DOBÝVÁNÍ ZNALOSTÍ Z DATABÁZÍ – PŘÍKLADY APLIKACÍ V KARDIOLOGICKÝCH DATECH

Jan Rauch

Anotace:

Příspěvek obsahuje základní informace o dobývání znalostí jakožto důležité disciplíně informatiky a ukazuje příklady jeho aplikací v medicínských datech. Jsou zmíněny hlavní rysy metodologie CRISP-DM a uvedeny příklady vyhledávání zajímavých asociačních pravidel za účelem orientace v neznámých datových souborech.

Klíčová slova:

Dobývání znalostí z databází, asociační pravidla, CRISP-DM, metoda GUHA, systém LISp-Miner.

1. Úvod

Dobývání znalostí z databází (DZD) je disciplína informatiky, která se intenzivně rozvíjí od počátku devadesátých let minulého století. Impulsem pro rozvoj bylo uvědomění si, že z rozsáhlých, někdy i desítky let shromažďovaných dat lze získat důležité nové informace a znalosti i když hlavní důvod shromažďování dat je jiný. Typickým příkladem takových dat jsou data uchovávaná v databázích bank a pojišťoven. Cílem DZD je nalézat neočekávané vztahy a sumarizovat data novými způsoby tak, že jsou srozumitelná a užitečná pro jejich majitele [1].

Jednou z oblastí s širokými možnostmi pro aplikace metod DZD je i medicínská informatika. Cílem tohoto příspěvku je stručně charakterizovat DZD a ukázat příklady aplikací DZD v kardiologických datech. Hlavní rysy DZD jsou stručně zmíněny v odstavci 2. Data, jichž se týkají příklady aplikací, jsou popsána v odstavci 3. Jedním z důležitých prostředků používaných pro analýzu dat při DZD jsou asociační pravidla. Dva příklady využití asociačních pravidel jsou v odstavci 4. Odstavec 5 obsahuje příklad hledání zajímavých rozdílů mezi dvěma skupinami pacientů.

2. Dobývání znalostí z databází

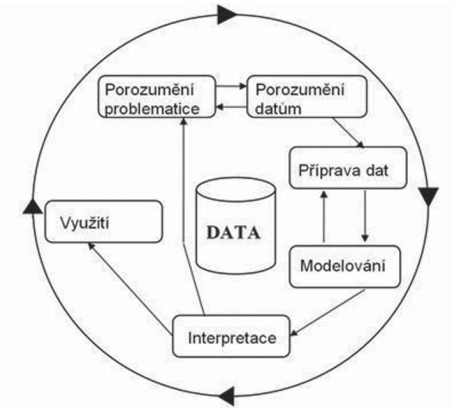
DZD je propojením tří samostatných disciplín – statistiky, databází a strojového učení [2], využívají se ale i poznatky z dalších oblastí. Je k dispozici řada metod včetně několika metodik podrobných metodik, jak DZD provádět. Podstatné je i to, že vývoj v této oblasti zdaleka není ukončen.

Podrobnější popis jednotlivých metod používaných v rámci DZD naprosto přesahuje rozsah tohoto příspěvku. Mezi metody a prostředky používané při DZD patří mimo jiné rozhodovací stromy, rozhodovací pravidla, shluková analýza, asociační pravidla, neuronové sítě, genetické algoritmy, a další. Uvedený seznam metod a prostředků DZD je velmi neúplný a nesystematický, jeho jediným cílem je naznačit rozmanitost těchto metod. Podrobný

a systematický přehled metod a prostředků DZD je uveden například v české monografii [2].

Dnes již jsou k dispozici rozsáhlé zkušenosti s aplikacemi metod DZD i různé softwarové systémy a to jak komerční tak i volně dostupné, viz http://www.kdnuggets.com/. Zkušenosti s prováděním analýz dat uchovávaných v databázích jednoznačně ukazují, že se jedná o složitý a pracný proces. Postupně pro tyto účely vzniklo několik metodologií. Patří mezi ně metoda „5A" firmy SPSS nebo metodologie SEMMA firmy SAS. Velmi známá je metodologie CRISP-DM (CRoss-Industry Standard Process for Data Mining), viz http://www.crisp-dm.org/.

Pro metodologii CRISP-DM je charakteristické chápání procesu DZD podle schématu uvedeného v Obrázku – 1. Životní cyklus projektu DZD je tvořen šesti fázemi.



Obrázek 1 – Životní cyklus projektu DZD dle metodologie CRISP-DM

- porozumění problematice
- porozumění datům
- příprava dat
- modelování
- interpretace
- využití

Pořadí jednotlivých fází není přesně dáno, výsledek jedné fáze ovlivňuje další postup, jednotlivé fáze se rozpadají na řadu dílčích kroků. Často je nutné se k některým krokům vracet. Vnější kruh na obrázku symbolizuje cyklickou povahu celého procesu dobývání znalostí z databází.

