Univerzita Karlova v Praze
Filozofická fakulta

Diplomová práce

2014 Martin Senft

Univerzita Karlova v Praze
Filozofická fakulta
Ústav knihovnictví a informačních studií

Diplomová práce

Martin Senft

Vytěžování databáze Poradny pro poruchy metabolismu

Data mining of the database of Consulting centre for metabolism disorders

vedoucí: prof. RNDr. Jiří Ivánek, CSc.

2014

Chtěl bych tímto poděkovat MUDr. Romanu Cibulkovi, PhD., MBA za poskytnutí dat a zhodnocení výsledků. Mým rodičům bych chtěl poděkovat za veškerou pomoc a podporu. V neposlední řadě bych chtěl poděkovat prof. RNDr. Jiřímu Ivánkovi, CSc. za trpělivé vedení mé práce.

Martin Senft

Prohlašuji, že jsem diplomovou práci vypracoval samostatně, že jsem řádně citoval všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Plzni dne 30.3.2014

Abstrakt:

Tato práce aplikuje data miningovou metodu rozhodovacích pravidel na data z Poradny pro poruchy metabolismu Fakultní nemocnice Plzeň. Jako nástroj slouží systém LISp-Miner, vyvinutý na VŠE Praha. Nalezená rozhodovací pravidla jsou zhodnocena s odborníkem.

Základní části této práce jsou: souhrn hlavních data miningových metod a metod pro hodnocení výsledků. Dále pak popis aplikace data minigu a popis a zhodnocení výsledků.

Klíčová slova: data mining, získávání znalostí z databází, rozhodovací pravidla, metoda GUHA, LISp-Miner,

Abstract:

This thesis applies the data mining method of decision rules on data from Consulting centre for Metabolism disorders from University hospital Pilsen. As a tool is used the system LISp-Miner, developed at University of Economics, Prague**.** Decision rules found are evaluated by a specialist.

The main parts of this thesis are followings: an overview on main data mining methods and results evalutation methods, description of the data mining method application on data and description and evaluation of results.

Key words: data mining, knowledge discovery from databases, decision rules, GUHA, LISp-Miner,

Úvod

Jedním z významných fenoménů současnosti je velmi rychle se zvětšující množství dat uchovávané v organizacích. Přeměna těchto dat na využitelné informace a znalosti patří mezi velké výzvy našeho století.

Cílem této práce byla ukázka aplikace data miningu na reálných datech. Za tímto účelem jsem zpracoval lékařské záznamy 115 pacientů z Poradny pro metabolické poruchy Fakultní nemocnice Plzeň.

Pro vytěžení znalostí z tohoto souboru jsem použil metodu GUHA, implementovanou v systému LISp-Miner. Systém LISp-Miner je akademický software vyvinutý na Vysoké škole ekonomické v Praze. Výběr a zpracování vstupních dat a zhodnocení konečných výsledků jsem provedl v rámci řízeného rozhovoru s lékařem. Součástí práce je i přehled metod data miningu. Tento přehled jsem zpracoval podle souhrnné publikace *Dobývání znalostí z databází* Petra Berky (2003).

Má práce se skládá z teoretické části (kapitoly 1-5) a praktické části (kapitoly 6,7). V teoretické části je souhrn metod data miningu a souhrn metod na vyhodnocení výsledků data miningu. Podrobněji jsem zde také popsal metodu GUHA. Následně je zde kapitola věnovaná nástrojům data miningu a systému LISp-Miner.

Praktická aplikace data miningu odpovídá metodice CRISP-DM, která má následující kroky: porozumění problematice (viz kapitola 6), porozumění datům (viz sekce 6.2.), příprava dat (viz sekce 6.3.), modelování (viz sekce 7.1.), vyhodnocení výsledků (viz sekce 7.2. - 7.4.10.), využití výsledků (viz sekce 7.5.).

Počet znaků: 120 283

Obsah

Seznam zkratek použitých v textu

Zkratky z lékařského záznamu:

ALT - alaninaminotransferáza

Apo-A - apolipoprotein A

Apo-B - apolipoprotein B

BMI - body mass index

CK - kreatinkináza

CRP - C-reaktivní protein

GHBA - glykohemoglobin

GMT - gama-glutamyltransferáza

HDL - HDL cholesterol

CHOL - cholesterol

KMOC - kyselina močová

KREA - kreatinin

LDL - LDL cholesterol

Lpa - lipoprotein malé a

MALB - mikroalbuminurie

OA - osobní anamnéza

RA - rodinná anamnéza

TG - triglyceridy

TSH - tyreotropin

Ostatní zkratky

API - application programming interface

CBR - case based reasoning

DEA - data envelopment analysis

FN Plzeň - Fakultní nemocnice Plzeň

KVO - kardiovaskulární onemocnění

PMML - predictive model markup language

ROC křivka - receiver operating characteristic

SVM algoritmus - support vector machine algoritmus

TDIDT - top down induction of decision tree algoritmus

1. Metody data miningu - modelování

Data mining definuje Berka (2003, s. 15) jako *„netriviální získávání implicitních, dříve neznámých a potencionálně užitečných informací z dat.“* Šarmanová (2002, s. 165) definuje metody data miningu s určitou nadsázkou jako metody pro vytváření hypotéz, zatímco statistické metody jsou zaměřené na potvrzování hypotéz.

Existuje celá řada metod, jak z dat vytvořit znalosti. Jednotlivé metody se pak liší nejen podle toho, jak reprezentují vytvořené znalosti, ale i tím pro jaký typ dat a úloh se hodí, jak složité vztahy dokáží reprezentovat, jak jsou tyto vztahy srozumitelné pro uživatele a jak jsou použitelné pro klasifikaci nových případů. (Berka 2003, s. 85)

Rozdělení metod data miningu podle typů úloh nabízí například Skalská (2010, s. 27). Metody dělí na explorační, verifikační a vizualizační. Cílem exploračních metod je nalézt v datech zákonitosti, netypické objekty nebo zcela neznámé a nepředpokládané souvislosti. Cílem verifikačních metod je pak zhodnotit správnost výsledků verifikačních metod. Vizualizační metody pak mají za cíl lepší porozumění výsledkům data miningu. Mimo tyto tradiční úlohy staví Skalská (2010) nové směry jako je text mining nebo web mining (Skalská 2010, s. 28-29).

Pro změnu Šarmanová (2002, s. 166) zmiňuje tři základní uživatelské skupiny data miningových metod. Jsou to průzkum (bankovnictví, marketing apod.), výzkum a sociologický výzkum. Rud a Magera (2001, s. 4-9) uvádějí mezi nejčastějšími úlohami „průzkumu“ analýzu profilu zákazníků, segmentaci zákazníků, zjišťování pravděpodobné odezvy, rizika a aktivace ze strany zákazníků, pravděpodobnosti odlivu zákazníků, pravděpodobnost následného nebo křížového prodeje nebo výpočet pravděpodobné čisté současné hodnoty produktu, resp. hodnoty produktu po dobu existence.

Následující sekce jsem zpracoval podle knihy Petra Berky *Dobývání znalostí z databází* (2003). Berka zde rozděluje nejpoužívanější metody pro data mining podle toho, s jakým typem informace pracují. Jsou to metody symbolické, subsymbolické a metody založené na analogii. K metodám data miningu se někdy přidávají také statistické metody diskriminační analýza, regresní analýza nebo shluková analýza - viz Šarmanová (2002) nebo Cios (2007). Ty ale Berka nechává stranou.

