Posudek na disertační práci **Markéty Trnečkové**

Factor analysis with ordinal attributes by similarity

Předložená disertační práce se zabývá algoritmy pro rozklady matic obsahujících prvky z nějaké uspořádané škály (např. z množiny $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$). Jde o důležitou metodu analýzy dat umožňující hledaní skrytých informací umožňujících vysvětlit vstupní data. Některé dosažené výsledky již byly publikovány ve sbornících konferencí i časopisecky, publikace dalších se připravuje.

Úvodní kapitola obsahuje základní popis řešených problémů a shrnutí existujících přístupů. V kapitole 2 nás autorka seznamuje se základními pojmy z oblasti fuzzy logiky, rozkladů matic a formální konceptuální analýzy. Kapitola 3 obsahuje teoretické výsledky, na kterých jsou založeny nové algoritmy rozkladu matic. Po kapitole 4 shrnující existují algoritmy pro rozklad matic s prvky z uspořádané škály následuje ústřední kapitola 5 se třemi novými algoritmy: GreEss_L, Asso_L a GreConD_L+. Nejobsáhlejší kapitola 6 přináší experimentální zhodnocení těchto algoritmů na ukázkovém příkladu (plemena psů a jejich vlastnosti), dále na reálných i syntetických datech. Při hodnocení chování navržených algoritmů je brán ohled i na interpretovatelnost dosažených výsledků. Z tohoto hlediska vycházejí algoritmy GreEss_L a GreConD_L+ lépe než Asso_L.

Formální úroveň práce by podle mého názoru mohla být lepší. Angličtina je často negativně ovlivněna českým slovosledem, objevují se překlepy i chyby z nepozornosti (např. definice suprema a infima na str. 7). Našel jsem i formulaci, která by v disertaci být neměla („In the rest of the paper unless otherwise stated...“ na str. 21) či větu začínající „A abundantly discussed the topicnatural option is...“, také na str. 21.

Po matematické stránce je práce podle mého názoru v pořádku. Práci pokládám za přínosnou z teoretického i aplikačního hlediska.

Otázky, komentáře a připomínky:

1. Na str. 9 není uvedeno, zda jsou příklady operací $\otimes$ platné pro interval $[0, 1]$ nebo i pro jeho konečné podmnožiny (např. Goguenova konjunkce).

2. Není mi jasné, co znamená věta „In addition to Theorem 3.“ na konci poznámky 1 (str. 25).

3. Co je F ve vzorci pro funkci, kterou minimalizujeme v metodě NMF (str. 27)?

4. Na str. 58 uvádíte: „Let us note that F_1, F_2, and F_3 obtained using Łukasiewicz operations also have their counterparts among the factors obtained using the Goguen operations.“. Výsledky pro Goguenovy operace však nejsou uvedeny. Bylo by možné uvést více podrobností?

5. Co znamenají čísla v tabulce 6.4 (str. 62)?

6. Na konci kapitoly 6 mi chybí přehledné shrnutí výhod a nevýhod navržených algoritmů a jejich srovnání s existujícími. Je to možné vydedukovat z informací v kapitole obsažených, ale přehlednější forma by podle mého názoru byla užitečná.

7. V kapitole 7 na str. 79 uvádíte několik námětů pro další výzkum v této oblasti (např. výběr vhodné škály, problematika šumu v datech). Máte už nějaké výsledky či ideje pro tyto problémy?

Práci **doporučuji** k obhajobě.

V Ostravě 29. 4. 2017

doc. Ing. Antonín Dvořák, Ph.D.

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

Oponentský posudek dizertační práce

Markéta Trnečková

Factor analysis with ordinal attributes

Předmětem dizertační práce jsou rozklady matic obsahujících hodnoty z uspořádané škály. Problém rozkladu je definovaný jako rozšíření problému rozkladu neboli faktorizace Booleovských matic (Boolean Matrix Factorization, BMF), ve které jsou vstupem matice obsahující pouze nuly a jedničky, na matice obsahující typicky hodnoty z intervalu od nuly do jedné. BMF je významný problém v oblasti dolování dat (data mining), cílem rozkladu je v datech reprezentovaných maticí nalézt skrytou informaci, ve formě tzv. faktorů, pomocí které lze data vysvětlit. Zatímco pro BMF vzniklo několik algoritmů rozkladu navržených speciálně pro binární data a vyznačujících se velmi dobrou interpretovatelností a kvalitou faktorů (z hlediska vysvětlení dat), pro nebinární data jsou běžně dostupné a používané převážně statistické metody rozkladu matic nad reálnými čísly. U nich je ovšem interpretovatelnost faktorů přinejmenším problematická a kvalita faktorů diskutabilní. Autorka v této práci ukazuje, že přímočaré rozšíření algoritmů pro BMF na ordinální data, tj. data reprezentovaná maticí s hodnotami z uspořádané škály, je pro takováto data vhodnější a interpretovatelnost a kvalita faktorů jsou znatelně lepší.

