DATA PREPROCESSING FOR DATA MINING SYSTEM

Michal Šebek
Master Degree Programme (2), FIT BUT
E-mail: xsebek00@stud.fit.vutbr.cz

Supervised by: Roman Lukáš
E-mail: lukas@fit.vutbr.cz

ABSTRACT

Databases grow up by new data continually. New hidden information can be formed in these databases. It is desirable to explore and process them. Because of this process called Knowledge Discovery in Databases has been defined and new complex systems has been developed for its support. These systems are formed by a lot of subsystems such as data preprocessing, data mining modules and results reporting. Main goal of this report is analysis and design of data preprocessing subsystem.

1 ÚVOD

V současné době se potýkáme s velkým nárůstem objemu dat, která je nezbytné v databázích uchovávat. Nabízí se pak otázka, jestli nelze tato uchovávaná data ještě nějak využít. Za tímto účelem vznikl proces ***získávání znalostí z databází***. Proces se zabývá získáváním nových užitečných informací, které se mohou v uložených datech vyskytovat. Typickým příkladem může být analýza nákupního košíku, kdy se z transakční databáze pokladen snažíme získat informace o často kupovaných kombinacích zboží.

Jak již bylo řečeno, získávání znalostí není atomická operace, ale jde o ucelený proces. Standardně se sestává z kroků předzpracování dat, samotného dolování z dat, vyhodnocení vzorů a prezentace získaných znalostí. Jelikož proces je značně komplexní a pracuje s velkými objemy dat, je snaha mít pro proces softwarový nástroj, který by analytikovi poskytl co největší možnou podporu. A právě analýza a návrh řešení předzpracování dat pro takovýto softwarový systém vyvíjený na FIT VUT v Brně je náplní tohoto článku.

2 PŘEDZPRACOVÁNÍ DAT V PROCESU DOLOVÁNÍ Z DAT

Data, která se ve zkoumaných databázích vyskytují, jsou však ve většině případů k přímé aplikaci dolovacích algoritmů nepoužitelná. Data často vlivem chyb při zadávání obsahují nesmyslné hodnoty, některé hodnoty chybí zcela, data jsou příliš objemná a atributy obsahují k analýze nevhodné hodnoty. Zmíněné problémy mohou zkreslit nebo zcela znehodnotit výstup některých dolovacích algoritmů, proto je nezbytné data vhodným způsobem upravit tak, aby proces dolování poskytl co nejlepší výsledky.

Předzpracování dat se děje v několika krocích. Ty jsou definovány poměrně komplexně a přesahují potřeby aplikace (např. kroky pro datové sklady), proto se zaměřuji spíše na rozbor kroků, které budou následně použity v návrhu systému. Podrobně o tématu pojednává [1].

2.1 ZJIŠTĚNÍ STATISTIK SOUBORU DAT

Nejprve je nutné data analyzovat. Pro popis středu souboru hodnot atributu se užívají základní míry jako *průměr*, *modus* a *medián*. Variaci hodnot pak lze vyjádřit pomocí *rozptylu* a *směrodatné odchylky*. Kvalitní přehled o rozložení hodnot atributu nám poskytují *kvantily*. Ty jsou definovány jako body hranic pravidelných intervalů kumulativní distribuční funkce náhodné veličiny. Dle počtu hranic definujeme nejčastěji používané 4-kvantily jako *kvartily* a 10-kvantily jako *decily*. Pomocí hodnot kvartilů Q_1 a Q_3 definujeme *mezikvartilovou vzdálenost* (IQR) jako $IQR = Q_3 - Q_1$. Více k definicím zmíněných veličin lze nalézt v literatuře [1].

2.2 ČIŠTĚNÍ DAT

Tento krok probíhá ve dvou fázích. V první je třeba **odstranit chybějící hodnoty**. Toto lze realizovat více způsoby. Nejjednodušší automatizovaný způsob je *ignorovat celou n-tici*. Tímto se však ztrácí cenná data k analýze. Proto se chybějící hodnota spíše nahrazuje jinou hodnotou. Tu lze například vyjádřit z ostatních n-tic jako *průměrnou hodnotou atributu všech záznamů či pouze náležejících dané třídě*, do které n-tice náleží.

Druhá fáze spočívá ve **vyhlazení zašuměných dat**. Za šum lze považovat hodnotu, která je nesmyslná nebo jde o hodnotu, která je od ostatních značně vzdálená - tzv. odlehlá hodnota. K vyhlazení se užívají techniky *regrese* (závislost jednoho atributu na druhém), *shlukování* (odlehlé hodnoty nenáleží k žádnému shluku) nebo *plnění*. V tomto případě jsou seřazená data rovnoměrně rozdělena do *košů* a hodnoty v koši vyhlazeny vybranou mírou středu.

