Datamining a AA (Above Average) kvantifikátor

Jan Burian

Laboratory of Intelligent Systems, Faculty of Informatics and Statistics, University of Economics, W. Churchill Sq. 4, 13067 Prague, Czech Republic, burianj@vse.cz

Abstrakt. Na základě frekvencí ze čtyřpolní kontingenční tabulky je definován AA kvantifikátor implementovaný v GUHA proceduře 4ftMiner, která je součástí analytického systému LISp-Miner. Je uveden motivační příklad. Dále je diskutována pravděpodobnostní interpretace AA kvantifikátoru. Nakonec jsou shrnuty vlastnosti AA kvantifikátoru a možnosti jejich využití.

1 Úvod. Cíle a obsah příspěvku.

Jedním ze způsobů dobývání znalostí z databází je vyhledávání důležitých vztahů, které se týkají vztahů dvou Booleovských atributů (často to jsou konjunkce literálů) derivovaných z analyzované databáze. Mezi dvěma konjunkcemi literálů vytvořených z kategorií atributů tabulky relační databáze, mohou existovat zajímavé vztahy (asociační pravidla), které je možno zkoumat na základě tak zvané čtyřpolní kontingenční tabulky (dále jen čtyřpolní tabulky). První skupinu literálů nazvěme antecedent(zkráceně budeme označovat jako φ),druhou sukcedent (zkráceně budeme označovat jako ψ). čtyřpolní tabulka má v takovém případě podobu:

	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Tabulka 1. čtyřpolní tabulka φ a ψ

Ve čtyřpolní tabulce "a" reprezentuje počet (frekvenci) všech případů (záznamů), kdy je splněn jak antecedent tak sukcedent. "b" je počet všech případů kdy je splněn antecedent a není splněn sukcedent a tak dále. Zobecněný kvantifikátor odráží nějakou charakteristiku vztahu mezi antecedentem a sukcedentem vypočítanou na základě frekvencí ze čtyřpolní tabulky a vnějších parametrů. Jedním z velmi praktických a snadno interpretovatelných zobecněných kvantifikátorů je takzvaný AA kvantifikátor (Above Average kvantifikátor). Tento příspěvek pojednává o motivaci pro jeho zavedení, jeho interpretaci a vlastnostech. AA kvantifikátor byl implementován v proceduře 4ft-Miner (součást dataminingového systému LISp-Miner [7]), který pracuje na základě metody GUHA [4].

2 Motivace pro zavedení AA kvantidikátoru

Asociační pravidla používaná k analýze dat procedurou 4ftMiner odrážejí souvislosti v datech, ale mohou být špatně interpretovatelná vzhledem k reálně existujícím vztahům ve světě. Vypovídací schopnost asociačních pravidel může být deformovaná strukturou dat. Jedním z probémů při interpretaci asociačních pravidel může být, že kategorie atributů mohou mít zcela nerovnoměrné zastoupení. Jako příklad vezměme data zkoumaná např. v [6]. Jednalo se o medicínská data (konkrétně databáze hypertoniků v rámci projektu EuroMISE [8]), kde záznamy v tabulce databáze představovala vyšetření a atributy hodnoty sledovaných charakteristik vyšetření, mimo jiné také měsíc ve kterém bylo vyšetření provedeno.
Dejme tomu, že antecedent bude představovat například tlak pacienta a sukcedent měsíc vyšetření. Počet kontrol v jednotlivých měsících (sukcedent) značně kolísá. Pak ale bude například vysoký tlak (stejně jako jiné hodnoty tlaku) velmi zřídka zastoupen v měsících, kdy je počet kontrol mizivý (typicky letní měsíce - pacienti i lékaři jsou na dovolených apod.).
Jak by dopadl výsledek analýzy pokud bychom použili jeden z klasických vztahů implemetovaných v proceduře 4ftMiner (a vůbec ve všech GUHA procedurách), tak zvanou fundovanou implikaci [4]?

Definice 1. *Mezi antecedentem a sukcedentem je vztah fundovanimplikace* $\Rightarrow_{p,s}$ *(formule* $\varphi \Rightarrow_{p,s} \psi$ *platí) tehdy a jen tehdy pokud pro* $0 < p \leq 1$ *a* $s > 0$ *platí* $\frac{a}{a+b} \geq p \ \wedge \ a \geq s$.

