Dolování ordinálních asociačních pravidel

Filip Karel, Jiří Kléma [1]

[1] Katedra kybernetiky, FEL, ČVUT - České vysoké učení
technické v Praze,
Technická 2, Praha 6, 166 27
{karelf1,klema}@fel.cvut.cz

Abstrakt. Asociační pravidla byla prvoplánově navržena jako nástroj pro vyhledávání vazeb mezi binárními atributy. Přestože je přechod na domény obsahující i jiné typy atributů relativně přímočarý, může při něm docházet ke ztrátě užitečné informace. To platí zejména v případě atributů, jejichž hodnoty lze uspořádat - ordinálních atributů. Rozličné způsoby jejich transformace na binární atributy mohou vést ke kombinatorické explozi a v konečném důsledku i k velkému množství nevýznamných pravidel. Článek diskutuje alternativní přístup, ve kterém cedenty nejsou tvořeny konjunkcí literálů, ale jednoduchými operacemi zachovávajícími uspořádání.

Klíčová slova: ordinální atribut, asociační pravidlo, dolování znalostí.

1 Úvod

V dosud publikované literatuře přistupují různí autoři k problematice generování asociačních pravidel s ordinálními atributy odlišně. Zavádí odlišnou terminologii i postupy. Shoda však panuje v tom, že standardní postupy, které se používají u binárních či kategoriálních atributů, jsou neefektivní s často nelogickými či nepoužitelnými výsledky. Spojité či diskrétní atributy, jejichž hodnoty lze uspořádat, s sebou přinášejí řadu specifik. Prvním důležitým faktorem je optimalizace automatické diskretizace spojitých atributů. Další otázkou je měření kvality získaných pravidel. Někteří autoři používají klasické míry jako podpora (Supp) a spolehlivost (Conf) [2], [3], [4], [8], případně je doplňují dalšími pomocnými „zajímavostními" měrami. Další autoři poukazují na to, že v případě ordinálních atributů je vhodnější použití jiných měr [5], [6], [9], [10] a [11].
Např. v [5] jsou navrženy míry založené na poloze jednotlivých bodů v prostoru a jejich vzdálenosti, v [10] jsou mírou kvality statistické ukazatele, v [9] a [11] jsou v první fázi pravidla hodnocena podobnými ukazateli jako podpora a spolehlivost, ale ve druhé fázi generování konkrétních pravidel je použita nově definovaná intenzita inklinace. My budeme k hodnocení pravidel používat podporu a spolehlivost doplněnou o kvantifikátor zdvihu (Lift).
Vlastní algoritmy generování pravidel jsou si v principu podobné. Nejdříve se naleznou vhodné intervaly hodnot ordinálních atributů a ty jsou pak využity v binárních testech tvořících základ tvorby tradičních asociačních pravidel. Odlišnosti spočívají v metodice vytváření intervalů a v určování kvality pravidel. Za nejvíce odlišný můžeme označit přístup prezentovaný v [9] a [11]. Ten je založen na myšlence mapování ordinálních kategorií na řadu za sebou následujících celých čísel

tak, že ordinální význam kategorií zůstane zachován (výška: malý → 0, střední → 1, velký → 2). Mapování plní současně roli transformace do množiny celých čísel a normalizace. Levé a pravé strany pravidel (v dalším textu označované jako cedenty) se vytvářejí tak, že se hodnoty jednotlivých atributů sčítají. Závislost cedentů se posuzuje na základě veličiny intenzity inklinace, asociace se vyhledávají pouze u závislých cedentů. Zjišťováním závislosti cedentů a identifikací oblastí zesílené asociace se redukuje prohledávaný prostor pravidel a omezuje možnost nalezení nahodilých souvislostí.

Právě popsaným postupem se inspiruje i přístup zvolený v tomto textu. Numerické atributy jsou nejprve převedeny na diskrétní ordinální. Z ordinálních atributů se operacemi sčítání a odčítání vytvářejí cedenty. Pouze u závislých cedentů jsou vyhledávny oblasti zesílených asociací, tyto oblasti jsou popsány a následně rozloženy na klasická asociační pravidla, tedy asociace mezi konjunkcemi literálů.

Diskretizace spojitých atributů není z prostorových důvodů diskutována. Kapitola 2 se zabývá vytvářením pravidel nad cedenty délky 1 (triviální cedenty). Testuje nezávislost cedentů a rekapituluje postup, jak pro závislé triviální cedenty vytvářet asociační pravidla. Klíčová kapitola 3 diskutuje vytváření netriviálních cedentů založených na více atributech. Současně zobecňuje postupy uvedené v kapitole 2. V kapitole 4 je navržený postup porovnán s tradičními metodami vytváření asociačních pravidel. Srovnání je provedeno nad reálnou doménou STULONG [12].