Jaké metody do zmíněných tří skupin patří? Symbolické metody v podstatě prohledávají prostor možných řešení. Patří mezi ně hlavně rozhodovací stromy, asociační a rozhodovací pravidla. Učení pomocí subsymbolických metod je možné chápat jako aproximaci nějaké funkce. Mezi tyto metody patří neuronové sítě, genetické algoritmy a bayesovské metody. Metody založené na analogii chápou učení jako zapamatování si typických příkladů. (Berka 2003, s. 85-88)

1.1. Rozhodovací stromy

Reprezentace znalostí pomocí rozhodovacích stromů je hojně používána v řadě oborů. Indukce rozhodovacích stromů probíhá následujícím způsobem. Data se rozdělují postupně na podmnožiny a to tak dlouho, až vzniknou podmnožiny s příklady jedné třídy. Jde tedy o metodu specializace shora dolů.[1] Cílem je nalézt pokud možno konzistentní strom s daty, který bude zároveň co nejmenší a nejjednodušší. (Berka 2003, s. 86)

Klíčová otázka celého postupu je, jak vybrat vhodný atribut pro větvení stromu. K dispozici je celá řada kritérií pocházejících z teorie informace a teorie pravděpodobnosti. Příkladem je třeba entropie, informační zisk, poměrný informační zisk atd. (Berka 2003, s. 87-90)

Pokud máme data s numerickými atributy, je nutné rozdělit obor hodnot na intervaly. V nejjednodušším případě se provádí rozdělení na dva intervaly (binarizace), kdy je pomocí entropie nalezen dělící řez (cut).

Problém chybějících hodnot je možné řešit následujícími způsoby. Chybějící hodnota je nahrazena nejčastější hodnotou daného atributu. Z příkladu s chybějící hodnotou se vytvoří více příkladů. Místo chybějící hodnoty se postupně doplňují hodnoty atributu s váhami. Tyto váhy jsou relativní četnosti těchto hodnot v datech. (Berka 2003, s. 98) Dalšími řešeními je chybějící hodnoty ignorovat, nahradit je hodnotou „nevím“ nebo libovolnou hodnotou. (Berka 2003, s. 147)

Rozhodovací stromy mohou také sloužit jak k rozdělování objektů do tříd, tak k odhadování hodnot nějakého numerického atributu (regresní stromy). Kritérium pro volbu atributu u regresních stromů není informační zisk, ale redukce směrodatné odchylky. (Berka 2003, s. 99)

Asi nejznámějšími algoritmy patřícími do této skupiny metod je vedle zmíněného TDIDT, algoritmus systému CART a C4.5 (Wu et al., 2008)

1.2. Asociační pravidla

Berka (2003, s. 102-155) rozlišuje dva typy pravidel: asociační a rozhodovací. Asociační pravidla hledají souvislosti mezi libovolnými kategoriemi bez toho, aby jednu kategorii upřednostňovaly jako závěr. Rozhodovací pravidla mají naopak jasně vymezený cílový atribut. O rozhodovacích pravidlech pojednává další sekce.

Základní charakteristiky

Pravidla můžeme charakterizovat podle toho, kolik z nich splňuje platnost předpokladu a kolik z nich platnost závěru, případně kolik z nich splňuje platnost obou částí (předpoklad i závěr) současně. Přehledně lze všechny tyto hodnoty zobrazit v kontingenční tabulce viz tabulka č. 1. (Berka 2003, s. 136)

1 Tento postup je známy pod anglickým názvem top down induction of decision trees (TDIDT).

	závěr platí	závěr neplatí
předpoklad platí	a	b
předpoklad neplatí	c	d

tabulka č. 1 čtyřpolní tabulka četností

Nejčastěji používanými charakteristikami asociačních pravidel jsou:

- *podpora* (support) – absolutní, popř. relativní počet případů splňujících předpoklad i závěr (a),
- *spolehlivost* (confidence) – podmíněná pravděpodobnost závěru, pokud platí předpoklad. (a/a+b),
- *absolutní, popř. relativní počet případů splňujících předpoklad* (a+b),
- *absolutní, popř. relativní počet případů splňujících závěr* (a+c),
- *pokrytí* (coverage) – podmíněná pravděpodobnost předpokladu, pokud platí závěr (a/a+c),

Generování kombinací

Základem všech algoritmů pro hledání asociačních pravidel je generování kombinací hodnot atributů. Metody generování kombinací je možné rozdělit na generování do šířky, generování do hloubky a heuristické metody.

Při generování do šířky se nejprve vygenerují všechny kombinace o délce 1, pak všechny kombinace o délce 2 atd. Hodnoty atributů jsou přitom řazeny podle abecedy.

Při generování do hloubky se vyjde od první kombinace délky 1. Ta se pak prodlužuje vždy o první kategorii dalšího atributu atd. Pokud již kombinace nelze prodloužit, hodnota posledního atributu se změní na další hodnotu. Hodnoty atributů jsou opět uspořádány podle abecedy.

Heuristikou se v procesu vytváření kombinací rozumí kombinace, která má nejčastější výskyt. Heuristické generování kombinací tedy spočívá ve vytvoření kombinací a jejich následné seřazení podle jejich výskytu v datech. (Berka 2003, s. 106-107).

Generování kombinací je výpočetně náročný proces. Obecně lze říci, že počet kombinací je exponenciálně závislý na počtu atributů. Většinou se proto při data miningu začíná od kombinací menších délek. (Berka 2003, s. 107-109).

Algoritmus Apriori

Nejznámějším algoritmem pro generování pravidel je algoritmus Apriori. Algoritmus se skládá ze dvou částí. Nejdříve se hledají kombinace hodnot kategorií, které splňují předem zadanou četnost (podporu). V druhé části se pak z množiny nalezených kombinací vytváří pravidla, která splňují

předem zadanou míru spolehlivosti. V následujících odstavcích je ukázáno, jak je možné tento algoritmus optimalizovat.

Kombinace délky K se vytváří spojováním kombinací délky K-1. Kombinace délky K splňuje předem zadanou četnost pouze tehdy, pokud všechny její pod-kombinace také splňují tuto četnost. Při generování nám tedy pomůže předešlá znalost četnosti kratších kombinací.

Z nalezených kombinací se nyní vytváří asociační pravidla tak, že se rozdělují na dvě části antecedent - sukcedent. Úkolem je rozlišit, která z těchto „rozdělení“ splňují zadanou míru spolehlivosti. Pokud pravidla s „delším“ antecedentem nesplňují zadanou míru spolehlivosti, nebudou ji splňovat ani pravidla s „kratším“ antecedentem. Dále pravidla s kratším sukcedentem musí splňovat předem zadanou míru spolehlivosti, aby tuto míru splňovala také pravidla s delším sukcedentem. Opět pomůže znalost již provedených kroků. Spolehlivost se může také snáze vypočítat z již známých četností z první fáze hledání kombinací. (Berka 2003, s. 109-111).

Zvláštní kapitolou při vytváření pravidel je vytváření takzvaných zobecněných pravidel. To vychází z faktu, že atributy se mohou sdružovat do obecnějších skupin různé míry obecnosti. Vzniklé skupiny tak tvoří jakousi taxonomii. Problém při vytváření obecnějších pravidel, je vlastně problém nastavení míry podpory, kterou musí tato pravidla splňovat. Pokud se nastaví míra podpory příliš vysoká, budou chybět atributy, které jsou v taxonomii níže. Pokud se naopak nastaví míra podpory příliš nízká, dojde k „explozi“ kombinací. (Berka 2003, s. 111-113)

Většina algoritmů na vytváření asociačních pravidel pracuje s jednou tabulkou, která se vytvoří z více tabulek ve fázi předzpracování dat. (Berka 2003, s. 116)

1.3. Rozhodovací pravidla

Zatímco asociační pravidla slouží k hledání zajímavých souvislostí, rozhodovací pravidla slouží ke klasifikaci. Formální tvar rozhodovacího pravidla je:

IF *Ant* THEN *Class*

Rozhodovací pravidla lze získat například pomocí převedení rozhodovací stromu na pravidla. (Berka 2003, s. 130-134).