Porozumění problematice je úvodní fáze, jejímž výsledkem musí být pochopení cílů projektu a požadavků na řešení formulovaných z hlediska zadavatele. Tato formulace musí být převedena do zadání úlohy pro dobývání

znalostí z databází. Například úlohu hledání zajímavých vztahů týkajících se krevního tlaku a fyzikálního a laboratorního nálezu v konkrétních datech lze formulovat jako úlohu na hledání silných asociačních pravidel týkajících se odpovídajících atributů pacientů.

Porozumění datům vyžaduje pochopení způsobu vzniku dat. Dále je třeba posoudit kvalitu dat a získat první přehled o struktuře hodnot. Obvykle se pracuje s histogramy četností hodnot atributů a deskriptivními charakteristikami jako jsou průměry nebo extrémní hodnoty, používají se i různé vizualizační techniky.

Podstatou **přípravy dat** je vytvoření datového souboru, který bude vstupem do jednotlivých analytických procedur. Většina současných analytických procedur pracuje s daty ve tvaru matice dat. Její vytvoření však může vyžadovat hluboké porozumění analyzované databázi a vyžadovat i náročnější datové operace jako je výpočet charakteristik časových řad.

Analytické procedury použité ve fázi **modelování** realizují různé metody a algoritmy pro dobývání znalostí. Obvykle je k dispozici více různých metod pro řešení dané úlohy. Je třeba vybrat ty nejvhodnější a vyladit nastavení jejich parametrů. Jedná se tedy o iterativní proces, jehož dílčí výsledky mohou vést k potřebě modifikovat data a tedy k návratu k přípravě dat. Této fázi se také někdy říká data *mining*. Termín data *mining* se ale někdy používá i jako synonymum pro DZD.

Ve fázi **interpretace** je třeba dosažené výsledky posoudit z pohledu zadavatele a ověřit, zda byly splněny cíle formulované na počátku projektu.

Vytvořením vhodného modelu celý projekt DZD nekončí. Podle typu projektu může fáze **využití** někdy vyžadovat prosté sepsání závěrečné zprávy, může se však jednat i o zavedení nového systému pro automatickou klasifikaci nových případů (například pacientů).

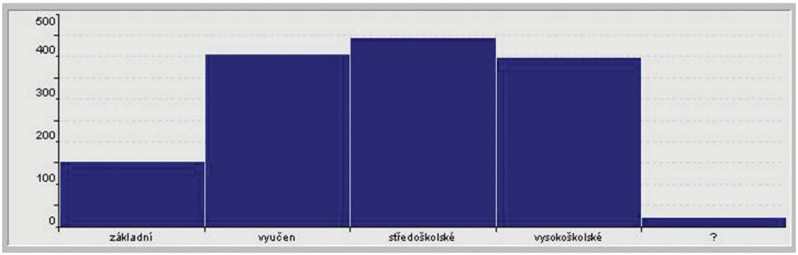
3. Data STULONG

Pro prezentaci příkladů použijeme data získaná v projektu STULONG[1]. Jeho podstatou je dlouhodobé sledování aterosklerózy u mužů středního věku. Podrobnosti jsou uvedeny na adrese *http://euromise.vse.cz/challenge2004/*. Použijeme výsledky vstupního vyšetření 1 417 mužů, Byly zjišťovány hodnoty 244 atributů pro každého pacienta, 64 atributů jsou číselné výsledky měření nebo kódy ať už získané přímo při vyšetření nebo jaké výsledky transformací ostatních atributů. Hodnoty těchto 64 atributů pro 1417 mužů tvoří matici dat Vstup, která má 1417 řádků a 64 sloupců. Řádky odpovídají pacientům a sloupce atributům popisujícím pacienty.

Těchto 64 atributů je rozděleno do 11 skupin. Pro naše účely použijeme dvě speciálně vytvořené skupiny atributů – Osobní údaje a Rizika. Jsou uvedeny v tabulce 1 spolu s frekvencemi jejich kategorií a dalšími údaji. Ukázka frekvencí kategorií atributu *Vzdělání* je v Obrázku 2. Jedná se o výstup systému LISp-Miner [3] (viz též *http://lispminer.vse.cz/*), který v příkladech používáme pro analýzy.