Parametr s umožňuje zohlednit strukturu dat jen velmi hrubě. Proto od něj nyní odhlédneme.
Pokud použijeme fundovanou implikaci k analýze dat, získáme spíše hypotézy (asociační pravidla) týkající se souvislosti vysokého tlaku s měsíci ve kterých bylo zaznamenáno větší procento celkového počtu vyšetření. Konkrétně pro hodnotu p=0.1 bychom při průměrném rozložení počtu vyšetření na měsíc očekávali že bude nalezen vztah mezi vysokým tlakem a 11. měsícem, ve kterém bylo zaznamenáno $12,68\%$ procent všech vyšetření, méně už bychom to čekali u 12.měsíce ve kterém bylo zaznamenáno $7,71\%$ procent všech vyšetření. A to už nemluvíme o letních měsících, kdy je procento všech zaznamenaných vyšetření menší než jedno procento.
Nejvíce vyšetření s extrémní hodnotou bude nejspíše v tom období, kdy je nejvíce vyšetření obecně. Tedy samotný počet vyšetření s vysokou hodnotou tlaku v jistém měsíci, nic neříká o tom, zda pacienti v tomto období skutečně mají vyšší sklon mít vysoký tlak. Důvod je v tom, že fundovaná implikace nezohledňuje podíl frekvence případů, kdy je platný sukcedent (a+c) k celkovému počtu vyšetření (a+b+c+d).
Pokud bychom chtěli generovat všechny platné hypotézy (asociační pravidla) pro jistým způsobem parametrizovanou skupinu atributů v antecedentu a sukcedentu (tak jak to umí pomocí tzv koeficientů např. procedura 4ftTask), pak

bychom pro smysluplnou analýzu závislosti vysokého tlaku pacientů v jistém měsíci, museli analýzu provádět zvlášť pro všechny různé kategorie sukcedentu (měsíce), pokaždé s příslušným parametrem p, který by odpovídal podílu počtu vyšetření v příslušném měsíci (a+c) na celkovém počtu vyšetření (a+b+c+d). Poznamenejme ještě, že pokud bychom zaměnili antecedent a sukcedent, situace se nezmění. Obecně může být atribut (či atributy) s více nerovnoměrně rozloženými hodnotami jak v sukcedentu tak v antecedentu.
Vlivu nerovnoměrného rozložení hodnot nás zbaví vhodně nadefinovaný zobecněný kvantifikátor. Jedním z takových kvantifikátorů je AA kvantifikátor. Na rozdíl od kvantifikátorů, které provádějí operace ekvivalentní statistickým testům (jako je F test či χ^2 test) je AA kvantifikátor výpočetně mnohem jednodušší.

3 AA kvantifikátor

AA kvantifikátor reprezentuje tento fakt:
Procento objektů které splňují φ i ψ z objektů které splňují ψ, je aspoň $(1 + p)$ krát větší než průměrné procento (procento objektů které splňují φ ze všech objektů v analyzované matici dat).

Definice 2. *Pro každou čtyřpolní tabulku $\langle a, b, c, d\rangle$ platí mezi antecedentem a sukcedentem vztah daný kvantifikátorem Above Average tehdy a pouze tehdy když:*

$$a + b > 0 \ \wedge \ a + c > 0$$

a zároveň

$$\frac{a * (a + b + c + d)}{(a + b) * (a + c)} \geq (1+p)$$

Parametr p je definován v intervalu $(-1; +\infty)$.
Poznámka 1: Výraz na levé straně nerovnice zmenšený o 1 nazýváme Average difference.
Poznámka 2: AA kvantifikátor je symetrický - je možno vzájemně zaměnit symboly b a c. Vyjadřuje tedy i fakt:
Procento objektů které splňují (mají atribut) ψ i φ z objektů které splňují φ, je aspoň $(1 + p)$ krát větší než průměrné procento (procento objektů které splňují ψ ze všech objektů v analyzované matici dat). Poznámka 3: V současné verzi procedury 4ftMiner je možno parametr p zadat pouze v intervalu $(0; +\infty)$. Pro záporné hodnoty parametru p není již název Above average relevantní, neboť jsou generována i pravidla, pro něž procento objektů které splňují (mají atribut) ψ i φ z objektů které splňují φ, je nižší než průměrné procento, ale stále vyšší než procento dané parametrem p.
Poznámka 4: Obdobně jako AA kvantifikátor je definován k němu opačný BA kvantifikátor (Below Average), který se liší znaménkem nerovnosti v definici. Poznámka 5: V [6] je definován AA kvantifikátor pomocí tzv. asociované fukkce. Její použití však vyžaduje zavedení tzv. 4FT kalkulu viz např [2], ten zde však vzhledem k rozsahu příspěvku nepoužijeme.