Asi nejznámějším algoritmem pro vytváření rozhodovacích pravidel je algoritmus pokrývání množin.[2] V prostoru atributů se postupně odebírají oblasti, které obsahují příklady pouze jedné třídy. Základní schéma algoritmu je následující.

2 Někdy se tomuto algoritmu také říká algoritmus „odděl a panuj“, na rozdíl od TDIDT algoritmu, který se nazývá algoritmem „rozděl a panuj.“

1. Najdi pravidlo, které pokrývá nějaké pozitivní příklady a žádný negativní.
2. Odstraň pokryté příklady z trénovací množiny.
3. Pokud v trénovací množině zbývají nějaké nepokryté příklady, vrať se k bodu 1, jinak skonči.

Klíčovým bodem algoritmu je krok číslo 1, tj. nalezení jednoho pravidla. Existují dvě varianty řešení:

- generalizace (odebírání kategorií z kombinace),
- specializace (přidávání kategorií do kombinace),

Pokud používáme generalizaci, je možné postupovat v trénovacích příkladech sekvenčně a jako první pravidlo vzít první pozitivní příklad. Následně toto pravidlo zkracovat, dokud jsou pokrývány pozitivní příklady, pak se vezme první následující zbylý pozitivní příklad jako další pravidlo atd.

Pokud se používá specializaci, nejčastěji se používá hledání gradientní (doplňuje se vždy „nejlepší“ kategorie „do hloubky“ bez navracení) nebo hledání paprskovité (dosazuje se vždy nejlepší kategorie ze seznamu nejlepších potencionálních předpokladů) (Berka 2003, s. 130-134). Charakteristikami kvality pravidel jsou negativní entropie, Laplaceův odhad očekávané spolehlivosti nebo m-odhad. Jde o míry vycházející z charakteristiky spolehlivost (Berka 2003, s. 135).

Seznam vzniklých pravidel může být buď neuspořádaný, nebo uspořádaný (tzv. rozhodovací seznam). Pak jde o strukturu typu:

IF *Ant1* THEN *Class1*,

ELSE IF *Ant2* THEN *Class2*,

ELSE IF *Ant3* THEN Cl*a*ss...

Hlavní cyklus algoritmu se upravuje podle toho, zda jde o to najít uspořádaný nebo neuspořádaný soubor pravidel. Pokud se vytváří neuspořádaný soubor pravidel, hledají se pravidla po jednotlivé třídy odděleně. Pokud se vytváří uspořádaný soubor pravidel (rozhodovací seznam), hledají se pravidla ke všem třídám najednou. (Berka 2003, s. 135-137)

1.4. Neuronové sítě

Neuronové sítě představují výpočetní model umožňující distribuované paralelní zpracování hlavně numerických dat. Základní jednotkou neuronových sítí je tzv. umělý neuron, který přijímá „podněty“ (numerické hodnoty), nelineárně je transformuje pomocí různých přenosových funkcí a předává dál ve formě výstupních hodnot. Jak napovídá název, tento model má za vzor biologické struktury. (Berka 2003, s. 157-160)

Zjednodušeně lze jeden neuron popsat takto. Numerické vstupy jsou násobené váhou a tento součin vstupuje do celkového součtu. Pokud tento součet přesáhne definovaný práh, neuron se aktivuje (vyšle podnět). Vstupy a výstupy neuronu mohou být binární nebo spojité. (Berka 2003, s. 158)

Popsaný neuron používá tzv. „skokovou“ přenosovou funkci. Nejpoužívanějšími tzv. „hladkými“ přenosovými (aktivačními) funkcemi jsou (Berka 2003, s. 159-160):

- sigmoidální funkce,
- hyperbolický tangens,

Neuronové sítě dokáží sami nastavit váhy ve svých neuronech na základě trénovacích dat. Mluví se zde o schopnosti neuronových sítí učit se. Metod, resp. algoritmů, jak toto učení probíhá, je celá řada. Příkladem může být tzv. gradientní metoda. Ta funguje tak, že se na začátku nastaví váhy neuronu a s těmi jsou zpracována trénovací data. Na základě porovnání trénovacích dat a právě vzniklé klasifikace se vypočítá tzv. střední kvadratická chyba. Data se pak opakovaně prochází tak dlouho, dokud není střední kvadratická hodnota minimální nebo dostatečně malá. Pro opakované procházení dat se používá parametr learning rate, který udává „krok“, kterým se mění váhy (Berka 2003, s. 163).

Neurony se spojují do sítí, které se skládají z vrstev. Existuje celá řada typologií sítí, které se od sebe liší počtem vrstev a typem vztahů, které mezi sebou mají jednotlivé neurony. Příkladem může být trojvrstevná síť typu Perceptron obsahující: vstupní vrstvu, skrytou vrstvu a výstupní vrstvu. Tří vrstevný model umožňuje totiž aproximovat libovolnou spojitou funkci. (Berka 2003, s. 173-176)

Neuronové sítě se využívají především u numerických atributů. Kategoriální atributy se před použitím binarizují. Tato metoda se používá jak ke klasifikaci, tak k predikci (cena akcí, spotřeba elektrické energie apod.). Problémem neuronových sítí je interpretace. Znalosti získané neuronovými sítěmi jsou totiž pro uživatele naprosto nesrozumitelné. (Berka 2003, s. 334-337)

Pravděpodobně nejznámější algoritmus patřící do rodiny neuronových sítí je algoritmus SVM (support vector machine) (Wu et al., 2008)

1.5. Evoluční algoritmy

Evoluční algoritmy je obecný název pro řešení, která se inspirují v biologické evoluci. Vychází z myšlenky, že evoluce představuje optimalizační úlohu. Stejně tak je hlavní využití těchto algoritmů v optimalizaci. Hlavními pojmy evolučních algoritmů jsou tzv. jedinci, generace, fitness funkce, selekce, křížení a mutace. (Berka 2003, s. 176)

Na začátku každého evolučního algoritmu je třeba definovat fitness funkci a zakódovat řešení.

Jednotlivá řešení se nejčastěji zakódují jako bitové řetězce. Jednotlivé atributy jsou reprezentovány takovým počtem bitů, kolik nabývají hodnot. Fitness funkce je hodnota, která udává, jak dobré je dané řešení vzhledem k našemu úkolu. Pokud je našim úkolem například nalezení co nejspolehlivějších pravidel, je funkce shodná s charakteristikou spolehlivosti. (Berka 2003, s. 180-181)

Algoritmus prochází tzv. generacemi. V každé generaci dojde k selekci, křížení a mutaci. Selekce představuje výběr nejlepších jedinců (řešení) z hlediska fitness funkce. Křížení znamená to, že ze dvou řešení (rodičů) vzniknou dvě řešení (děti). Jde o velký zásah do řešení. Mutace naopak představuje náhodnou změnu jednoho bitu. Generace se většinou opakují tak dlouho, dokud daná řešení nedosáhnou předem definovaného optima. (Berka 2003, s. 178-179)

1.6. Bayesovská klasifikace

Bayesovská klasifikace vychází z Bayesovy věty o podmíněné pravděpodobnosti. Základními pojmy jsou apriorní pravděpodobnost (četnost získaná z učení s učitelem), aposteriorní pravděpodobnost (podmíněná pravděpodobnost). Při klasifikaci se hledá hypotéza (třída), která má pro danou evidenci maximální aposteriorní pravděpodobnost. (Berka 2003, s. 182-184)

Co dělat, pokud je evidencí víc? První řešením je koncept bayesovský naivní klasifikátor. Druhým řešením je koncept bayesovské sítě. Naivní bayesovský klasifikátor vychází z představy, že jednotlivé evidence jsou navzájem nezávislé. Snadno lze tedy vypočítat aposteriorní pravděpodobnost hypotézy při platnosti všech evidencí. (Berka 2003, 184-188)