Skupina	Atribut		Kategorie
	Jméno	Krátké jméno	název – počet pacientů, název – počet pacientů, …
Osobní údaje	Stav	Stav	ženatý – 1207, rozvedený – 104, svobodný – 95, vdovec – 10, chybí – 1
	Věk	Věk	(35;40) – 111, (40;45) – 450, (45;50) – 680, (50;55) – 176
	Vzdělání	Vzdělání	základní – 151, vyučen – 405, středoškolské – 444, vysokoškolské – 397, chybí – 20
	Zodpovědnost v zaměstnání	Zodpovědnost	řídící pracovník – 286, částečně samostatný pracovník – 435, ostatní – 636, důchodce pro ICHS – 6 , důchodce, jiné důvody – 19 , chybí – 35
	Tělesná aktivita v zaměstnání	Telaktza	převážně sedí – 739, převážně stojí – 167, převážně chodí – 373, přenáší těžká břemena – 100, chybí – 38
	Tělesná aktivita po zaměstnání	Aktpozam	převážně sedí – 266, mírná – 1028, velká – 118, chybí – 5
Rizika	Diabetes	Diabetes	ano – 30, ne – 1378, chybí – 9
	Hypertenze	Hypertenze	ano – 220, ne – 1192, chybí – 5
	Hyperlipidemie	Hyperlipidemie	ano – 54, ne – 815, chybí – 548
	Infarkt myokardu	Infarkt	ano – 34, ne – 1378, chybí – 5

Tabulka 1 – Přehled atributů skupin Osobní údaje a Rizika.



Obrázek 2 – Frekvence kategorií atributu Vzdělání

4. Aplikace asociačních pravidel

Pojem asociační pravidlo byl zaveden Agrawalem, viz např. [4] v souvislosti s analýzou dat o nákupních košících. Pro získávání asociačních pravidel se používá algoritmus apriori, podle [5] se jedná o čtvrtý nejrozšířenější algoritmus pro data mining. Asociační pravidla jsou vhodná, mimo jiné, pro úvodní orientaci o vztazích mezi atributy v daných datech.

Téměř 30 let před Agrawalem vznikla metoda GUHA – explorační analýzy dat vyvíjená od šedesátých let dvacátého století [6,7]. Jejím principem je nabízet vše zajímavé, co lze k danému problému odvodit z daných dat. Metoda je implementována realizována pomocí GUHA procedur. Vstupem GUHA procedury jsou analyzovaná data spolu s jednoduchým zadáním rozsáhlé množiny potenciálně zajímavých tvrzení o datech. GUHA procedura vygeneruje všechna zadané tvrzení a verifikuje je v daných datech. Výstupem jsou všechna prostá tvrzení. Tvrzení je prosté, pokud je pravdivé v datech a pokud přímo neplyne z nějakého jiného jednoduššího tvrzení, které již je součástí výstupu.

Nejstarší GUHA procedura je procedura ASSOC [7], která pracuje se vztahy, které lze chápat jako podstatně zobecněná asociační pravidla. Procedura ASSOC pracuje s asociačními pravidly tvaru A ≈ S. Zde A a S jsou booleovské atributy odvozené ze sloupců analyzované matice dat. A se nazývá *antecedent* a S se nazývá *sukcedent*. Symbol ≈ je *4ft-kvantifikator*, určuje vztah mezi A a S.

Asociační pravidlo A » S je *pravdivé v matici* M, pokud podmínka přiřazená 4ftkvantifikátoru » je splněna pro kontingenční tabulku atributů A a S v matici M. V opačném případě je asociační pravidlo A ≈ S *nepravdivé v matici* M. Kontingenční tabulku atributů A a S v matici M značíme 4ft(A,S,M). Jedná se o čtveřici čísel (a,b,c,d), viz tabulka 2.

M	S	Ø S
A	a	b
Ø A	c	d

Tabulka 2 – Kontingenční tabulka 4ft(φ, ψ, M) atributů A a S v matici dat M

Zde *a* je počet řádků matice M, pro které jsou pravdivé A i S, *b* je počet řádků, pro které je A pravdivé a S nepravdivé, *c* je počet řádků, pro které je A nepravdivé a S pravdivé a *d* je počet řádků, pro které jsou A i S nepravdivé.

Nejčastěji používané 4ft-kvantifikátory jsou definovány v [7, 8, 9]. Většina z nich je implementována v proceduře 4ftMiner [9], kterou použijeme v dále uvedeném příkladu. Řada z nich je definována na základě statistických testů hypotéz. Jsou uvedeny definice čtyř často používaných 4ft-kvantifikátorů,

4ft-kvantifikátor $\Rightarrow_{p,Base}$ *fundované implikace* je pro $0 < p \leq 1$ a $Base > 0$ definován v [8] podmínkou $\frac{a}{a+b} \geq p \wedge a \geq Base$. To znamená, že je asociační pravidlo $A \Rightarrow_{p,Base} S$ je pravdivé v matici dat M pokud alespoň 100*p* procent z řádků splňujících booleovský atribut A splňuje i booleovský atribut S a pokud zároveň nejméně *Base* řádků splňuje jak A tak i S. Podíl se $\frac{a}{a+b}$ nazývá konfidence.

4ft-kvantifikátor $\Leftrightarrow_{p,Base}$ *dvojité fundované implikace* je pro $0 < p \leq 1$ a $Base > 0$ definován v [7] podmínkou $\frac{a}{a+b+c} \geq p \wedge a \geq Base$. To znamená, že asociační pravidlo $A \Leftrightarrow_{p,Base} S$ je pravdivé v matici dat M pokud alespoň 100*p* procent z řádků splňujících A nebo S splňuje jak A tak i S a zároveň nejméně *Base* řádků splňuje A i S.

