KATEDRA INFORMATIKY
UNIVERZITA PALACKÉHO V OLOMOUCI
www.inf.upol.cz

Oponentský posudek dizertační práce

Radek Janoštík

A new perspective on the Close-by-One algorithm

Práce shrnuje tři výsledky výzkumu autora v oblastech analýzy a získávání znalostí z dat (data mining), využívající aparátu a algoritmů z formální konceptuální analýzy (FCA). Dva z nich se přímo dotýkají známého algoritmu pro výpočet (původně) všech formálních konceptů z objekt-atributových dat, algoritmu Close-by-One (CbO).

První výsledek představuje nový algoritmus LinCbO pro výpočet minimální báze atributových implikací, Duquenne-Guigues báze. V něm je vcelku přímočaře využit algoritmus CbO ve spojení se známým algoritmem LinClosure pro výpočet jistého uzávěru množiny atributů, s využitím kterého jsou vytvořeny implikace báze. Algoritmus LinClosure je zde mírně upravený jednak pro urychlení testu kanonicity vypočteného uzávěru v algoritmu CbO, ale především pro efektivnější výpočet uzávěru množiny, která je sjednocením dříve vypočteného uzávěru a množiny dalších atributů (tímto způsobem jsou uzávěry počítány algoritmem CbO). Toho je dosaženo inovativním využitím počtů atributů počítaných algoritmem LinClosure zjištěných během výpočtu dřívějšího uzávěru pro výpočet nového uzávěru (vylepšení označované jako *attribute counters reuse*).

Očekávané výrazné (pro některá data i několikařádové) zrychlení výpočtu je v práci ukázáno na výsledcích experimentů nad náhodně generovanými i reálnými daty, stejně jako výkonnostní překonání jiných existujících algoritmů (a jiných uzávěrů než LinClosure). Výrazným nedostatkem práce ovšem je, že podstata úpravy algoritmu LinClosure, umožňující toto zrychlení, není v práci objasněna a není ukázána korektnost upraveného algoritmu. K tomu je totiž potřebná znalost principu původního algoritmu LinClosure, který v práci bohužel není vysvětlen, ale pouze popsán. Rovněž není ukázána korektnost celého algoritmu LinCbO.

Proto mám dotaz: Jak slouží počty atributů (tzv. attribute counters) počítané algoritmem LinClosure pro vyhnutí se porovnávání množin atributů a dosažení lineární časové složitosti algoritmu, jak je uvedeno v poznámce 1 na straně 22 práce? Tzn. vysvětlete podstatu algoritmu LinClosure a jeho nové úpravy (attribute counters reuse).

Autor se věnuje také využití, v algoritmu LinCbO, vylepšení základního algoritmu CbO týkajících se ušetření výpočtu uzávěrů, zde označované jako metody prořezávání (výpočetního stromu uzávěrů, pruning methods), známých z algoritmů FCbO a InClose3+; v mírně obecnější podobě, včetně experimentálního porovnání se základním CbO. Zde mi ovšem chybí rozbor jistého rozdílu oproti těmto algoritmům přiblížený v následujícím dotazu.

Druhým výsledkem práce je objasnění a reformulace známého výkonného (jednoho z nejvýkonnějších) algoritmu LCM pro výpočet častých uzavřených množin atributů (frequent closed itemsets) v objekt-atributových transakčních datech v termínech FCA a zejména ukázání toho,

že LCM je ve své podstatě CbO se dvěma optimalizacemi pro efektivnější zpracování řídkých dat a navíc také ušetřením výpočtu uzávěrů na způsob FCbO a InClose3+. Výsledek má svůj význam, neboť osvětluje a rozkrývá komunitě kolem FCA poněkud „tajemný" velice výkonný algoritmus dobře známý lidem z oblasti data mining. Problém, stejně jako u algoritmu LinCbO (resp. LinClosure), ovšem je, že v práci opět není ukázána ekvivalence reformulovaného LCM s CbO a rozebrán rozdíl oproti FCbO a InClose3+ (uvedený již výše), ve smyslu ušetřených výpočtů uzávěrů. Práce také neobsahuje výkonnostní srovnání zmíněných algoritmů.

Zmíněný dotaz k rozdílu algoritmů: Jaký vliv na průběh algoritmu (jaké ušetřené výpočty uzávěrů) má „zapomenutí" posledně přidaného atributu do uzávěru (alias formálního konceptu) jako atributu způsobujícího selhání testu kanonicity, tj. procedura `RemoveRulesByRightSide` v Algoritmech 8 a 10?

Třetím výsledkem práce je pak propojení FCA s logickou analýzou dat (LAD). Je ukázaný vztah mezi základními prvky obou analýz, intenty formálních konceptů v (dvouhodnotové) FCA a vzory generovanými množinou pozorování (spanned patterns) v (binární) LAD, umožňující vzájemné propojení obou disciplín a využití poznatků jedné v druhé (např. v práci zmíněné redukce počtu konceptů/vzorů a zobecnění na nebinární data). V práci je toto demonstrováno využitím algoritmů FCA, konkrétně CbO a FCbO, pro výpočet vzorů v LAD, kde výsledky experimentů ukazují, že algoritmy FCA jsou řádově výkonnější než algoritmy používané v LAD.