Bayesovské sítě představují grafickou faktorizaci pravděpodobnostních modelů, kdy jejich cílem je kompaktní reprezentace společného pravděpodobnostního rozložení všech proměnných s využitím známých podmíněných nezávislostí. Jde o acyklicky orientované grafy, kde uzly reprezentují proměnné a hrany podmíněné nezávislosti mezi těmito proměnnými. (Berka 2003, s. 187-189)

Pokud existuje evidence a jsou známy pravděpodobnostní vztahy s dalšími proměnnými, je možné pomocí různých algoritmů provádět celou řadu abduktivních a deduktivních závěrů. Tato inference může být kauzální, diagnostická, interkauzální nebo smíšená. Podle typu zvoleného algoritmu, může být exaktní nebo aproximační. (Berka 2003, s. 187-189)

Jak patrné, bayesovské sítě v sobě obsahují znalosti dvou typů: o vztazích mezi atributy a o pravděpodobnostech atributů. Učení bayesovských tedy probíhá jak po rovině struktury, tak po úrovně parametrů. Podle typu úlohy je k dispozici celá řada algoritmů. (Berka 2003, s. 191-195)

Bayesovské sítě mají uplatnění v expertních systémech používaných například v bioinformatice, analýze vzorů nebo medicíně. Jejich použití má ale také tradici v modelování dat například v epidemiologii nebo sociálních vědách. (Berka 2003, s. 196-197)

Nejznámějším algoritmem z rodiny bayesovských sítí je algoritmus EM (Wu et al., 2008).

1.7. Metody založené na analogii

Základní charakteristiky těchto metod jsou následující:

- ve fázi učení nedochází ke generalizaci,
- klasifikace pomocí těchto metod je založena na podobnosti,
- příklady jsou chápány jako body v n-rozměrném prostoru (učení založené na instancích), nebo jako složité struktury, rámce (usuzování založené na příkladech), (Berka 2003, s. 206)

Problémy, které je třeba u těchto metod vyřešit, jsou následující: Jak zvolit příklady do databáze příkladů. Jak tyto příklady ukládat. Jak určit (měřit) podobnost příkladů. Podle jakého postupu provést klasifikaci nových příkladů. Asi nejznámějším algoritmem patřící do skupiny těchto metod je algoritmus k-NN (Wu et al., 2008).

Výběr příkladů

Berka (2003, s. 204) uvádí tři možnosti:

- Ukládá se každý příklad z databáze i příklady zatížené šumem.
- Nejdříve se příklady klasifikují pomocí systému a do databáze se ukládají pouze chybně zařazené příklady. Zde je počet uložených příkladů menší, ale je zde problém se šumem.
- Používá se složité kritérium založené na souhrnné správnosti klasifikace. Příklady dobře klasifikující se v databázi nechají, špatně klasifikující příklady se naopak z databáze vyškrtnou.

Trochu jiná situace je u systémů příkladového usuzování (case-based reasoning, dále také jako CBR). Zde nejsou příklady reprezentovány například řádkem v tabulce, ale komplikovanými strukturami. Výběr příkladů je iterativní proces. (Berka 2003, s. 210)

Ukládání příkladů

Výběr příkladů úzce souvisí s ukládáním příkladů v databázi. Proto, aby vyhledávání bylo dostatečně rychlé, se využívá celá řada indexačních technik. Může jít třeba o k-d stromy (atributy tvoří nelistové uzly) nebo IGTree (využívá kritérium informačního zisku). Během ukládání dat může probíhat také komprese. V případě velkého počtu příkladů se také neukládá každý příklad, ale

pouze tzv. centroidy. V nejjednodušším případě může jít o průměr z příkladů dané třídy.[3] (Berka 2003, s. 204-208)

Měření podobnosti

Měření podobnosti je klíčový koncept. Nejpoužívanější metriky jsou:

- eukleidovská vzdálenost,
- Hammingova vzdálenost,
- metrika překrytí („overlap“) (Berka 2003, s. 198).

Nevýhodou těchto metrik je to, že každý atribut má stejnou váhu, což vede k přeučení systému. Další nevýhodou je to, že metriky jsou příliš jednoduché a pouhá shoda či neshoda atributů nedokáže zachytit složitost problému. (Berka 2003, s. 199)

Na řešení zmíněných problémů se používají váhy a složitější metriky. Váhy atributů se zjišťují podle informačního zisku, pravděpodobnosti nebo inkrementálně podle výsledků klasifikace (podobně jako u neuronových sítí). Mezi pokročilejší metriky pak může patřit například metrika MVDM. Tato metrika bere do úvahy celkovou podobnost příkladů patřících do různých tříd v celé trénovací množině a zároveň využívá i váhy příkladů. (Berka 2003, s. 199-201)

Klasifikace nových příkladů

Jako příklad klasifikačního algoritmu může sloužit algoritmus k-nejbližších sousedů. Systém zde nalezne podle použité metriky k-nejbližších příkladů, které pak „hlasují“ o zařazení příkladu do třídy. Jako hlasování zde většinou slouží vážený průměr. (Berka 2003, s. 206)

1.8. Induktivní logické programování

U této metody nejsou příklady reprezentovány pomocí řádků tabulky databáze, ale pomocí predikátové logiky prvního řádu. Příklad je tak možné popsat libovolným počtem atributů, které mohou mít mezi sebou celou řadu vztahů. Většina algoritmů induktivního logického programování vychází z tradičních algoritmů dobývání znalostí. (Berka 2003, s. 211)

Výhody induktivního logického programování jsou (Berka 2003, s. 211):

- nalezené znalosti mají kompaktnější a srozumitelnější hodnotu,
- umožňuje řešit úlohy v oblastech vyžadujících reprezentaci více relací (např. analýza strukturálních nebo prostorových dat, (klasifikace chemických sloučenin),
- umožňuje využít doménových znalostí (v podobě předem známých pravidel), cílem tohoto typu učení je na základě příkladů a dílčí doménové znalosti, odvodit znalosti dokonalejší.

3 Není tak zcela pravda, že u metod založených na analogii nedochází vůbec ke generalizaci.

2. Metoda GUHA

Ve své práci jsem použil softwarový nástroj pro dobývání znalostí LISp-Miner (více viz kapitola 5) Ten je jednou z implementací metody GUHA. Tato kapitola stručně charakterizuje metodu GUHA, zvláště pak její proceduru 4ft-Miner.

Metoda GUHA (General Unary Hypotheses Automaton)[4] vznikla v šedesátých letech minulého století v týmu vedeném P. Hájkem. Od začátku byla chápána nejen jako algoritmus nebo softwarový nástroj, ale jako celistvý metodický systém. Řešení analytického problému bylo rozděleno do dílčích úkolů, např. vstup dat do stroje, strojové rozšíření dat o některé složené veličiny, vlastní generování hypotéz, tisk výsledků atd. Každý dílčí úkol byl pak řešen pomocí nějakého postupu. Pokud je takový postup prováděn, strojově je v terminologii GUHA nazýván procedura, ostatní postupy jsou nazývány pravidla. GUHA obsahuje čtyři typy postupů: vstupní, filtrační, samotné GUHA procedury a výstupní. (Hájek et al. 1983, s. 23-54). Samotné GUHA procedury jsou počítačové programy, které generují a následně verifikují nalezené struktury v datech. (Hájek et al. 2010, s. 37)

Od již zmíněného algoritmu Apriori se GUHA liší především větším počtem možností, které nabízí. GUHA představuje spojení celé řady metod generování asociačních pravidel a celé řady metod pro jejich statistické ověřování. Základní myšlenka metody GUHA je ostatně následující. V datech, které jsou k dispozici, najít všechny zajímavé vztahy, které mají zadanou logickou formu a jsou podporovány daty. (Hájek et al. 2010, s. 34)

Systém LISp-Miner vychází ze zkušeností z předchozích implementací a postupně budované teorie. Zároveň však systém LISp-Miner přináší celou řadu dalších možností v zadávání úloh, prezentaci výsledků atd. (Šimůnek 2010, s. 13)

Systém LISp-Miner je v současnosti složen z osmi GUHA procedur, dále pak procedury pro strojové učení KEX, modulu pro zpracování dat a z několika dalších modulů, které různým způsobem podporují analytické procedury. (Rauch 2011, s. 4) Ty umožňují jak generování asociačních pravidel zmíněných v předchozí kapitole, tak modelování dalších datových struktur následně verifikovaných pomocí různých kontingenčních tabulek. (Hájek et al. 2010, s. 36)

Ve své práci se systémem LISp-Miner jsem použil proceduru 4ft-Miner. Ta je vylepšenou implementací původní GUHA procedury ASSOC.