4 Pravděpodobnostní interpretace AA kvantifikátoru

Obdobu AA kvantifikátoru definoval jako tak zvanou zajímavost (interestingness) Kodratoff [5]. Kodratoff zkoumal různé charakteristiky asociačních pravidel na základě vztahů mezi pravděpodobnostmi a podmíněnými pravděpodobnostmi sukcedentu (S) a antecedentu (A). K těmto vztahům se poté pokoušel nalézt odpovídající algebraický výraz sestavený z frekvencí obsažených ve čtyřpolní tabulce. Pro názornější ilustraci uveďme různé pravděpodobnosti týkající se antecedentu a sukcedentu vyjádřené pomocí algebraických výrazů sestavených z frekvencí obsažených ve čtyřpolní tabulce.

$$P(A \wedge S) = \frac{a}{a+b+c+d}$$

$$P(A) = \frac{a+b}{a+b+c+d}$$

$$P(S) = \frac{a+c}{a+b+c+d}$$

$$P(S|A) = \frac{a}{a+b}$$

$$P(A|S) = \frac{a}{a+c}$$

Zajímavost (interestingness) reprezentuje tento vztah:

$$\frac{P(A \wedge S)}{P(A) * P(S)} = \frac{a * (a+b+c+d)}{(a+b) * (a+c)}$$

Vztah na pravé straně rovnice je pouze jinak formulovaný vztah z definice AA kvantifikátoru. Kodratoff neuvádí, i když to z výše uvedeného vyplývá, že jeho zajímavost můžeme vyjádřit za pomocí podmíněných pravděpodobností také jako:

$$\frac{P(A|S)}{P(A)} = \frac{P(S|A)}{P(S)}$$

Kodratoff také nezkoumá žádné další vlastnosti tohoto vztahu relevantní pro metodu GUHA. Pravděpodobnostní vyjádření je však velmi praktické pro pochopení reálného uplatnění AA kvantifikátoru. Pokud by veličiny, které představuje antecedent a sukcedent byly navzájem zcela nezávislé, pak by

$$P(A \wedge S) = P(A) * P(S)$$

A tedy

$$\frac{P(A \wedge S)}{P(A) * P(S)} = \frac{a * (a+b+c+d)}{(a+b) * (a+c)} = 1$$

Zajímavost popisuje míru vzájemné závislosti antecedentu a sukcedentu. Hodnota zajímavosti vyšší než 1 znamená vzájemnou závislost antecedentu a sukcedentu v tom smyslu, že tyto mají sklon vyskytovat se v jednom záznamu častěji,

než při vzájemné nezávislosti (tedy antecedent a sukcedent se navzájem "přitahují").
Naopak zajímavost menší než 1 znamená vzájemnou závislost antecedentu a sukcedentu v tom smyslu, že tyto mají sklon vyskytovat se v jednozm záznamu méně často než při vzájemné nezávislosti (tedy antecedent a sukcedent se navzájem "odpuzují").
Pokud vztáhneme tuto pravděpodobnostní interpretaci k parametru p AA kvantifikátoru, tak jak jsme ho definovali v předchozí kapitole, pak například:
Hypotézy nalezené pro p=0.5 dávají do vztahu ty antecedenty a sukcedenty, které jsou splněny zároveň v nejméně o 50% více záznamench, než by tomu bylo při jejich vzájmené nezávislosti.

5 Vlastnosti AA kvantifikátoru

V [4] jsou definovány vlastnosti symetričnosti a asociačnosti zobecněných kvantifikátorů (třídy kvantifikátorů). V [6] jsou tyto vlastnosti definovány pomocí tzv. TPC (Truth preservation condition). Ve [2] je pak definována F-vlastnost.

Definice 3. *Zobecněný kvantifikátor $\sim$ patří do třídy asociačních kvantifikátorů, jestliže pro každé dvě čtyřpolní tabulky $\langle a, b, c, d\rangle$ a $\langle a', b', c', d'\rangle$ platí:*

Platí-li vztah daný kvantifikátorem $\sim$ pro čtyřpolní tabulku $\langle a, b, c, d\rangle$ a zároveň

$$a' = a \ \wedge \ b' = c \ \wedge \ c' = b \ \wedge \ d' = d$$

pak platí tento vztah i pro čtyřpolní tabulku $\langle a', b', c', d'\rangle$

Definice 4. *Zobecněný kvantifikátor $\sim$ patří do třídy asociačních kvantifikátorů, jestliže pro každé dvě čtyřpolní tabulky $\langle a, b, c, d\rangle$ a $\langle a', b', c', d'\rangle$ platí:*