4ft-kvantifikátor $\equiv_{p,Base}$ *fundované ekvivalence* je pro $0 < p \leq 1$ a $Base > 0$ definován v [7] podmínkou . $\frac{a+d}{a+b+c+d} \geq p \wedge a \geq Base$ To znamená, že asociační pravidlo $A \equiv_{p,Base} S$ je pravdivé v matici dat M pouze pokud alespoň 100*p* procent z řádků má stejnou hodnotu (buď *pravda* nebo *nepravda*) pro oba booleovské atributy A nebo S a zároveň nejméně *Base* řádků splňuje A i S.

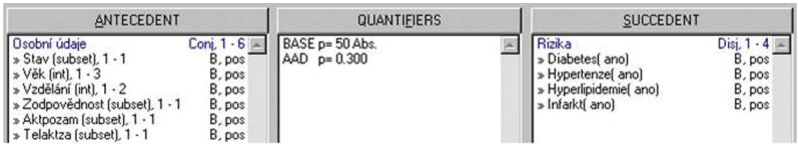
4ft-kvantifikátor $\Rightarrow^{+}_{p,Base}$ nadprůměrné závislosti, který se často nazývá *AA-kvantifikátor* (z anglického *Above Average Dependence*) je pro $0 < p$ a $Base > 0$ definován v [9] podmínkou $\frac{a}{a+b} \geq (1+p)\frac{a+c}{a+b+c+d} \wedge a \geq Base$. To znamená, že asociační pravidlo $A \Rightarrow^{+}_{p,Base} S$ je pravdivé v matici dat M pokud relativní četnost řádků splňujících booleovský atribut S mezi řádky splňujícími booleovský atribut A je alespoň o 100*p* procent vyšší než relativní četnost řádků splňujících S v celé matici a pokud zároveň nejméně *Base* řádků splňuje jak A tak i S. Vynecháme-li podmínku $a \geq Base$, pak lze říct, že AA-kvantifikátor vyjadřuje nárůst relativní četnosti alespoň o 100p procent pokud je splněna podmínka daná booleovský atributem A.

Možnosti aplikace asociačních pravidel ukážeme na řešení analytické otázky

Mají v matici dat *Vstup* některé kombinace hodnot atributů skupiny Osobní údaje vliv na nárůst relativní četnosti výskytu některých rizik nebo jejich kombinací?

Použijeme již zmíněnou proceduru 4ft-Miner. Zadanou analytickou otázku budeme řešit tak, že budeme hledat všechna asociační pravidla $A \Rightarrow^{+}_{0.3,50} S$ pravdivá v matici dat Vstup, taková že A je booleovská charakteristika skupiny atributů Osobní údaje a S je booleovský atribut vyjadřující nějakou vhodnou kombinaci rizik. Použití 4ft-kvantifikátoru $\Rightarrow^{+}_{0.3,50}$ znamená, že nás zajímají takové booleovské charakteristiky A osobních údajů a kombinace rizik S, které jsou v matici dat Vstup současně splněny alespoň pro 50 pacientů a v matici dat Vstup zároveň platí, že relativní četnost řádků pacientů s kombinací rizik S je mezi pacienty splňujícími booleovskou charakteristiku A osobních údajů alespoň o 30% vyšší, než v celé matici dat Vstup.

Procedura 4ft-Miner, stejně jako další implementace procedury ASSOC, vycházejí z reprezentace analyzovaných dat pomocí vhodných bitových řetízků [9], nepoužívá se algoritmus apriori běžně využívaný při analýze dat o nákupních košících. Tento přístup umožňuje velmi jemně definovat množinu relevantních asociačních pravidel. Rozsáhlé možnosti definice množiny relevantních asociačních pravidel, která mají být generována a verifikována jsou popsány např. v [9]. Jejich podrobnější popis přesahuje rozsah tohoto příspěvku. Jsou ale naznačeny v Obrázku 3, kde je přehledně uvedeno jedno z možných zadání procedury 4ft-Miner pro řešení naší analytické otázky.



Obrázek 3 – Příklad zadání parametrů procedury 4ft-Miner

Zadání v Obrázku – 3 znamená, že booleovské charakteristiky skupiny atributů Osobní údaje se vytvářejí jako konjunkce booleovských atributů odvozených z atributů – sloupců matice dat Vstup. V jedné konjunkci se může vyskytovat 1–6 booleovských atributů vytvořených z jednotlivých sloupců (ale maximálně jeden booleovský atribut pro každý sloupec), viz výraz Osobní údaje Conj, 1–6.