4 Místo pojmu asociační pravidlo se zde používá pojem hypotéza.

2.1. Procedura 4ft-Miner

Procedura 4ft-Miner pracuje s booleovskými atributy, které jsou odvozené z analyzované matice dat. Matice dat je odvozena z databázové tabulky. Počet řádků databázové tabulky se rovná počtu řádků matice dat.

V systému lze nastavit typ a délku tzv. koeficientu. Kategorie K je jedna hodnota z množiny všech hodnot, které může atribut A nabýt. Množina všech hodnot atributu A se nazývá koeficient. Na rozdíl od algoritmu Apriori, může být koeficientem podmnožina omezené délky, interval omezené délky nebo řez (interval obsahující krajní hodnotu). (Berka 2003, s. 121)

Základní booleovský atribut je dán dvojicí *atribut A(kategorie K*). Základní booleovský atribut je pravdivý v řádku matice, pokud atribut A nabývá hodnotu kategorie K. Booleovský atribut se skládá ze základních atributů pomocí logických spojek konjunkce, disjunkce a negace.(Rauch 2011, s. 5-6)

Pravidla mají v proceduře 4ft-Miner následující podobu:

Antecedent ~ Sukcedent/ Podmínka

Antecedent, sukcedent a podmínka jsou booleovskými atributy. Vlnovka značí libovolný zobecněný 4ft-kvantifikátor. Procedura může generovat asociační pravidla s i bez podmínky.

Kvantifikátor je v podstatě zobrazení z kontingenční tabulky na hodnoty 0, 1. (Berka 2003, s. 118) Pravidlo má hodnotu 1, tzn. je pravdivé, pokud hodnota funkce definující kvantifikátor splňuje zadané podmínky a naopak. (Berka 2003, s. 123) Kvantifikátor slouží k tomu, abychom mohli definovat, kolik objektů musí splnit nějaké pravidlo. Kvantifikátor tak určuje druh a sílu pravidla (Burda 2004, s. 6). Podle podobných (logických) vlastností je možné kvantifikátory v proceduře rozdělit na:

- implikační,
- dvojité implikační,
- ekvivalenční,
- symetrické asociační.

Podle způsobu výpočtu dělí Šimůnek (2010, s. 90) kvantifikátory na

- funkční – složitější než agregační kvantifikátory,
- agregační – jedna frekvence z příslušné tabulky četnosti.

Vstupy do procedury 4ft-Miner jsou analyzovaná data a jednoduchá definice velké množiny relevantních vzorů (patterns). Analyzovaná data jsou ve formě tabulky implementované v systému LISp-Miner. Definice relevantních vzorů probíhá pomocí definice množin relevantních antecedentů, sukcedentů, podmínek a definice kvantifikátoru.

K jemnému definování a manipulaci s literály nabízí procedura možnost vytvářet tzv. dílčí cedenty. Jde o podmnožiny literálů, ve kterých je možné hromadně definovat typy a délky koeficientů, to zda jsou literály ve vztahu konjunkce nebo disjunkce, maximální počet literálů atd. (Hájek et al. 2010, s. 38). Dílčí cedenty by ale měly mít význam i sémantický. Měly by odrážet obecně přijímané členění analyzované oblasti. (Rauch, 2013, s. 353)

Algoritmus procedury 4ft-Miner začíná pravidly o délce antecedentu 1. Postupně se počítají čtyřpolní tabulky a vyhodnocuje se funkce kvantifikátoru. Následuje prodlužování antecedentu, kdy se testuje opět pravdivost všech přípustných kombinací, a to až do maximálně zadané délky. (Burda 2004, s. 26)

Procedura 4ft-Miner nabízí tři pojetí pro práci s chybějícími hodnotami (Berka 2003, s. 129):

- **konzervativní**: Chybějící hodnoty se ignorují.
- **optimistické**: Chybějící hodnoty jsou brány jako podporující platnost daného konkrétního pravidla. Počet platných výskytů pravidla se tedy zvýší a tím se „zlepší“ i charakteristiky pravidla.[5]
- **pesimistické**: Opak optimistického pojetí, počet neplatných výskytů pravidla se zvýší.

Metoda GUHA se dále rozvíjí. Rauch (2013, s. 384) uvádí následující témata současného rozvoje:

- získávání zkušeností z aplikace metody,
- studium vzájemných závislostí výsledků jednotlivých procedur,
- shromáždění a využití doménových znalostí,
- automatizace běhu GUHA procedur,
- využití gridových technologií,
- rozvoj observačních kalkulů (speciální logické kalkuly),
- GUHA procedury pro multirelační data mining,
- aplikace fuzzy přístupů.

5 Charakteristikou je zde míněna hodnota funkce definující vybraný kvantifikátor.

Nejvýznamnějšími implementacemi metody GUHA je systém *LISp-Miner* a systém *Ferda Data Miner.* (Rauch 2013, s. 384)

2.2. Medicínské aplikace metody GUHA

Asi největší a nejznámější aplikaci metody GUHA na medicínská data představuje projekt STULONG. Projekt probíhal mezi lety 1979 až 1999. Cílem projektu bylo podrobnější studium rizikových faktorů arterosklerózy u mužů středního věku.

Datovým vstupem bylo 244 atributů u 1419 sledovaných pacientů z několika medicínských pracovišť. Z toho 219 atributů byly výsledky měření nebo číselné kódy laboratorních vyšetření. Data byla ze čtyř zdrojů: vstupních vyšetření (1419 pacientů s 244 atributy), kontrolní vyšetření (10 610 záznamů po dobu 20 let, každý záznam má 66 atributů), dotazník rozesílaný poštou (403 pacientů, 62 atributů), informace o úmrtí (389 pacientů, 5 atributů). (Projekt STULONG 2003).

Atributy u vstupních vyšetření byly rozděleny do následujících skupin: identifikační údaje, sociální charakteristiky[6], tělesná aktivita, kouření, pití alkoholu, cukr-káva-čaj, rodinná anamnéza, osobní anamnéza, dotazník A2[7], fyzikální vyšetření, biochemická vyšetření. U kontrolních vyšetření byla celá řada atributů popisujících změnu v životním stylu, abúzu, farmakoterapii, zaměstnání atd. Informace o úmrtí byly zaměřeny hlavně na příčinu úmrtí a čas úmrtí. (Dolejší 2002)

Projekt měl několik okruhů analytických otázek. Tyto okruhy vznikly na základě původního cíle projektu, potřeb a požadavků odborníků a vytvořeného systému analytických otázek. Jak tento systém fungoval? Jsou dva datové soubory *Vstupní* vyšetření a *Kontrolní* vyšetření. Z těchto dvou souborů vzniknou tři matice: *Vstupní-Vstupní, Vstupní-Kontrolní, Kontrolní-Kontrolní*. Atributy v dvou zmíněných datových souborech je možné seskupit do skupin, ze kterých jsou znovu vytvořeny jednotlivé matice. Poslední matice je sestavena z jednotlivých atributů vybrané dvojice skupin atributů. Systematicky je tak možné postihnout všechny možné vztahy. (Černý et al. 2003, s. 184-187)

