CLUSTER ANALYSIS MODULE OF DATA MINING SYSTEM

Martin Hlosta
Master Degree Programme (2), FIT BUT
E-mail: xhlost00@stud.fit.vutbr.cz

Supervised by: Jaroslav Zendulka
E-mail: zendulka@fit.vutbr.cz

ABSTRACT

Knowledge discovery in databases is a process, that deals with the problem of enormously growing amount of data in databases. One phase of this process is data mining. Clustering is one of data mining task types, which can be executed on preprocessed data. The goal of this paper is to describe analysis and design of cluster analysis module for a data mining system currently being developed at Faculty of Information Technologies.

1 ÚVOD

S rostoucím množstvím dat, která jsou uchovávána v různých databázích roste i potřeba se v těchto datech vyznat a nalézat v nich předem neznámé, zajímavé, potenciálně užitečné znalosti. Tento proces se nazývá **získávání znalostí z databází** a skládá se z několika fází, a to předzpracování dat, dolování z dat, vyhodnocení a prezentace výsledných znalostí.

Z hlediska tohoto příspěvku je důležitá fáze dolování z dat, která někdy dává jméno celému procesu. Na předzpracovaný datech, tj. vyčištěných, integrovaných, vhodně transformovaných, je provedena dolovací úloha. Existuje více typů takových dolovacích úloh. Mezi nejznámější patří asociační analýza (např. analýza nákupního košíku), klasifikace (např. detekce podvodného jednání), predikce (např. předpověď vývoje teploty počasí) a shluková analýza. Rozšíření momentálně vyíjeného dolovacího systému na FIT o modul shlukové analýzy je předmětem tohoto příspěvku.

2 SHLUKOVÁ ANALÝZA

Shlukování je jednou z dolovacích úloh, která je založená na učení bez učitele. Data se snažíme na základě podobnosti zařazovat do předem neznámých tříd, které nazýváme **shluky**. To je rozdíl oproti klasifikaci, kde tyto třídy známe. Existuje mnoho způsobů jak takovéto shluky nalézt a metody, které při jejich hledání používáme. Na tyto metody klademe také několik podmínek, např. škálovatelnost, tvorba shluků různých tvarů apod.

Velké množství existujících algoritmů lze rozdělit do několika kategorií. Jedná se o metody založené na rozdělování, hustotě, mřížce, modelu, hierarchické metody a popřípadě některé další metody. Některé algoritmy lze současně zařadit do více kategorií.

3 SHLUKOVACÍ METODY ZALOŽENÉ NA HUSTOTĚ

Pro rozšíření byly vybrány tři algoritmy z metod založených na hustotě. Tyto algoritmy přidávají datové objekty do shluků, pokud je hustota dat větší než určitý práh. Jejich hlavní výhodou oproti ostatním metodám je schopnost nalézt shluky obecně libovolného tvaru. V praxi se tyto algoritmy používají např. při shlukování webových sezení. Systém bude rozšířen o algoritmy DBSCAN, OPTICS a DENCLUE. Kromě následujícího popisu lze detailnější podrobnosti najít např. v [1].

3.1 DBSCAN

Principem algoritmu je postupné přidávání bodů s dostatečně velkou hustotou do shluků. Prohledává se okolí bodů, a pokud obsahuje větší než určitý počet bodů, pak je vytvořen nový shluk a daný bod prohlášen jako jádro tohoto shluku.

3.2 OPTICS

Algoritmus je založen na obdobném prinipu jako DBSCAN, ale řeší jeho problém hledání shluků s různou hustotou. Vytváří pomocnou strukturu se seřazenými vstupními daty tak, že body, které jsou si nejblíž, jsou v této struktuře sousedy. V tomto pořadí se pak body při zpracování procházejí.

3.3 DENCLUE

Narozdíl od předchozích se využívá distribučních funkcí k matematickému popisu shluků i různého tvaru, konkrétně zachycení vlivu bodů na ostatní objekty. Je vhodný pro data s vysokou dimenzionalitou, která jsou hodně zašumělá. Autoři rovněž uvádějí větší výkonnost oproti algoritmu DBSCAN.