2 Pravidla s triviálními cedenty

Pro jednoduchost uvažujme nejprve pravidla vyjadřující vztah pouze mezi dvěma triviálními cedenty. Testování nezávislosti cedentů se redukuje na dobře známou úlohu testování nezávislosti kategoriálních atributů, použijeme χ^2 test dobré shody. Vyřazujeme vzájemně nezávislé cedenty, čímž zabráníme vytváření potenciálně nahodilých pravidel. Zkusme otestovat závislost cedentů výška a váha osoby. Na obrázku 1 je příslušná kontingenční tabulka (oba atributy jsou diskretizovány do 5 kategorií) a rozdílová tabulka skutečných a očekávaných hodnot za předpokladu nezávislosti.

skutečné		VAHA					
		1	2	3	4	5	
VYSKA	1	77	77	34	12	3	203
	2	55	127	102	36	6	326
	3	24	93	143	60	17	337
	4	14	72	127	96	35	344
	5	1	17	57	80	44	199
		171	386	463	284	105	1409

rozdíl		VAHA					
		1	2	3	4	5	
VYSKA	1	53	21	-33	-29	-12	0
	2	15	38	-5	-30	-18	0
	3	-17	1	32	-8	-8	0
	4	-28	-22	14	27	9	0
	5	-23	-38	-8	40	29	0
		0	0	0	0	0	0

Obrázek. 1. Kontingenční tabulky skutečných a rozdílových hodnot cedentů VYSKA a VAHA

Výška a váha jsou evidentně závislé ($p<<0.001$), pokračujeme tedy vyhledáváním asociačních pravidel. Vycházíme z rozdílové kontingenční tabulky, zaměřujeme se na

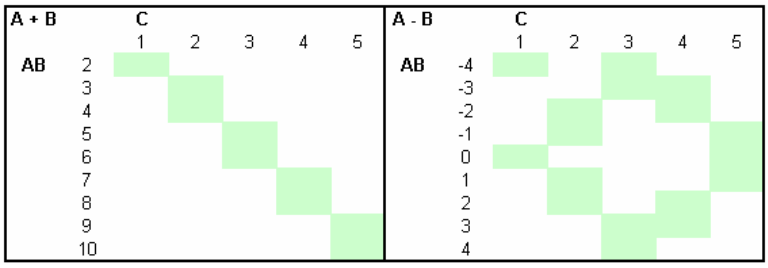
oblasti kladných hodnot, které odpovídají oblastem zesílené asociace. Hledáme největší obdélníky obsahující výhradně kladné hodnoty (viz. obrázek 1). Z každého obdélníku získáváme jedno pravidlo, pravidla ohodnotíme pomocí klasických měr kvality jako je podpora, spolehlivost a zdvih [2].

Tabulka 1 – pravidla a jejich měřítka kvality

č.	pravidlo	spolehlivost	zdvih	podpora
1	vyska = 1..2 → vaha = 1..2	0,64	1,61	0,24
2	vyska = 1..3 → vaha = 2	0,34	1,25	0,21
3	vyska = 3 → vaha = 2..3	0,70	1,16	0,17
4	vyska = 3..4 → vaha = 3	0,39	1,19	0,19
5	vyska = 4 → vaha = 3..4	0,75	1,24	0,18
6	vyska = 4..5 → vaha = 4..5	0,47	1,70	0,18

3 Generování pravidel s netriviálními cedenty

Pro cedenty jednotkové délky jde o triviální postup. Asociační pravidla však obvykle tvoříme především pro netriviální cedenty, tj. chceme kombinovat více ordinálních atributů na straně antecedentu nebo sukcedentu. Problém, který potřebujeme vyřešit, je, jak reprezentovat hodnoty více atributů hodnotou jedinou.
Řešením může být prosté sečtení nebo odečtení hodnot jednotlivých atributů. V případě dvou atributů tvořících antecedent můžeme uvažovat buď sčítání A+B nebo odčítání A-B, zbylé možnosti -A-B, B-A jsou pouhým doplňkem předchozích dvou. Jaké jsou rozdíly mezi těmito dvěmi možnostmi? Mějme následující situaci. A a B jsou ordinální diskretizované atributy, nabývají náhodně a na sobě nezávisle hodnot 1 až 5. Atribut C nabývá také hodnot 1 až 5 a je na atributech A a B závislý. V 80% hodnot je přímo úměrný hodnotě A a B. Tato závislost je lineární. Ve zbylých 20% případů je hodnota atributu C náhodná. Na obrázku 2 máme vyznačené oblasti kladných hodnot v rozdílových kontingenčních. V prvním případě získáme hodnotu antecedentu AB součtem A+B, ve druhém případě rozdílem A-B.