Součástí projektu STULONG byly i dva výzkumné podprojekty. Prvním z nich je prezentace výsledků analýz v přirozeném jazyce. Cílem byla lepší srozumitelnost výsledků pro uživatele. Prováděné pokusy se týkaly asociačních pravidel ve formě fundované implikace. Převod do přirozeného jazyka byl pokusně implementován na silné implikace pro úlohy vycházející ze vstupních vyšetření. Převod pravidel uskutečňuje systém AR2NL, který zpracovává výsledky z modulu 4ft-Miner systému LISp-Miner. (Projekt STULONG 2003)

6 Např. rodinný stav, vzdělání, zodpovědnost v zaměstnání.
7 Např. bolest hrudníku, bolest končetin, dušnost.

Druhým projektem byla tzv. on-line analýza. Šlo o vytvoření webového rozhraní pro systém LISp-Miner, které by umožňovalo zadávat jednoduché úlohy on-line a získávat obratem odpovědi. Cílem tohoto projektu bylo umožnit uživateli zadávat vlastní úkoly bez speciálních znalostí daného nástroje data miningu. Díky tomu byly vyzkoušeny možnosti webové technologie a možnosti zapojení odborníků. (Černý et al. 2003, s. 187-188)

Určitým druhem podprojektu byla také analýza příčin úmrtí, kterou provedl Burian a Rauch (2003). Vstupem do této analýzy bylo 389 pacientů z projektu STULONG. U každého z pacientů bylo zachyceno 39 atributů, které byly rozděleny do tří skupin *obecné charakteristiky*, *vyšetření* a *neřesti*[8]. Cílem projektu bylo nalezení významných souvislostí vzhledem k příčinám úmrtí jako je infarkt, poruchy srdečních chlopní, úraz, nádorové onemocnění, obecná arteroskleróza atd. V rámci analýzy se podařilo najít potenciálně zajímavá asociační pravidla s kvantifikátory fundovaná implikace a horní kritické implikace. Naopak nepodařilo se nalézt žádná zajímavá pravidla s kvantifikátory dvojitá fundovaná implikace a fundovaná ekvivalence. (Burian a Rauch 2003)

Další aplikací metody GUHA, která nijak nesouvisela s projektem STULONG, byl data mining databáze katetrizací z II. interní kliniky kardiologie a angiologie Všeobecné fakultní nemocnice Praha. Autorem této analýzy byl Štochl (2003). Cílem analýzy bylo najít nové zajímavé souvislosti v zmíněné databázi. Jak již název databáze napovídá, sledovanou problematikou bylo zužování tepen (stenóza). Pro sledování stenóz jsou tepny rozděleny na segmenty. Některé segmenty tepen jsou významnější než jiné a některé mají dané pevné pořadí. (Štochl 2003, s. 192-195)

Okruhy analytických otázek byly následující: rozdíly poškození tepen mezi skupinami pacientů, trendy v rozložení stenóz (např. počet stenóz v závislosti na pořadí segmentu), tendence kombinací segmentů se stenózou k důležitým segmentům se stenózou a konečně tendence kombinací jednotlivých segmentů se stenózou. První dva okruhy analytických otázek byly řešeny statistickými metodami a nástroji. Třetí okruh analytických otázek byl řešen pomocí procedury 4ft-Miner a poslední okruh analytických otázek byl řešen pomocí statistických metod a složitějších SQL dotazů. (Štochl 2003, s. 198-201)

8 Např. alkohol, kouření, kofeinové nápoje.

3. Metody data miningu – vyhodnocení

Konečné rozhodnutí o použitelnosti výsledků data miningu náleží vlastníkovi/ uživateli dat. Ten pro vyhodnocení používá především svých doménových znalostí. Protože analytici provádějící data minig disponují pouze omezenými znalostmi dané domény, je používají při vyhodnocování relativně jednoduché formální metody. Tato kapitola slouží jako přehled těchto metod.

Je nezbytné dodat, že v praxi práce s výsledky probíhá většinou opakovaně. Nejdříve jsou výsledky podrobeny formálnímu vyhodnocení a pak teprve vyhodnocení odborníka. Tento cyklus se opakuje tak dlouho, dokud vlastník/uživatel dat neshledává výsledky jako užitečné. (Cios et al. 2007, s. 470)

Metody na vyhodnocování výsledků data miningu je možné kategorizovat podle různých kritérií. Jedním z těchto kritérií může být fakt, zda data mining probíhá se známými hodnotami cílových atributů (tzv. s učitelem) nebo ne (tzv. bez učitele) (Cios et al. 2007, s. 471). Další rozdělení může být na metody pro selekci výsledků a metody pro zjištění míry správnosti klasifikace/predikce (Cios et al. 2007, s. 469)

Dalším kritériem pak může být rozdělení podle podstaty jednotlivých metod. Cios (2007, s. 471) uvádí následující kategorie:

- znovupoužití dat (např. křížová validace),
- heuristické metody (např. Occamova břitva),
- analytické metody (různá informační kritéria např. Akaike, Bayes),
- míry zajímavosti (např. senzitivita).

Zmíněné metody v závorkách jsou více vysvětleny v následujících sekcích.

V praxi se vyhodnocování výsledků děje dlouhodobě a opakovaně v rámci procesů sledování a udržování modelu. (Rud a Magera 2001, s. 148-154)

Metod pro vyhodnocování výsledků existuje celá řada. Cios (2007, s. 485) uvádí následující posloupnost metod jako nejpoužívanější.

1. Occamova břitva,
2. křížová validace,
3. analýza senzitivity, specificity a křivky ROC.

Následující přehled vychází ze souhrnné monografie Petra Berky *Dobývání znalostí z databází* (2003). Berka rozděluje metody vyhodnocení podle toho, zda se používají u deskriptivních úloh či úloh klasifikačních/předpovědních (dále jen jako klasifikační úlohy).

3.1. Vyhodnocení deskriptivních úloh

U deskriptivních úloh je hlavním kritériem: novost, zajímavost, užitečnost a srozumitelnost. Znalosti získané v rámci řešení deskriptivních úloh je možné rozdělit do následujících skupin:

- zřejmé znalosti ve shodě se „zdravým selským rozumem“,
- zřejmé znalosti ve shodě se znalostmi experta dané oblasti,
- nové zajímavé znalosti,
- znalosti, které musí expert podrobit bližší analýze,
- znalosti, které jsou v rozporu se znalostmi experta (většinou jde o nahodilé shody). (Berka 2003, s. 223)

Pokud jsou nalezené znalosti ve formě asociačních pravidel, je hodnocení založené na již zmíněných numerických parametrech jako je spolehlivost, podpora, pokrytí atd. (viz sekce 1.2).

Pomocí při vyhodnocování znalostí jsou také různé vizualizace. Může jít o různé typy rozhodovacích stromů. Berka dále uvádí jako příklad tzv. „webový uzel“ systému Clementine, ve kterém je síla vztahu znázorněna silou čáry. Pro vyjádření podpory a spolehlivosti asociačního pravidla může sloužit koláčový graf. (Berka 2003, s. 236-238)

3.2. Vyhodnocení klasifikačních úloh

Základní míra pro hodnocení klasifikátorů je správnost klasifikace. Její výpočet vychází z počtu chybných a správných klasifikací na testovacích datech.