Booleovské atributy odvozené ze atributu Stav určuje výrazem Stav(subset), 1–1. To znamená, že se vytvoří všechny základní booleovské atributy Stav(α) kde α je podmnožina množiny všech kategorií atributu Stav a a obsahuje právě jednu kategorii (úplně přesně řečeno 1–1 kategorii, tedy minimálně jednu a zároveň maximálně jednu kategorii). Takto jsou zadány čtyři booleovské atributy Stav(ženatý), Stav(rozvedený), Stav(svobodný) a Stav(vdovec), viz. též tabulka 2. Pokud by byl použit výraz Stav(subset), 2–2, bylo by zadáno

šest booleovských atributů Stav(ženatý,rozvedený), Stav(ženatý,svobodný), Stav(ženatý,vdovec), Stav(rozvedený, svobodný), Stav(rozvedený,vdovec) a Stav(svobodný,vdovec). Výrazem Stav(subset), 1–2 by bylo zadáno všech deset výše uvedených základních booleovských atributů.

Základní booleovský atribut Stav(ženatý) je pro daného pacienta pravdivý, pokud je hodnota atributu Stav kategorie ženatý. Základní booleovský atribut Stav(ženatý,vdovec) je pro daného pacienta pravdivý, pokud je hodnota atributu Stav kategorie ženatý nebo vdovec.

Výrazy Zodpovědnost (subset), 1–1, Aktpozam (subset), 1–1 a Telaktza (subset), 1–1 definují analogickým způsobem množiny booleovských atributů odvozených ze sloupců Zodpovědnost, Aktpozam a Telaktza.

Booleovské atributy odvozené z atributu Věk jsou dány výrazem Věk(int), 1–3. To znamená, že z atributu Věk se vytvoří všechny základní atributy Věk(*α*) kde *α* je interval 1–3 kategorií (tedy 1–3 po sobě jdoucí kategorie atributu Věk). Příklady takových booleovských atributů jsou Věk((35;40⟩) a Věk((35;40⟩, (40;45⟩), které píšeme stručněji Věk(35;40⟩ a Věk(35;45⟩. Výrazem Věk(int), 1–3 je tedy definováno celkem 9 základních booleovských atributů Věk(35;40⟩, … , Věk(40;55⟩. Booleovské atributy odvozené ze sloupce Vzdělání jsou definovány analogicky.

Zadání v Obrázku 3 také znamená, že booleovské atributy vyjadřující kombinaci rizik jsou definovány jako disjunkce 1–4 booleovských atributů Diabetes(ano), Hypertenze(ano), Hyperlipidemie(ano), Infarkt(ano), viz. výraz Rizika Disj, 1–4.

Běh procedury 4ft-Miner zadaný dle Obrázku 3 trval 1 vteřinu na PC s procesorem Intel(R) Core(TM)2Duo, 1.33 GHz a 2 GB RAM. Bylo verifikováno téměř 36 000 asociačních pravidel, 44 z nich bylo pravdivých. Přehled deseti nejsilnějších je v Obrázku 4, ukázka detailního výstupu nejsilnějšího pravidla je v Obrázku 5.

Actual group of hypotheses: All hypothesis

Hypotheses in group: 44 Shown hypotheses: 44 Highlighted: 0

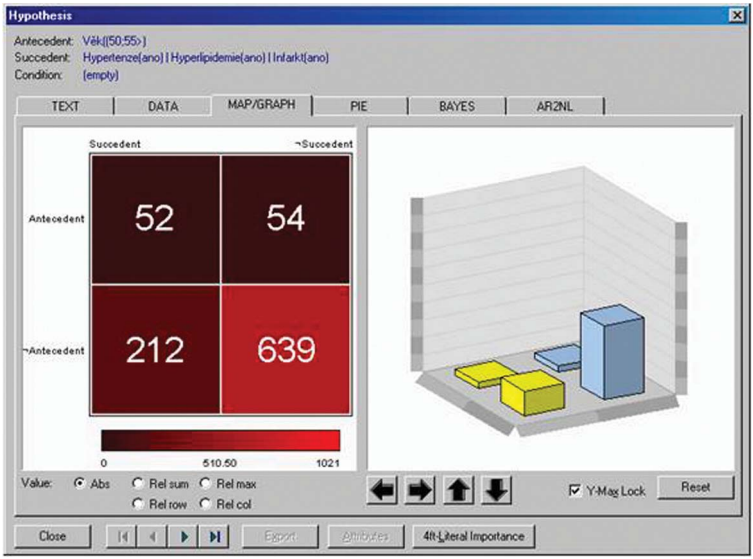
Nr.	Id	AvgDf	Hypothesis
1	40	0.778	Věk([50;55⟩) ••• Hypertenze(ano) | Hyperlipidemie(ano) | Infarkt(ano)
2	37	0.726	Věk([50;55⟩) ••• Diabetes(ano) | Hypertenze(ano) | Hyperlipidemie(ano)
3	38	0.724	Věk([50;55⟩) ••• Diabetes(ano) | Hypertenze(ano) | Hyperlipidemie(ano) | Infarkt(ano)
4	39	0.540	Věk([50;55⟩) ••• Diabetes(ano) | Hypertenze(ano) | Infarkt(ano)
5	9	0.434	Stav(ženatý) & Věk([45;50⟩, [50;55⟩) & Vzdělání(základní, vyučen) ••• Hypertenze(ano) | Infarkt(ano)
6	5	0.407	Stav(ženatý) & Věk([45;50⟩, [50;55⟩) & Vzdělání(základní, vyučen) ••• Diabetes(ano) | Hypertenze(ano) | Infarkt(ano)
7	13	0.406	Stav(ženatý) & Věk([45;50⟩, [50;55⟩) & Vzdělání(vyučen) ••• Diabetes(ano) | Hypertenze(ano) | Infarkt(ano)
8	29	0.398	Věk([45;50⟩, [50;55⟩) & Vzdělání(středoškolské, vysokoškolské) ••• Diabetes(ano) | Hyperlipidemie(ano) | Infarkt(ano)
9	6	0.391	Stav(ženatý) & Věk([45;50⟩, [50;55⟩) & Vzdělání(základní, vyučen) ••• Hypertenze(ano)
10	42	0.387	Vzdělání(vysokoškolské) & Aktpozam (mírná) & Telaktza (převážně sedí) ••• Hypertenze(ano)