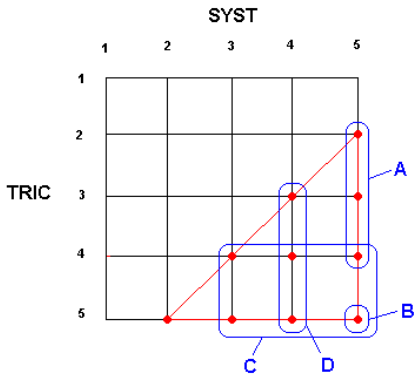


Obrázek 2. Oblasti kladných hodnot v rozdílových kontingernčních tabulkách

Z obrázků je zřejmé, že nejčitelnější a nejvýhodnější obrazec pro další generování je získán pokud je netriviální atribut utvořen podle toho jak jeho jednotlivé triviální atributy ovlivňují atribut na straně sukcedentu. A to tak, že pokud je hodnota atributu

sukcedentu přímo úměrná hodnotě atributu antecedentu, tak hodnotu ve výpočtu netriviálního atributu přičítáme a pokud je nepřímo úměrná, hodnotu odečítáme. Hodnota χ^2 je také vyšší pokud je cedent utvořen ve shodě s tím, jak ovlivňuje cedent na opačné straně pravidla. Pokud jej vytvoříme nevhodně, pak hrozí riziko, že nenalezneme existující závislost a pomineme významná pravidla. Vyvstává tedy problém optimálního vytváření cedentů. Jednoduchým řešením je hodnocení pomocí testu nezávislosti. Antecedentová kombinace, která dosáhne vyšší hodnoty χ^2, zřejmě více ovlivňuje sukcedent a je zde předpoklad nalezení „kompaktnějších" pravidel. Připomeňme, že vícenásobné testování nezávislosti k časové náročnosti celého algoritmu přispívá minimálně.

Nyní již přistupme ke generování pravidel. Budeme hledat závislost sukcedentu SUBSC (kožní řasy subscalpular) na antecedentu TRIC/SYST (kombinace kožní řasy triceps a systolického tlaku). Hodnota χ^2 je vyšší vytvoříme-li antecedent TRIC/SYST sečtením hodnot atributů TRIC a SYST než jejich odečtením. Oba antecedentové atributy nabývají po diskretizaci hodnot 1 až 5, součtový antecedent TRIC/SYST tedy nabývá hodnot 2 až 10. Aplikujeme postup uvedený v kapitole 2, pro všechny obdélníky rozdílové kontingenční tabulky zavádíme navíc dvě měřítka kvality Tc a Pp [9]. Tc je podpora daného obdélníku a Pp je poměr kladných hodnot obsažených v daném obdélníku a součtu všech kladných hodnot v celé tabulce. Z těchto obdélníků (potažmo pravidel) vyřadíme ty, které nesplňují podmínky minimální Pp a Tc. Vhodnou volbou měřítek Tc a Pp dochází k další výrazné úspoře prohledávaného prostoru při téměř nulové ztrátě generovaných pravidel a nalezených závislostí.



Obrázek 3. Vybrané obdélníky v zajímavé oblasti catributů TRIC/SYST.

Uvažujme, že jedním z pravidel získaných dosavadními kroky je i pravidlo TRIC/SYST = 7..10 → SUBSC = 5. Nyní je třeba provést dekompozici cedentu TRIC/SYST. Situaci demonstruje obrázek 3. Červeně je vyznačena oblast, která odpovídá součtu atributů TRIC + SYST = 7..10, tj. oblast odpovídající antecedentu našeho pravidla. K rozkladu musíme evidentně testovat všechny obdélníky, které lze vytvořit uvnitř inkriminované oblasti. Na obrázku 3 jsou jako příklad vybrané čtyři obdélníky. V tabulce 2 jsou pravidla, která odpovídají těmto obdélníkům i s jejich měřítky kvality. Měřítka opět slouží k eliminaci pravidel. Lze použít například

jednoduchou heuristiku, která pro každý obdélník vybírá maximálně 2 pravidla. První s maximální podporou, druhé s maximálním zdvihem (resp. spolehlivostí), oboje pouze v případě překročení prahových hodnot. V našem příkladu jde o pravidla B a C. Závislost počtu generovaných pravidel na veličinách minimální podpora, spolehlivost a zdvih je téměř lineární.