Existuje celá řada způsobů jak data, která jsou k dispozici, rozdělit na trénovací a testovací část. Tyto způsoby se liší především podle toho, zda dochází k iteraci nebo nedochází, případně ke kolika iteracím dochází. Dále pak podle toho, kolik procent dat je zařazeno do trénovací a kolik do testovací části dat. (Berka 2003, s. 224-226)

Počet chybných a správných klasifikací je možné zaznamenat v kontingenční tabulce. Pokud jde o klasifikátor klasifikující pouze do dvou skupin (*ano*, *ne*), pak je tato tabulka opět čtyřpolní a jsou v ní hodnoty *ano-dobře*, *ano-špatně*, *ne-dobře*, *ne-špatně*. (Berka 2003, s. 226)

Z této čtyřpolní tabulky vychází nejjednodušší numerické parametry pro hodnocení klasifikátorů. Berka (2003, s. 227-229) uvádí následující:

- *celková správnost* (overall accuracy) nebo též *úspěšnost* (succesfulness),
- *celková chyba* (komplementární charakteristika k celkové správnosti),
- *přesnost a úplnost* (pojmy převzaté z oblasti vyhledávání informací),
- *senzitivita a specificita* (pojmy převzaté z medicíny).

Jak naznačují informace v závorkách, zmíněné parametry se často překrývají. Jde vlastně jen o rozdílné přístupy vedoucí ke stejnému cíli. Celková správnost je poměr správně klasifikovaných příkladů ke všem příkladům. Úplnost je pak poměrem mezi hodnotami *ano-dobře* a součtem hodnot *ano-dobře* a *ne-špatně* z kontingenční tabulky atd.

Čtyřpolní tabulka ukazuje pouze počty chybných a správných zařazení do tříd. V okamžiku, kdy některé chyby mohou mít pro danou firmu horší dopad než jiné, nás také zajímá typ chyby. Proto se k počtům v čtyřpolní tabulce někdy přidávají váhy, resp. ceny. Jiná situace nastává, v případě, kdy jsou jednotlivé třídy rozloženy výrazně nerovnoměrně. Pak je vhodné vypočítat jednotlivé numerické charakteristiky pro jednotlivé třídy zvlášť. (Berka 2003, s. 227-228)

Vedle zmíněných numerických charakteristik je možné popisovat chování modelu ještě komplexněji. Pro příklad může sloužit křivka učení, křivka navýšení, křivka ROC nebo analýza DEA. Křivka učení dává do souvislosti počet příkladů v trénovací množině a správnost klasifikace. Křivka navýšení dává do souvislosti hodnoty *ano-dobře* s všemi hodnotami z kontingenční tabulky. Křivka ROC dává do souvislosti hodnoty *ano-dobře* a *ne-dobře* a stejně jako analýza DEA slouží k optimalizaci nastavení učícího se algoritmu. (Berka 2003, s. 230-235)

Jiné míry se používají, když nejde o klasifikaci do tříd, ale o numerickou predikci. Zde se používá například střední kvadratická chyba, odmocnina ze střední kvadratické chyby, střední absolutní chyba nebo korelační koeficient. (Berka 2003, s. 235-236)

3.3. Použitá metoda vyhodnocení

Cílem mé práce byla deskripce, to znamená nalezení vztahů vedoucích k daným rozhodnutím. Zhodnocení tak spočívalo v konfrontaci zjištěných vztahů se znalostmi lékaře. Lékař výsledky zařadil do pěti skupin: *obecně známé, odborně známé, překvapivé, nesouvisející*. Vyhodnocení proběhlo formou řízeného rozhovoru nad výsledky. Osnova řízeného rozhovoru je v příloze číslo 4.

4. Nástroje data miningu

Základní znaky nástrojů data miningu jsou následující (Berka 2003, s. 271)

- systémy pokrývají celý proces dobývání znalostí,
- většinou nabízejí více algoritmů pro vytváření znalostí,
- kladou důraz na vizualizaci (jak při práci se systémem, tak při interpretaci).

Obecně lze systémy rozdělit na komerční a výzkumné. Podle oblasti zaměření pak na univerzální nebo specializované. (Berka 2003, s. 271, 29) Mimo systémy komerční a výzkumné existuje ještě celá řada systémů v režimu open source. (Kováč 2012) Jiné kritérium pro dělení data miningových nástrojů nabízí například Mikut a Reischl (2011, s. 3) a to podle skupin uživatelů: korporace, aplikovaný výzkum, vývoj nových data miningových algoritmů a vzdělávání.

Kritérií, která vedou k rozhodnutí, který systém zvolit, může být vedle aplikační oblasti celá řada, počínaje možnostmi přístupu k externím datům nebo možnostmi práce s velkými soubory. Kováč (2002, 29-49) vybírá u popisu jednotlivých nástrojů tato kritéria: nabízené API, podpora různých databázových systémů, možnosti vizualizace, možnosti statistické analýzy, volně dostupné softwarové knihovny a pluginy. V neposlední řadě jsou rozhodující i cena, dokumentace, podpora nebo snadnost ovládání, stejně jako u kteréhokoliv jiného softwaru (Berka 2003, 289-290). Vedle těchto charakteristik uvádějí Mikut a Reischl (2011, s. 6) také závislost na platformě: Windows, MAC OS, Linux, mainframe. Vedle těchto platforem se také často používají na platformě nezávislá Java - řešení nebo řešení běžící v cloudu.

Mikut a Reischl (2011, s. 6-11) uvádějí následující kategorizaci:

- **data miningové suity** – řešení zaměřená přímo na data mining, nabízí širokou funkcionalitu, většinou jsou poměrně drahé,
- **řešení pro business intelligence** – nejsou přímo zaměřené na data mining, ale obsahují základní funkcionalitu pro data mining, především týkající se statistických metod,
- **řešení pro matematiku** – nejsou přímo zaměřené na data mining, ale nabízejí velké možnosti vytváření algoritmů a vizualizace, jejich výstupy jsou často drobná rozšíření (add-ons) jiných aplikací,
- **integrační balíky** – rozšiřitelné balíky různých open-sourcových algoritmů, fungující jako samostatný software (většinou na platformě Java) nebo jako větší rozšíření pro jiné řešení,

- **drobná rozšíření** (add-ons) – určená pro nástroje jako jsou Excel, Matlab apod. většinou s omezenou, ale užitečnou funkcionalitou,
- **data miningové knihovny** – balíky funkcí, které je možné využívat prostřednictvím API,
- **speciální nástroje** – podobně jako data miningové suity, jen zaměřené na implementaci speciální skupiny metod jako jsou například neuronové sítě,
- **výzkumné nástroje** – většinou první a ne vždy stabilní implementace nového a inovativního algoritmu,
- **speciální řešení** – nástroje upravené pro úzkou oblast použití jako například vytěžování textů.

5. Systém LISp-Miner

5.1. Historie

Systém LISp-Miner je jednou z implementací metody GUHA, respektive implementací několika GUHA-procedur. (Šimůnek 2010, s. 13) Práce na tomto systému začala na přelomu let 1995-96. První procedura, která byla implementována, byla procedura DB-ASSOC.

Zajímavou, ačkoliv neúspěšnou kapitolou ve vývoji systému byla snaha o multi-relační mining v letech 1998-1999. Multi-relační data mining se ale ukázal jako příliš náročný, pro běžného uživatele v podstatě nepoužitelný. (Šimůnek 2010, s. 18)

Od roku 1998, kdy byla opuštěna cesta multi-relačního miningu, se začalo pracovat na nové koncepci systému LISp-Miner. Tato nová koncepce je charakteristická dvěma hlavními prvky modularitou a metabází. V systému LISp-Miner jsou v současnosti implementovány procedury 4ft-Miner, CF-Miner, KL-Miner, SetDifference procedury a AC4ft-Miner. Od začátku vývoje systému, obsahuje LISp-Miner také proceduru strojového učení KEX, která ale nepatří mezi původní GUHA-procedury.