4 VYVÍJENÝ SYSTÉM

Na FIT se několik let vyvíjí systém pro získávání znalostí z databází, který je primárně určený pro podporu výuky. Architektura systému je typu klient/server. Serverová část uchovává data pro dolování a také poskytuje některé funkce pro předzpracování a dolování - v současnosti podpora dolování ODM serveru Oracle. Klientská část je napsána v jazyce Java a má modulární strukturu. Jádro obstarává grafické uživatelské rozhraní a předzpracování dat a k němu se vyvíjí moduly, které implementují dolovací úlohy. Celá dolovací úloha je pak sdílena mezi moduly v XML dokumentu v jazyce DMSL, který byl vyvinut na FIT a umožňuje tak přenositelnost se systémy, které by jej také podporovaly. Jazyk je detailně popsán v [2].

Je také důležité zmínit, že dolovací úloha se v uživatelském rozhraní tvoří spojováním dostupných komponent v orientovaném grafu - od předzpracování, přes dolování až po vizualizaci výsledků.

Stávající implementace je výsledkem úprav jádra Ing.Šebka, které jsou popsány v [3], a několika dalších absolventů, kteří jsou autory dolovacích modulů. V současné době celkem osm studentů v rámci řešení bakalářských a diplomových projektů pracuje jak na rozšíření funkcionality jádra, tak na přídavných dolovacích modulech.

5 KONCEPCE ŘEŠENÍ

Do systému bude přidán nový modul obstarávající shlukovou analýzu. Bude implementovat rozhraní *MiningPiece*, které představuje komponentu, kterou lze vložit do palety v procesu dolování. Jak bylo zmíněno, existuje mnoho shlukovacích algoritmů. Bude tedy vytvořeno společné rozhraní, které bude pokrývat společné vlastnosti všech shlukovacích algoritmů. To se týká jak pro zadávání vstupních parametrů algoritmů, tak i modelu, který bude výsledkem shlukování včetně vizualizace. Každý algoritmus pak implementuje toto rozhraní a přidá specifické vlastnosti. U metod založených na hustotě to jsou dva vstupní parametry pro úlohu. Na výstupu půjde např. o vizualizaci pomocné struktury algoritmu OPTICS. Společné vlastnosti algoritmů budou zohledněny i v definici schématu DMSL dokumentu jak pro uzly zadávání dolovací úlohy, tak i reprezentaci znalosti. Na řešení modulu se společně se mnou podílí P. Riedl, který řeší rozšíření o jiné shlukovací algoritmy, vizualizaci a validaci výsledných shluků.

Všechny doposud vytvořené dolovací moduly využívaly k dolování podporu, kterou poskytuje server Oracle - ODM (Oracle Data Mining). Mnou přidané algoritmy založené na hustotě v této podpoře obsaženy nejsou. Do budoucna se počítá především o rozšíření o další algoritmy neobsažené v ODM a použití dalších databázových serverů, popřípadě jiných uložišť dat. Z hlediska přenositelnosti je ale vhodné zachovat již použité rozhraní algoritmů, konkrétně dle standardu JDM 1.0 (Java Data Mining) vyvinutého pod JSR-73. Dnes už existuje i verze 2.0 pod JSR-241, ale z hlediska shlukování nepřináší nic důležitého. Proto všechny nově přidané algoritmy budou zapouzdřeny tak, aby implementovaly rozhraní *javax.datamining.algorithm* nebo také nastavení parametrů shlukovací úlohy pomocí rozhraní *clusteringSettings*.

Pro algoritmy DBSCAN a OPTICS existujuje implementace v dolovacím systému Weka, který je rovněž napsán v jazyce Java. Není potřeba je proto implementovat celé a lze dostupné kódy využít pro integraci do systému. Naopak algoritmus DENCLUE bude potřeba implementovat s pomocí popisu autorů a popřípadě jedné existující implementace v jazyce C++.

6 ZÁVĚR

Tento příspěvek představil analýzu a návrh koncepce řešení při tvorbě modulu pro shlukovou analýzu systému pro dolování z dat, vyvíjeného na FIT. Vznikl jako součást mé diplomové práce, kde bude tento návrh dále implementován, integrován do systému a důkladně otestován na vhodném vzorku dat.

Poděkování: Tato práce vznikla částečně za podpory grantu VUT FIT, FIT-S-10-2 a specifického výzkumu MSM0021630528.

REFERENCE

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

[2] Kotásek, P.: DMSL: Data Mining Specification Languag. Dizertační práce, FIT VUT v Brně, Brno, 2003

[3] Šebek, M.: Rozšíření funkcionality systému pro dolování z dat na platformě NetBeans. Diplomová práce, FIT VUT v Brně, Brno, 2009