Tabulka 2 - pravidla získaná z vyznačených obdélníků A, B, C a D

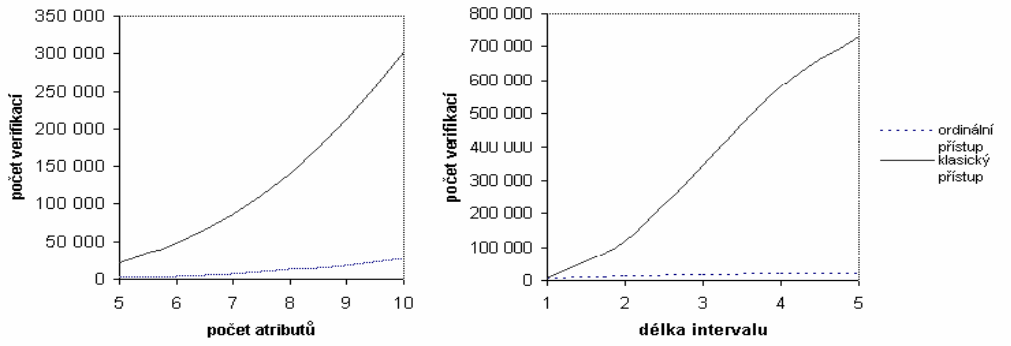
obd.	pravidlo	spolehlivost	zdvih	podpora
A	TRIC = 2..4 ∧ SYST = 5 → SUBSC = 5	0,18	1,64	0,02
B	**TRIC = 5 ∧ SYST = 5 → SUBSC = 5**	**0,60**	**2,10**	**0,09**
C	**TRIC = 4..5 ∧ SYST = 3..5 → SUBSC = 5**	**0,46**	**1,60**	**0,17**
D	TRIC = 3..5 ∧ SYST = 4 → SUBSC = 5	0,38	1,31	0,05

4 Srovnání s klasickým přístupem

V tomto odstavci srovnáme ordinální přístup s přístupem klasickým. Ve zvolené modifikaci klasický přístup využívá stejné diskretizace jako ten ordinální. Z ordinálních atributů ovšem generuje atributy binární. Používá metody intervalů, tj. binární atribut vyjadřuje příslušnost k množině sousedících kategorií tvořících interval dané délky. Parametrem je maximální délka intervalu. Pracujeme-li s atributem, jehož hodnoty jsou rozděleny do kategorií 1,..,5 a maximální délkou intervalu 2, můžeme generovat intervaly {1}, {1,2}, {2}, {2,3}, atd. Následně ověřujeme všechny jevy vyjádřitelné konjunkcí daného počtu binárních testů. Klasický přístup budeme reprezentovat systémem LISp Miner, procedurou 4ft-Miner [13]. Srovnání provedeme nad reálnou doménou STULONG. Jde o epidemiologickou studii primární prevence ateroskleróz. Reálné výsledky zaměřené na analýzu trendů využívající ordinálních asociačních pravidel byly zpracovány a prezentovány na Discovery Challenge ECML/PKDD 2004 [14]. V rámci srovnání budeme pracovat s databází obsahující 859 záznamů. Jeden záznam odpovídá jedné sledované osobě, každá osoba je popsána 10 ordinálními atributy jako např. váha, výška, tlak nebo cholesterol. Každý atribut je diskretizován do 5 kategorií.

Časovou náročnost obou přístupů budeme hodnotit na základě počtu provedených verifikací, tj. u kolika kandidátských pravidel ověřujeme splnění příslušných měr kvality. Nejprve budeme studovat nárůst počtu verifikací s rostoucím počtem atributů v databázi (levá část obrázku 4). Počet atributů nejprve uměle omezíme na 5, pak postupně budeme zvyšovat až do maximálního počtu 10. Antecedent je v tomto experimentu tvořen 2 atributy (má délku 2), sukcedent má délku 1. Délka intervalů jednotlivých atributů je menší nebo rovna 3.