Obrázek 4 — Deset nejsilnějších asociačních pravidel nalezených dle zadání v Obrázku 3

Nejsilnější pravidlo je pravidlo

$$\text{Věk}(50;55) \Rightarrow^{+}_{0.78,\,52} \text{Hypertenze(ano)} \vee \text{Hyperlipidemie(ano)} \vee \text{Infarkt(ano)},$$

které vychází z toho, že mezi muži ve věku 50 až 55 let je relativní četnost



Obrázek 5 – Ukázka detailu nejsilnějšího asociačního pravidla z Obrázku 4

přítomnosti alespoň jednoho rizika z Hypertenze, nebo Hyperlipidemie nebo Infarkt rovna $\frac{52}{52+54}=0.490$ a relativní četnost přítomnosti alespoň jednoho z těchto rizik mezi všemi pacienty v matici dat.

Vstup je $\frac{52+212}{52+54+212+639}=0.276$. To znamená, že relativní četnost těchto rizik je mezi muži ve věku 50 až 55 let je o 78 procent vyšší než v celé matici.

Poznamenejme, že v matici Vstup některé údaje chybí, proto je tedy celkový součet pacientů pro některá pravidla menší než 1417.

Podrobnější interpretace výsledků je však náročná, vyžaduje mimo jiné využití věcných znalostí, pokud například chceme oddělit překvapující výsledky. V každém případě však interpretace přesahuje rozsah tohoto příspěvku. Lze ale učinit závěr že takováto asociační pravidla pomůžou při orientaci v datech.

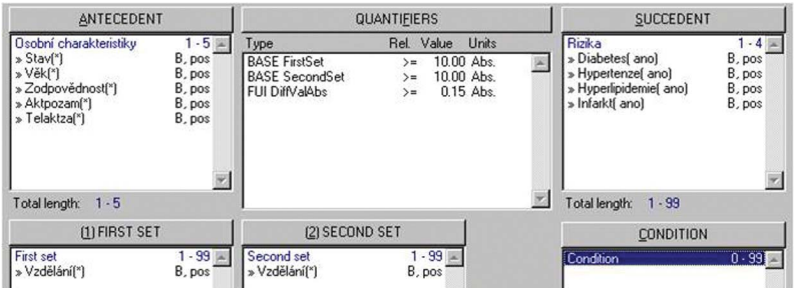
Jiný příklad kdy asociační pravidla pomohla k orientaci v neznámých datech se týká analýzy databáze katetrizací obsahující údaje o téměř třech tisících pacientů II. interní kliniky kardiologie a angiologie Všeobecné fakultní nemocnice v Praze. Aplikace procedury 4ft-Miner umožnila orientaci v těchto datech a vedla k následné aplikace dalších metod. Výsledkem bylo nalezení zajímavých vztahů týkajících se stenóz. Některé podrobnosti jsou uvedeny na adresách
http://euromise.vse.cz/dzmd/prehled/abstrakty/index.php?page=stochl
http://euromise.vse.cz/dzmd/prehled/abstrakty/index.php?page=mrazek

5. Hledání zajímavých rozdílů mezi skupinami pacientů

Jinou důležitou úlohou je hledání rozdílů skupinami pacientů. I při jejím řešení mohou pro orientaci v datech posloužit asociační pravidla. Na příkladu ukážeme využití procedury SD4ft-Miner [10] která je, stejně jako procedura 4ft-Miner, součástí systému LISp-Miner. Předpokládejme, že nás zajímá otázka.

