sektoru pro podporu řízení a rozhodování.

Existuje řada metod, které využívají různé techniky získávání znalostí. Nejčastěji se používají metody využívající asociační analýzu, shlukování, klasifikaci a prognózu.

Pro svou snadnou interpretaci a názornost jsou obvykle používána asociační pravidla. Ta byla vytvořena pro použití nad transakčními databázemi, ale protože se nejčastěji používají relační databáze, byly vytvořeny metody umožňující jejich využití i nad relačními daty.

2 ZÍSKÁVÁNÍ ASOCIAČNÍCH PRAVIDEL NAD RELAČNÍMI DATABÁZEMI

2.1 ASOCIAČNÍ PRAVIDLA V TRANSAKČNÍCH DATECH

Máme-li transakční databázi, pak cílem je získat asociační pravidla ve tvaru implikace A=>B, kde A i B jsou množiny položek z množiny atributů $I = I_1, I_2, \ldots, I_m$. Množiny A a B musí být disjunktní.

Pravidla mají vlastnosti:

- **Podpora (support):** asociační pravidlo $A \Rightarrow B$ má podporu s, jestliže (s*100)% transakcí v databázi odpovídá sjednocení množin A a B. Jde tedy o frekvenci výskytu pravidla v databázi. Podpora tedy může nabývat hodnot z intervalu <0,1>.

- **Spolehlivost (confidence):** pravidlo A$\Rightarrow$B má spolehlivost c, jestliže (c*100)% transakcí v databázi, které odpovídají množině A, také odpovídají množině B. Tento parametr udává sílu implikace v asociačním pravidle. Spolehlivost může tedy také nabývat hodnot z intervalu <0,1>.

$$\text{conf} = \frac{\text{supp}(A \cup B)}{\text{supp}(A)} \qquad 3.1$$

2.2 RELAČNÍ DATABÁZE

Relační databáze může obsahovat dva typy atributů: kategorické a kvantitativní. Kategorické atributy (např. město, značka auta apod.) mají doménu obsahující konečný počet hodnot, kterých může daný atribut nabýt. Nad těmito hodnotami není definována žádná relace uspořádání. Naproti tomu kvantitativní (numerické) atributy (např. věk, cena apod.) mají doménu obsahující nekonečný počet hodnot a zpravidla je nad těmito hodnotami definována relace uspořádání.

3 METODY PRO ZÍSKÁVÁNÍ KVANTITATIVNÍCH ASOCIAČNÍCH PRAVIDEL

Vybraná metoda byla popsána v [1]. Základem metody je analýza atributů, rozdělení jejich hodnot na intervaly a následné zpracování generujícím algoritmem.

3.1 URČENÍ POČTU INTERVALŮ PRO KAŽDÝ KVANTITATIVNÍ ATRIBUT.

Pro určení počtu intervalů použijme K-metriku, která je zadána uživatelem a udává ztrátu informace při rozdělení na intervaly, kdy n je počet kvantitativních hodnot, m je minimální podpora a K>1 hodnota udávající tolerovanou ztrátu informace.

$$NumberofIntervals = \frac{2 \times n}{m \times (K-1)} \qquad 3.2$$

Metrika určuje na kolik intervalů, nebo zda vůbec, budou hodnoty sledovaného atributu rozděleny.

3.2 MAPOVÁNÍ ATRIBUTŮ NA ČÍSLA

- Kategorické atributy – hodnoty jsou přímo mapovány na celá čísla
- Kvantitativní atributy bez intervalů – obsahují málo hodnot, které mohou být také mapována přímo na celá čísla
- Kvantitativní atributy s intervaly – intervalům jsou přiřazeny celá čísla a poté jsou hodnoty atributů rozděleny do jednotlivých intervalů

3.3 VÝPOČET PODPORY INTERVALŮ, URČENÍ FREKVENTOVANÝCH MNOŽIN KATEGORICKÝCH ATRIBUTŮ.

V tomto kroku jsou slučovány jednotlivé sousední intervaly, dokud nedosáhnou maximální podpory a je spočítána podpora kategorických a kvantitativních atributů. Tak jsou

nalezeny frekventované položky, ze kterých jsou dále vypočítávány všechny frekventované množiny mající minimální podporu. Ty jsou využity v dalším kroku.

3.4 GENEROVÁNÍ PRAVIDEL

Na základě vygenerovaných frekventovaných množin, které mají minimální podporu, jsou vytvářena asociační pravidla. Pro každé pravidlo je spočítána podpora a spolehlivost, které určí, zda pravidlo splnilo minimální hodnoty a bude-li ponecháno.

3.5 URČENÍ ZAJÍMAVÝCH PRAVIDEL.

Při kombinování intervalů pro kvantitativní atributy může nastat taková situace, že vznikne příliš mnoho pravidel, která jsou si podobná. Proto zavedeme hodnotu R, pomocí které uživatel určí, o jak moc zajímavá pravidla má zájem. Definujme I_R jako množinu trojic <x,l,v> charakterizující jednotlivé atributy a jejich intervaly. A nechť X je frekventovaná množina, a $\hat{X}$ je generalizace X (a X je specializace $\hat{X}$) pokud *atributy(X) = atributy($\hat{X}$)* a všechny intervaly v X jsou podmnožinami intervalů v $\hat{X}$. Potom X je R-zajímavá s ohledem k $\hat{X}$, pokud je podpora X větší nebo rovna R-násobku očekávané podpory založené na $\hat{X}$ a pro jakoukoliv specializaci X‘, kde X‘ má minimální podporu a X-X‘ $\subseteq I_R$, X-X‘ je R-zajímavá vzhledem ke $\hat{X}$. Stejně tak pravidlo X=>Z je R-zajímavé vzhledem k pravidlu $\hat{X} \Rightarrow \hat{Y}$, pokud je podpora pravidla X=>Y R-násobkem očekávané podpory $\hat{X} \Rightarrow \hat{Y}$ nebo spolehlivost je R-násobkem spolehlivosti $\hat{X} \Rightarrow \hat{Y}$ a frekventovaná množina X$\cup$Y je R-zajímavá vzhledem k $\hat{X} \cup \hat{Y}$

4 IMPLEMENTACE

Diplomová práce navazuje na jiný projekt vyvíjený na fakultě FIT, zabývajícím se předzpracováním relačních dat podle zvolených kritérií [3]. Tento projekt definuje rozhraní, kterým komunikuje s dolovací metodou. Systém používá k definici úlohy jazyk DMSL, který je založen na XML standardu. Úkolem implementované úlohy je převzetí definice dat a parametrů pro dolování ze vstupního DMSL dokumentu a poskytnout grafické rozhraní pro jejich definici. Tyto data zanalyzuje a určí typu atributů a jejich intervalů. Poté provede výpočet frekventovaných množin a vlastní generování asociačních pravidel. Výsledná pravidla jsou uložena zpět do souboru DSML, který slouží k přenosu dat zpět do původního modulu, který provede zobrazení získaných informací. Výhodou tohoto postupu je možnost nahradit jakýkoliv modul jiným, který podporuje komunikaci pomocí standardu DMSL. Celý systém je naprogramován v jazyce Java a využívá pro ukládání dat relační databázi se kterou komunikuje pomocí univerzálního rozhraní JDBC.

LITERATURA

[1] Srikant, R., Agrawal, R.: Mining Quantitative Association Rules In Large Relational Tables, Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, 1996

[2] Bartík, V.: Získávání asociačních pravidel z relačních dat, FIT VUT 2003

[3] Hromčík, P.: Systém pro získávání znalostí z databází, FIT VUT 2003