Toto rozšíření spočívá v použití fuzzy logiky místo logiky dvouhodnotové, kdy škálou hodnot je reziduovaný svaz. To přináší několik netriviálních problémů. Například u tzv. esenciálních prvků matice, na kterých je založen jeden z úspěšných BMF algoritmů rozkladu, GreEss. Role prvků matice a vysvětlení dat pomocí faktorů jsou předmětem nových teoretických výsledků práce, nový algoritmus rozšiřující GreEss, označený GreEss_L, je pak hlavním algoritmickým výsledkem práce. A vedle něj ještě další dva, Asso_L a GreCond_L+, pro které také není rozšíření z jejich původních BMF podob zcela triviální (i když sčítání hodnot škály jako reziduovaného svazu není z pohledu fuzzy logiky zcela „košer"). Pro první a poslední uvedený algoritmus práce navazuje na práce kolegů R. Bělohlávka a M. Trnečky. Zajímavý je ovšem také vlastní představený problém úspěšnosti vysvětlení dat pomocí faktorů nalezených ve vybrané menší části těchto dat, včetně navrženého řešení (s využitím esenciálních prvků). Toto řešení vykazuje v experimentech překvapivě dobré výsledky.

Experimentální vyhodnocení úspěšnosti nových algoritmů, na ukázkových datech, na (reálných) datech z praxe i na syntetických náhodně vygenerovaných datech, tvoří druhou část práce. Toto vyhodnocení je provedeno velmi kvalitně, se (správným) zaměřením se nejen na počet získaných faktorů, ale hlavně na jejich kvalitu z hlediska vysvětlení dat a na jejich interpretovatelnost. Nejsou opomenuta ani porovnání rozšířeného algoritmu (existujícího GreCond_L) s jeho původní BMF verzí aplikovanou pro binární data vzniklá z dat ordinálních po naškálování a porovnání se statistickými algoritmy pro matice s reálnými čísly.

Práce má vysokou úroveň formálního zpracování, zahrnující korektní matematiku, srozumitelné pseudokódy algoritmů, názorné ilustrace faktorizací. Drobné výhrady, které mírně snižují celkovou kvalitu práce, mám k jazyku a obsahu. Úroveň angličtiny je kolísavá, zatímco například

v částech 3 a 5 je dobrá, v částech 4 a 6 už je horší. Po obsahové stránce mi hlavně v teoretické části 3 často chyběly příklady a příklady ve stěžejní algoritmické části 5 mohly být více rozvedeny (příklad 8 je neúplný). Dále pak používané fuzzy množiny a jejich notace nejsou zavedeny.

K novým algoritmům a experimentům mám následující dotazy:

1. Rozšířené algoritmy pro ordinální data nezaručují menší počet faktorů než původní verze těchto algoritmů pro BMF pro binární data po naškálování, že ano? Rank ordinálních dat je vždy menší (Věta 3), to ano, ale v souhrnech výsledků z experimentů s reálnými daty je pro některá data pro nižší práh pokrytí uváděno méně faktorů vypočítaných BMF algoritmem GreConD než rozšířeným algoritmem GreConD$_L$.

2. Jaké jsou vhodné stupně K pro algoritmus Asso$_L$? Dále, algoritmus nevykazuje v experimentech příliš dobré výsledky. Proč uváděný problém, faktory s hodnotami „okolo středu“ škály L, není problém i u ostatních algoritmů? A nelze to řešit tak, že tyto hodnoty nebudou mezi těmi vhodnými?

3. V případě binární matice lze tzv. esenciální prvky snadno (efektivně) určit. Lze je snadno určit i pro matice s hodnotami z uspořádané škály? Jak?

Práce, vytvořená na základě tří publikovaných a tří (zatím) nepublikovaných článků, jejichž je autorka spoluautorkou, obsahuje netriviální výsledky a je až na obsahové výtky výše napsaná pečlivě (včetně dobrého přehledu souvisejících prací). I s ohledem na další práce autorky *doporučuji* tuto práci u obhajoby uznat jako úspěšnou dizertační práci.

V Olomouci dne 16. května 2017
doc. Mgr. Jan Outrata, Ph.D.