Jsou v matici dat Vstup patrné nějaké rozdíly mezi pacienty s různým vzděláním ohledně vlivu kombinací hodnot atributů skupiny Osobní údaje na výskyt některých rizik nebo jejich kombinací?

Procedura SD4ft-Miner hledá dvojice asociačních pravidel, kriterium zajímavosti se týká celé dvojice. Použijeme zadání procedury dle Obrázku 6.



Obrázek 6 – Příklad zadání parametrů procedury SD4ft-Miner

Tímto zadáním hledáme dvojice podmíněných asociačních pravidel $(A \Rightarrow_{p1,a1} S) / V_1$ a $(A \Rightarrow_{p2,a2} S) / V_2$ takových, že

- $a_1 \geq 10$, viz výraz `Base FirstSet >= 10.00 abs` ve sloupci QUANTIFIERS
- $a_2 \geq 10$, viz výraz `Base SecondSet >= 10.00 abs` ve sloupci QUANTIFIERS
- $|\frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2}| \geq 0.15$ viz výraz `FUIDiffValAbs` ve sloupci QUANTIFIERS

Podmíněné asociační pravidlo $(A \Rightarrow_{p1,a1} S) / V_1$ se týká matice dat Vstup / V_1 která je tvořena těmi řádky, které splňují booleovský atribut V_1. Kontingenční tabulka atributů A a S v matici Vstup / V_1 je v tabulce 3. Podobně pro asociační pravidlo $(A \Rightarrow_{p2.a2} S) / V_2$.

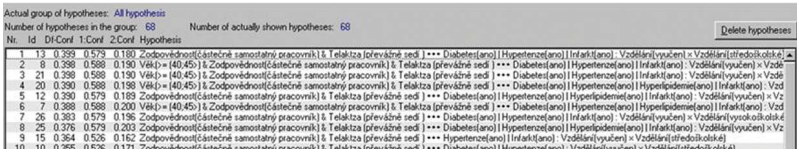
Vstup / V_1		S	Ø S
A	a1	b1	
Ø A	c1	d1	

Tabulka 3 – Kontingenční tabulka atributů A a S v matici dat Vstup / V_1

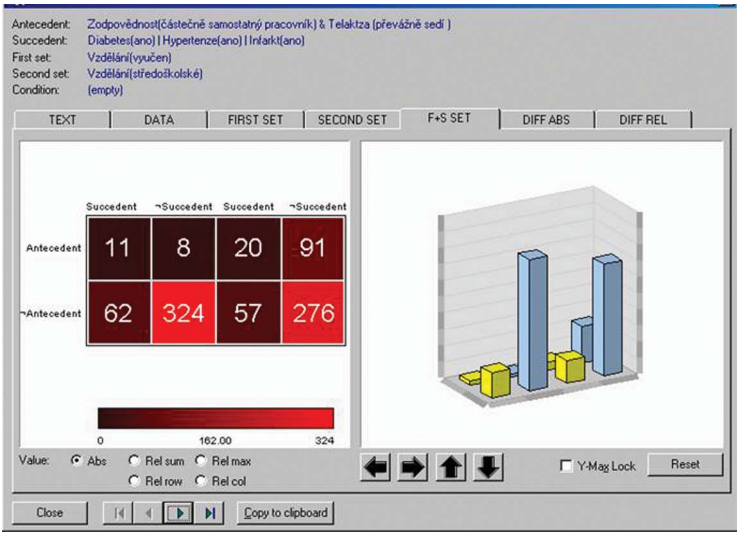
Přípustné booleovské atributy A jsou zadány ve sloupci ANTECEDENT v Obrázku 6, přípustné booleovské atributy S ve sloupci SUKCEDENT v Obrázku 6, podobně jako v Obrázku 3. Booleovské atributy V_1 a V_2 jsou zadány ve sloupcích FIRST SET a SECOND SET v Obrázku 6, vždy se jedná o jeden z atributů Vzdělání(základní), Vzdělání(vyučen), Vzdělání(středoškolské), Vzdělání(vysokoškolské).

Podmínka $|\frac{a_1}{a_1+b_1} - \frac{a_2}{a_2+b_2}| \geq 0.15$ znamená, že rozdíl konfidencí obou pravidel je nejméně 0.15.

Běh procedury SD4ft-Miner zadaný dle Obrázek 6 trval 9 vteřin na PC s procesorem Intel(R) Core(TM)2Duo, 1.33 GHz a 2 GB RAM. Bylo verifikováno téměř 274 000 dvojic asociačních pravidel, 68 z nich splňovalo zadaná kriteria. Přehled deseti dvojic s největším rozdílem konfidencí je v Obrázku 7, ukázka detailního výstupu pro dvojici s největším rozdílem je v Obrázku 8.