Je vidět, že u klasického přístupu roste počet verifikací velmi rychle. Pro 10 atributů je počet verifikací vyšší než 300 000. U ordinálního přístupu (Pp = 0,1; Tc = 0,1) je počet verifikací přibližně 30 000. Počet nalezených pravidel je přitom řádově stejný (viz. dále). Ordinální přístup řadu verifikací eliminuje testováním nezávislosti cedentů, k dalšímu zjednodušení dochází zaměřením na oblasti zesílené asociace - obdélníky s kladnými hodnotami v kontingenční tabulce.



Obrázek 4. Závislost počtu verifikací na počtu atributů a délce intervalu.

U klasického přístupu tvoří verifikace pravidel prakticky 100% časové náročnosti. U ordinálního přístupu je tomu jinak, verifikace konkrétních pravidel tvoří od 60% (u relativně nezávislých dat) do 90% (u velmi závislých dat) celkové časové náročnosti. Zbylý čas je spotřebován na optimalizaci konstrukce cedentů, testování jejich nezávislosti a identifikaci oblastí zesílené asociace. I tak však dosahuje celková časová náročnost maximálně 15 až 20% klasického přístupu.

Výhoda ordinálního přístupu dále narůstá pokud zvyšujeme maximální povolenou délku intervalu (uvažujeme všech 10 atributů, zbylé parametry jsou nastaveny stejně jako v předchozím experimentu). Z pravé části obrázku 4 je zřejmé, že již při délce 5 je počet verifikací v klasickém přístupu až 35 násobný ve srovnání s ordinálním.

Pro větší délky intervalu je u klasického přístupu počet pravidel zhruba třikrát vyšší. Jedná se však většinou o opakující se pravidla, která nijak nepřispívající k celkovému porozumění závislostí mezi cedenty. U ordinálního přístupu stoupá počet pravidel pomalu zejména proto, že z každého obdélníku vybíráme pouze 2 nejlepší pravidla. Omezení zajišťuje, že nejsme zahlceni mnoha podobnými pravidly popisujícími stejnou nebo příbuznou asociaci.

Srovnávání počtu nalezených pravidel (asociací) však obecně nemůžeme provádět při stejných hodnotách parametrů minimální podpory, spolehlivosti a zdvihu. Jak již bylo uvedeno dříve, u ordinálního přístupu slouží tato měřítka spíše k eliminaci velmi špatných pravidel. U ordinálního přístupu při mnohem nižších měřítkách kvality dosahujeme až o 90% méně verifikací, řádově stejného počtu generovaných pravidel a stejného či vyššího počtu nalezených asociací. Z toho také vyplývá, že poměry počet pravidel / počet verifikací, počet asociací / počet verifikací a počet asociací / počet pravidel je u ordinálního přístupu mnohem vyšší.

Jak bychom nastavili jednotlivé parametry tak, abychom obdrželi přibližně stejné počty pravidel? Pokud ve zvolené doméně u klasického přístupu volíme MinSupp=0.1, MinConf=0.6 a MinLift=0.8, získáme při 212 631 verifikacích 449 pravidel postihujících 94 různých asociací (odlišnost asociací byla hodnocena subjektivně). Pokud volíme u ordinálního přístupu MinSupp=0.05, MinConf=0.4, MinLift=0.5, Pp=0.1 a Tc=0.1 získáme při 19 321 verifikacích 509 pravidel postihujících 207 různých asociací. Je tedy zřejmé, že u ordinálního přístupu lze vyhledávat i slabší asociace bez kombinatorické exploze spojené se snižováním prahových hodnot.

Riziko vynechání asociace u ordinálního přístupu hrozí pokud je sukcedent závislý na jednom ze dvou antecedentových atributů a na druhém závisí minimálně. Pokud např. platí A=vysoké & B=nízké → C=nízké a A=nízké & B=vysoké → C=vysoké, pak je součet A+B při vytváření antecedentu pokaždé stejný. Přitom C je jednou nízké a jednou vysoké. To by však znamenalo, že celý řádek v rozdílové tabulce by byl kladný. To je však nemožné, protože rozdílové tabulce musí být součty řádků a sloupců nulový. Většinou je však z ostatních získaných pravidel zřejmá klíčová role daného jednoho atributu.