Platí-li vztah daný kvantifikátorem $\sim$ pro čtyřpolní tabulku $\langle a, b, c, d\rangle$ a zároveň

$$a' \geq a \ \wedge \ b' \leq b \ \wedge \ c' \leq c \ \wedge \ d' \geq d$$

pak platí tento vztah i pro čtyřpolní tabulku $\langle a', b', c', d'\rangle$

Definice 5. *Zobecněný kvantifikátor $\sim$ patří do třídy kvantifikátorů s vlastností F, jestliže pro každou čtyřpolní tabulku $\langle a, b, c, d\rangle$ platí:*
- Platí-li vztah daný kvantifikátorem $\sim$ pro čtyřpolní tabulku $\langle a, b, c, d\rangle$ a zároveň $b \geq c - 1 \geq 0$, pak platí i pro čtyřpolní tabulku $\langle a, b + 1, c - 1, d\rangle$.
- Platí-li vztah daný kvantifikátorem $\sim$ pro čtyřpolní tabulku $\langle a, b, c, d\rangle$ a zároveň $c \geq b - 1 \geq 0$, pak platí i pro čtyřpolní tabulku $\langle a, b - 1, c + 1, d\rangle$.

V [6] jsou dokázány tyto vlastnosti AA kvantifikátoru (respektive AA kvantifikátor patří do těchto tříd kvantifikátorů):

Věta 1. Vlastnosti AA kvantifikátoru:
1. AA kvantifikátor je symetrický.
2. AA kvantifikátor není asociační.
3. AA kvantifikátor má vlastnost F, až na výjimku, kdy:
Average difference $= 0 \wedge a = 0 \wedge (b = 1 \vee c = 1)$

Poznamenejme, že mnoho dříve používaných "praktických" kvantifikátorů patřila do třídy asociačních kvantifikátorů. Například kvantifikátory implikační, dvojitě implikační, ekvivalenční atd. viz např. [2]. Je zajímavé, že některé z ekvivalenčních kvantifikátorů (např. kvantifikátor prostého vychýlení a Fisherův) patří stejně jako AA kvantifikátor do třídy kvantifikátorů s vlastností F.

6 Závěr

AA kvantifikátor byl implementován v GUHA proceduře 4ftMiner, která je součástí systému LispMiner. V rámci tohto systému je v současnosti využíván například při analýzách medicínských dat v rámci projektu EuroMISE či analýzách dopravních nehod (projekt Traffic).
Zjištěné vlastnosti AA kvantifikátoru bude možno využít při hledání nových (třeba i složitějších - statistické testy) kvantifikátorů vhodných pro implementaci do systému pro datamining na bázi hledání asociačních pravidel.
Vlastnosti AA kvantifikátoru byly použity ke zkoumání možností optimalizace analytického software (například systému LISp-Miner), např. pomocí předpočítání tzv. tabulek kritických frekvencí [6].

English anotation:
The AA quantifier implemetnted in GUHA procedure 4ftTask (part of analytical system LISp-Miner is defined.) A motivation example illustrate its use. Then the probabilistic interpretation of AA quantifier is discussed. Finally the properies of AA quantifier are resumed.

Reference

1. Brachman T. - Anand Y.: The Process of Knowledge Discovery in Databases. In Fayad, U. M. er al.: Asvances in Knowledge Discovery in and data mining. AAAI Press/ The MIT Press, 1996, s. 37 -57
2. Rauch, J.: Příspěvek k logickým základům KDD, habilitiační práce, VŠE, Praha 1998
3. Rauch, J.: Classes of Four Fold Table Quantifiers. In Principles of Data Mining and Knowledge Discovery. Red. Zytkow, J - Quafafou, M. Berlin, Springer Verlag 1998, pp. 203-211
4. Hájek, P., Havránek T.: Mechanising Hypothesis Formation - Mathematical Foundations for a General Theory. Berlin - Heidelberg - New York, Springer-Verlag, 1978
5. Kodratoff, I. :Comparing machine learning and knowwledge discovery in databases. On: Lecture Notes from Machine Learning and Applications, ACAI99, Chania, Vol.1 1999.
6. Burian J.: Guha datamining, od praxe k teorii a zase zpět. Diplomová práce, VŠE, Praha 2002
7. Systém LISp-Miner, URL http://lispminer.vse.cz, 2002
8. Evropské centrum pro medicínskou informatiku, statistiku a epidemiologii - Kardio, URL http://euromise.vse.cz, 2002