Obrázek 7 – Deset dvojic pravidel s největším rozdílem konfidencí dle zadání v Obrázku 6



Obrázek 8 – Ukázka detailu dvojice pravidel z Obrázku 7 s největším rozdílem konfidencí

V Obrázku 8 je dvojice podmíněných pravidel

Kombinace $\Rightarrow_{0.58,11}$ *Rizika* / Vzdělání(vyučen) a

Kombinace $\Rightarrow_{018,20}$ *Rizika* / Vzdělání(středoškolské) kde

Kombinace znamená Zodpovědnost(částečně samostatný pracovník) $\wedge$ Telaktza(převážně sedí) a

Rizika znamená Diabetes (ano) $\vee$ Hypertenze(ano) $\vee$ Infarkt(ano).

Konfidence prvního pravidla je 0.58 což znamená, že v daných datech je mezi vyučenými pacienty splňujícími Zodpovědnost(částečně samostatný pracovník) a zároveň Telaktza(převážně sedí) 58 procent pacientů ohrožených alespoň jedním z rizik Diabetes(ano), Hypertenze(ano), Infarkt(ano). Konfidence druhého pravidla je 0.18 což znamená, že v daných datech je mezi pacienty se středoškolským vzděláním splňujícími Zodpovědnost(částečně samostatný pracovník) a zároveň Telaktza(převážně sedí) pouze 18 procent pacientů ohrožených alespoň jedním z rizik Diabetes(ano), Hypertenze(ano), Infarkt(ano).

Podrobnější interpretace uvedených výsledků procedury SD4ft-Miner však přesahuje rozsah tohoto příspěvku.

6. Závěr

Naznačili jsme důležité rysy dobývání znalostí z databází a ukázali tři jednoduché příklady aplikací. Je třeba zdůraznit, že se jedná pouze o náznak možností aplikací dobývání znalostí z databází pro analýzu medicínských dat. Další informace lze získat například v [11].

Poznámky

(1.) Projekt STULONG byl realizován na II. interní klinice 1. lékařské fakulty Univerzity Karlovy a ve Všeobecné fakultní nemocnice v Praze pod vedením Prof. MUDr. F. Boudíka, DrSc. ve spolupráci s MUDr. M. Tomečkovou, CSc. a Prof. MUDr. J. Bultasem, CSc. Data byla převedena do elektronické podoby Evropským centrem pro medicínskou informatiku, statistiku a epidemiologii Univerzity Karlovy a Akademie věd ČR pod vedením Prof. RNDr. Jany Zvárové,DrSc..

Literatura

[1.] Hand, D., Manilla, H. and Smyth, P. (2001). Principles of Data Mining, MIT

[2.] Berka, P. (2003). Dobývání znalostí z databází. Academia, Praha,

[3.] Šimůnek, M. (2003). Academic KDD Project LISp-Miner. In Abraham A et al (eds), Advances in Soft Computing - Intelligent Systems Design and Applications, Springer, Berlin Heidelberg New York,

[4.] Agrawal, R., Imielinski, T. and SWAMI, A. (1993). Mining associations between sets of items in massive databases. In: Proc. of the ACM-SIGMOD 1993 Int. Conference on Management of Data. 207-216, Washington D.C..

[5.] Xindong, W. et al. (2008). Top 10 Algorithms in Data Mining. Knowledge and Information Systems, Vol. 14 1 – 37

[6.] Hájek P. (2004). Metoda GUHA v minulém století a dnes. In: Snášel V. (ed.) ZNALOSTI 2004. VŠB TU, Ostrava

[7.] Hájek,P., Havránek,T. and Chytil,M.K. (1983). Metoda GUHA: Automatická tvorba hypotéz. Academia, Praha

[8.] Hájek,P. and Havránek,T. (1978). Mechanising Hypothesis Formation- Mathematical Foundations for a General Theory, Springer-Verlag: Berlin - Heidelberg - New York,

[9.] Rauch, J. and Šimůnek, M. (2005). An Alternative Approach to Mining Association Rules. In: Lin, T. Y. et al. (eds) Data Mining: Foundations, Methods, and Applications, 219 – 238, Springer-Verlag, Berlin - Heidelberg - New York

[10.] PRICKL Rauch, J. and Šimůnek, M. (2009). Dealing with Background Knowledge in the SEWEBAR Project. In: Berendt B. et al.: Knowledge Discovery Enhanced with Semantic and Social Information, 89 – 106, Springer-Verlag, Berlin - Heidelberg - New York

[11.] Berka, P. and Rauch, J. (2009) Dobývání znalostí V databázích. In: Zvárová, J., Svačina, Š. and Valenta, Z. Biomedicínská informatika – Systém pro podporu lékařského rozhodování, 63–119, UK, Praha

Prohlášeni:

Vytvoření tohoto příspěvku bylo podporováno projektem 1M06014 Ministerstva školství, mládeže a tělovýchovy ČR.

Kontakt:

Doc. RNDr. Jan Rauch, CSc.
Centrum biomedicínské informatiky
Ústav informatiky AV ČR, v.v.i.
Pod Vodárenskou věží 2, 182 07 Praha 8
rauch@vse.cz