Dotaz: V Poznámce 8 na straně 66 (a také v závěru k této části práce) je zdůrazněno, že algoritmy FCA nelze jednoznačně považovat za výkonnější než algoritmy LAD, protože produkují vzory v jiném pořadí. Zřejmě ale jakoukoliv počáteční část celé množiny vzorů spočítají výrazně rychleji. Je tedy pořadí získaných vzorů v LAD důležité?

Tyto, bezesporu přínosné, výsledky jsou publikovány v jednom časopiseckém (druhý aktuálně v recenzním řízení) a jednom konferenčním článku, jejichž je autor práce spoluautorem. Autor je také spoluautorem dalšího publikovaného článku v časopise, jehož předmět již není v práci obsažen. Práce má velmi dobrou formální i jazykovou úroveň, použité i nové algoritmy jsou dobře popsány (včetně srozumitelných pseudokódů), leč nejsou vysvětleny některé jejich stěžejní aspekty a není ukázána jejich korektnost. Přesto se domnívám, že se autorovi podařilo splnit cíl: ukázání dalších možností využití populárního algoritmu CbO a jeho derivátů.

Autor dosažením těchto výsledků prokázal schopnost vědecké práce, včetně práce s literaturou, a tudíž, za podmínky uspokojivého zodpovězení výše uvedených dotazů (hlavně prvních dvou), **doporučuji** práci uznat jako úspěšnou dizertační práci.

Ve Veselíčku dne 2. května 2021
doc. Mgr. Jan Outrata, Ph.D.

Review on the dissertation thesis

A new perspective on the Close-by-One algorithm

by Radek Janostik from Department of Computer Science,
Faculty of Science, Palacky University Olomouc

The thesis of Mr. Radek Janostik consists of five chapters. In Chapter 1 the author gives basic definitions and facts of the research domain. In Chapter 2 a new algorithm for computing the Duquenne-Guigues (also known as canonical or stem) implication basis, called LinCbO, is presented. The algorithm is based on the CbO algorithm where Lin-Closure algorithm computes the closure by reusing the values of counters from the previous calls of the closure. It is shown that the reuse of attribute counters can significantly speed up the performance of LinClosure algorithm. Moreover, the algorithm was extended to different data types and to accounting for background knowledge. Chapter 3 provides a description of the LCM algorithm from the viewpoint of Formal Concept Analysis (FCA). It is demonstrated that LCM is basically CbO algorithm with certain engineering implementation features, which do not change theoretical complexity, but improve the experimental performance. Unveiling unclear features that make LCM so efficient is a very important contribution of the thesis. In Chapter 4 the relationship between two methodologies of data analysis and data mining – Formal Concept Analysis and Logical Analysis of Data (LAD) – is studied. It is shown that patterns of LAD can be interpreted as FCA intents covering some positive and no negative examples. Using this relationship, one can adapt algorithms from FCA for solving problems from LAD. Chapter 5 presents a conclusion of the thesis and possible further research directions.

In general, the thesis provides a deep thorough study of the algorithmic issues of FCA and LAD, comparing existing approaches and proposing new more efficient solutions to well-known problems. The thesis proposes both important theoretical results and comprehensive experimental comparative study of the proposed algorithms and those well-known in the literature. The results are presented in a very clear form, with detailed explanations and numerous examples. This gives the work an additional propedeutical value.

I would make two remarks to the author. The first one concerns the relationship of the pruning techniques mentioned in page 57 (66). What is the relative computational costs of these techniques and what data are most suited for each of these techniques? The second remark concerns the relationship of LAD and FCA. Patterns of LAD which cover positive examples and no negative examples are well studied in FCA as concept-based hypotheses, both in binary setting, and in various non-binary settings in terms of pattern structures, see

Bernhard Ganter, Sergei O. Kuznetsov:
Hypotheses and Version Spaces. *ICCS 2003: 83-95,* **2003**

Sergei O. Kuznetsov:
Pattern Structures for Analyzing Complex Data. *RSFDGrC 2009: 33-44*

In these and related papers both closed and related minimal classifiers (based on minimal generators) were studied. The formalism of pattern structures first described in
Bernhard Ganter, Sergei O. Kuznetsov:
Pattern Structures and Their Projections. *ICCS 2001: 129-142*

allows for patterns of different nature, e.g. based on tuples of numerical intervals (hyperrectangles), which are studied in the thesis.

These remarks do not affect the general value of the presented work. The thesis of Mr. Radek Janostik presents an excellent solid original scientific contribution, which shows that the author - to the reviewer's opinion – deserves the doctoral degree of Palacky University Olomouc.

Moscow, May 5, 2021

Professor Dr. Sergei O. Kuznetsov

Department of Data Analysis and Artificial Intelligence

National Research University Higher School of Economics, Moscow

https://www.hse.ru/en/staff/skuznetsov