5 Závěr

Článek diskutuje problematiku získávání ordinálních asociačních pravidel. Inspiruje se přístupy navrženými v literatuře a navrhuje vlastní algoritmus pro dolování asociačních pravidel v ordinálních doménách. Postup je založen na myšlence testování závislosti cedentů. Případné numerické atributy jsou nejprve převedeny na diskrétní ordinální. Z ordinálních atributů se vytvářejí cedenty jednoduchými operacemi sčítání a odčítání. Pouze u závislých cedentů jsou vyhledávny oblasti zesílených asociací, tyto oblasti jsou popsány a následně rozloženy na klasická asociační pravidla, tedy asociace mezi konjunkcemi literálů. Článek se zabývá volbou vhodného algoritmu diskretizace numerických atributů, optimalizací metod vytváření a testování závislosti cedentů a také postupy jejich následného rozkladu. Navržený postup je porovnán s tradičními metodami vytváření asociačních pravidel.
Navrhovaný postup nelze chápat jako postup čistě konkurenční k tradičním metodám dolování asociačních pravidel založených na variantách algoritmu APRIORI. Důvodem je to, že se nejedná o úplný algoritmus zaručující nalezení všech pravidel vyhovujících vstupním mírám zajímavosti pravidla. Problematická je také kombinace ordinálních a nominálních atributů majících více než 2 kategorie. Tyto vlastnosti však mohou být v řadě případů „plně" vyváženy menší složitostí algoritmu danou významně menším počtem ověřovaných asociací. Ta v důsledku vede k menšímu počtu generovaných pravidel při současném omezení jinak obvyklých podobností mezi jednotlivými pravidly. Tyto vlastnosti se mohou příznivě projevit zejména při práci s rozsáhlými databázemi, popřípadě v situacích, kdy vyhledáváme složité asociace vyjádřené kombinací většího počtu literálů.

Reference

1. Agrawal R., Imeliski T., Swami A. Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Conference on Management of Data.* pp. 207-216, Washington, D.C., 1993.
2. Srikant R., Agrawal R. Mining Quantitative Association Rules in Large Relational Databases. *ACM SIGMOD Conference on Management of Data.* Montreal, Canada, 1996.

3. Fukuda T., Morimoto Y., Morishita S., Tokuyama T., Mining Optimized Association Rules for Numeric Attributes, *ACM SIGMOD Conference on Management of Data*, Montreal, Canada, 1996.
4. Fukuda T., Morimoto Y., Morishita S., Tokuyama T., Data Mining Using Two-dimensional Optimized Association Rules: Schemes, Algorithms and Visualization, *ACM SIGMOD Conf. on Management of Data*, Tuscon, AZ, 1997.
5. Miller R.J., Yang Y., Association Rules over Interval Data, *ACM SIGMOD Conference on Management of Data*, Tuscon, AZ, 1997.
6. Imberman S., Domanski B., Finding Association Rules from Quantitative Data Using Data Booleanization, 1999.
7. Webb G.I., Discovering Associations with Numeric Variables, *ACM SIGMOD Conference on Management of Data*, San Francisco, CA, 2001.
8. Rastogi R. Shim K., Mining Optimized Association Rules with Categorical and Numeric Attributes, *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, 2002.
9. Guillaume S., Discovery of Ordinal Association Rules, *Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD'02), 322-327, Taipei, Taiwan, 6-8 May 2002.
10. Aumann Y., Lindell Y., A Statistical Theory for Quantitative Association Rules, *Journal of Intelligent Information Systems*, vol. 20, 255--283, 2003.
11. Guillaume S., Ordinal Association Rules towards Association Rules, *Data Warehousing and Knowledge Discovery: 5th International Conference DaWaK*, Prague, Czech Republic, 3-5 September 2003.
12. Projekt STULONG, WWW page, http://euromise.vse.cz/stulong.
13. Projekt LISp Miner, WWW page, http://lispminer.vse.cz/.
14. Kléma J., Nováková L., Karel F., Štěpánková O., Trend Analysis in Stulong Data, *In ECML/PKDD'04 workshop proceedings: A Collaborative Effort in Knowledge Discovery from Databases*, 56--67, 2004.

Annotation:

Ordinal association rules mining

Association rules have exhibited an excellent ability to identify interesting association relationships among a set of binary variables describing huge amount of transactions. Although the rules can be relatively easily generalized to other variable types, the generalization can result in a computationally expensive algorithm generating a prohibitive number of redundant rules of little significance. This danger especially applies to ordinal variables. The paper presents and verifies an alternative approach to the ordinal association rule mining. In this approach, ordinal variables are not immediately transformed into a set of binary variables. Instead, it applies simple arithmetic operations in order to construct the cedents, tests their independence and searches for areas of increased association which are finally decomposed into conjunctions of literals. This scenario outputs rules that do not syntactically differentiate from classical association rules.