2.3 TRANSFORMACE A REDUKCE DAT

Některé metody dolování (např. shlukování) mohou do značné míry ovlivnit řádově odlišné rozsahy hodnot atributů použitých k dolování. Aby metody pracovaly správně, je nutné rozsahy numerických atributů normalizovat na podobné rozsahy (nejčastěji interval hodnot 0-1). Zde se užívá různých metod **normalizace**. Přirozenou transformací atributu s předem známým rozsahem je *min-max normalizace*. Pro nedefinované rozsahy lze užít normalizaci *z-score*.

Celý proces dolování probíhá nad velkým objemem dat. Urychlení výpočtu je možné dosáhnout vhodnou transformací a redukcí počtu dat. Transformace může spočívat v **konstrukci nových atributů**, které budou agregovat či jinak vycházet z původních atributů. Jiný přístup k optimalizaci hodnot v databázi je **diskretizace** numerických atributů, která původní atribut převede na kategorický tak, že hodnoty atributu zařadí do malého počtu definovaných intervalů. Přirozeně lze množství dat redukovat ještě tak, že z původního souboru jsou vybrány pouze některé n-tice a to buď náhodným, nebo stratifikovaným *vzorkováním*.

3 NÁVRH ŘEŠENÍ PODPORY PŘEDZPRACOVÁNÍ DAT

Kvalitní podpora předzpracování dat je nezbytná pro každý systém dolování z dat. Při návrhu rozšíření v oblasti předzpracování pro stávající systém je nutné vycházet ze stávající koncepce

podrobně popsané v [2]. Systém je navržený tak, že dolovací proces uživatel definuje spojováním uzlů grafu. Každý uzel reprezentuje konkrétní komponentu systému s možností definovat akci nad vstupními daty do uzlu. S ohledem na toto je třeba problematiku předzpracování dat nejprve logicky rozčlenit do jednotlivých grafových komponent a jim definovat příslušnou funkčnost (systém některé komponenty již částečně definuje, ale implementace není kompletní).

- **Komponenta výběru dat** - je základní komponentou pro načtení dat do dolovacího procesu. Komponenta by měla jak umožnit načtení dat ze stávajících tabulek na serveru, tak i umožnit import z některých datových souborů (např. formátu CSV).
- **Komponenta VIMEO funkcí** - je v systému již částečně implementována. Vychází ze zobecněného principu zpracování VIMEO (Valid, Invalid, Missing, Empty, Outlier) hodnot atributů. Ve výsledku je nezbytné, aby poskytovala možnosti náhrad hodnot zadanou metodou.
- **Komponenta transformací dat** - poskytuje v současné době základní metody transformace atributů v oblasti konstrukce atributů a normalizace. Následně je nezbytné dodefinovat především proces diskretizace hodnot.
- **Komponenta náhledu dat (Insight)** - bude uživateli poskytovat podporu pro statistickou analýzu souboru dat. Funkčnost bude řešena podle analýzy v části 2.1. Kromě numerického popisu dat bude komponenta realizovat vykreslování statistických grafů (krabicové, kvantilové, bodové grafy a histogramy).
- **Komponenta redukce dat** - bude poskytovat metody pro snížení počtu n-tic vstupujících do dalších kroků dolovacího procesu. Jelikož některé klasifikační a predikční algoritmy vyžadují více množin dat (pro učení a pro testování), je třeba aby komponenta umožňovala rozdělení dat zvoleným způsobem do těchto množin.

Pro každou z těchto komponent budou řešeny optimalizační problémy, neboť množství zpracovávaných dat je značné. V případě složitějších výpočtů nad celým souborem dat je nezbytné zvážit přenesení části výpočtu na databázový server v podobě uložených procedur PL/SQL v případě serveru Oracle a v maximální míře využívat možností jazyka SQL pro efektivní zpracování dotazů.

4 ZÁVĚR

Článek čtenáři představuje současné metody předzpracování dat procesu získávání znalostí z databází především v oblasti čištění, tranformace a redukce dat. Z tohoto rozboru následně vychází návrh rozšíření vyvíjeného systému pro dolování v podobě specifikace funkčnosti komponent systému.

REFERENCE

[1] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Elsevier Inc., second edition, 2006, 770s., ISBN 978-1-55860-901-3

[2] Krásný, M.: Systém pro dolování z dat v prostředí Oracle, Diplomová práce, FIT VUT v Brně, Brno, 2008