5.2. Architektura

Jak již bylo zmíněno, hlavní rysy systému LISp-Miner jsou modularita a metabáze. Metabáze slouží k ukládání nastavení jednotlivých úloh a jejich výsledků. Na jedněch datech je tak možné mít více metabází. Hlavní moduly byly vytvořeny na základě metodiky CRISP-DM. Základní moduly systému LISp-Miner jsou:

- moduly pro porozumění a transformaci dat
    - LM DATASOURCE
    - LM TIMETRANSF
- moduly jednotlivých GUHA-procedur (4ft-miner, CF-miner, KL-miner, procedury SetDifference)
    - XXTASK - modul pro zadání úloh
    - XXRESULT - modul pro interpretaci výsledků
    - XXGEN - modul pro dávkové zpracování úloh z externích aplikací

- moduly pro proceduru strojového učení
    - KEXTASK
    - KEXRESULT
- moduly pro porozumění problematice
    - LM LAQ MANAGER
    - LM KNOWLEDGESOURCE
- administrativní modul
    - LM ADMIN
- modul pro podporu prezentace získaných dat
    - LM SWBEXPORTER

5.3. Základní pojmy používané v systému LISp-Miner

Pro lepší srozumitelnost především kapitoly 7 uvádím definici vybraných pojmů ze systému LISp-Miner. Jejich popis vychází z přehledu používaných pojmů v publikaci „*Systém LISp-Miner, akademický systém pro dobývání znalostí z databází*“ Šimůnka (2010, s. 88-90).

- **délka antecedentu** – maximální počet literálů
- **literál** – kombinace atributu a koeficientu
- **koeficient** – množina kategorií, definována typem koeficientu, případně délkou koeficientu
- **typ koeficientu** – Definuje přípustné kombinace kategorií, které mohou být zároveň vloženy do koeficientu literálu. LISp-Miner nabízí celou řadu typů viz sekce 2.1. Zde zmíním pouze ty, které jsem použil ve své práci.
- **typ koeficientu subset** – libovolná kombinace kategorií
- **typ koeficientu interval** – V koeficientu mohou být pouze sousední kategorie (tak jak jsou definovány v modulu LM DataSource).
- **typ koeficientu left cut** – jako typ koeficientu interval, ale vždy musí začínat první kategorií
- **typ koeficientu right cut** – jako typ koeficientu interval, ale vždy musí končit poslední kategorií

- **typ koeficientu one category** – Koeficient může tvořit pouze jedna explicitně zadaná kategorie.
- **délka koeficientu** – počet kategorií v koeficientu (je možné nastavit podle typu koeficientu)
- **kvantifikátor** – V LISp-Mineru se kvantifikátor skládá z číselné charakteristiky pravidel (resp. funkce, do které vstupují hodnoty z čtyřpolní tabulky četností), z relačního operátoru (*vetší než, menší než*…) a prahové hodnoty. Kvantifikátory se v systému LISp-Miner dělí na agregační a funkční.
- **agregační kvantifikátor** – číselná charakteristika je pouze jednou frekvencí z čtyřpolní tabulky četností
- **funkční kvantifikátor** – Číselná charakteristika pravidla je složitější než jedna frekvence z čtyřpolní tabulky četností. Na základě podobných logických vlastností je možné tyto charakteristiky rozdělit do tříd: implikační, dvojité implikační, ekvivalenční a symetricky asociační.

6. Data

6.1. Poradna pro poruchy metabolismu

Data pocházejí z *Poradny pro poruchy metabolismu tuků* (dále také jako poradna). Tato poradna funguje při *Ústavu klinické biochemie a hematologie Fakultní nemocnice Plzeň*. Cílem poradny je prevence, vyšetření, pravidelné léčení a sledování převážně pacientů s hyperlipidémií a jinými závažnými rizikovými faktory arterosklerózy.

Cílem předzpracování dat je vybrat pro dané úlohy relevantní atributy a převést data do formátu, který vyhovuje zvolenému systém (Berka 2003b). V následujících sekcích je nejdříve popsán výběr dat a pak způsoby jejich zpracování.

6.2. Výběr dat

Při výběru dat jsem řešil následující otázky:

- počet pacientů,
- výběr pacientů,
- výběr záznamů,
- výběr vyšetření,
- předchozí farmakoterapie.

6.2.1 Počet pacientů

Spolupracující lékař MUDr. Roman Cibulka, PhD., MBA (dále také jako lékař) převzal poradnu v roce 2011. Celkem jsem získal 105 záznamů návštěv (dále také jako záznamy) rozdílných pacientů. Zhruba 75 záznamů se týká pacientů lékaře (jsou z let 2011-2013). Zbylých 20 záznamů se týká pacientů, které ošetřovala předchůdkyně MUDr. Cibulky. Vzhledem k tomu, že lékař používal při výběru záznamů osobní znalost pacientů (viz níže) a vzhledem k celkovému počtu pacientů, kteří poradnu navštíví, byl počet 105 záznamů reálné maximum, které jsem mohl získat.

6.2.2 Výběr pacientů

Z pacientů poradny jsme museli vyfiltrovat výjimečné pacienty. Výjimečným pacientem byl například pacient po transplantaci ledvin a to z důvodu, že transplantované ledviny mohou zapříčinit změnu řady hodnot laboratorních vyšetření. Dalším výjimečným pacientem může být například pacient jiné národnosti.[9] Dalším problémem mohou být například pacienti

9 Normální rozmezí některých laboratorních vyšetření je vázané na konkrétní populaci.

s psychiatrickou diagnózou apod. Výběr pacientů jsem nechal na lékaři.

6.2.3 Výběr záznamů

Těžké poruchy metabolismu tuků, které jsou v poradně léčeny, jsou léčeny řadu let, většinou pak celý život. Jde o poruchy z velké části způsobené geneticky. Důvodů, proč pacient do poradny přijde, je celá řada. Může jít například o obezitu nebo zvýšenou hladinu krevních tuků, která je zjištěna během preventivního vyšetření u praktického lékaře. Jen někdy je farmakoterapie nasazena hned během první návštěvy. Většinou je po vstupním laboratorním vyšetření a objektivním vyšetření probrán životní styl a doporučena pohybová aktivita nebo změna stravovacích návyků. Teprve když tato doporučení nepomohou a naměřené hodnoty vyšetření se nezlepší, je nasazena léčba léky. Mezi vstupním vyšetřením a kontrolním vyšetřením, kdy jsou nasazena farmaka, často uběhne až několik let a uskuteční se řada kontrolních vyšetření. Z hlediska cíle mé práce vybral lékař do datového souboru záznam právě toho kontrolního vyšetření, kde došlo k předepsání léků. Protože lékař mohl pracovat v informačním systému jen s jednoduchým dotazovacím rozhraním, výběr takového záznamu prováděl „ručně“ na základě osobní znalosti pacientů. U pacientů, kde zatím nedošlo k předepsání léků, jsme vybrali vždy poslední záznam.

6.2.4 Výběr vyšetření

Lékařské záznamy obsahují celou řadu údajů. Jsou to osobní údaje, anamnézy, objektivní vyšetření a laboratorní vyšetření.

Osobní údaje

Vedle osobních údajů věk a pohlaví tu jsou ještě adresa, rodné číslo a zdravotní pojišťovna. Tyto údaje nebyly použity z důvodu anonymizace dat a jejich malé nebo žádné relevance k řešenému úkolu.

Anamnézy

Z anamnéz byla vypuštěna farmakologická anamnéza, alergologická anamnéza, gynekologická anamnéza, pracovní anamnéza, sociální anamnéza. Alergologická anamnéza a gynekologická anamnéza mají z hlediska řešené problematiky jen okrajový význam. V záznamech tyto údaje také většinou chybí. Farmakologická anamnéza byla vypuštěna pro zjednodušení problematiky (viz 6.2.5.) Údaje o pracovní a sociální anamnéze byly vypuštěny pro malou objektivnost, pro zjednodušení problematiky, a protože v záznamech často chybí.[10] Těžké poruchy metabolismu,

10 Z výzkumů je známo, že lidé žijící osaměle nebo pracující na směny mají vyšší riziko poruch metabolismu. Stejně tak se zvyšuje riziko při nedostatku pohybu nebo špatném stravování. Bohužel údaje o životním stylu není jednoduché ověřit. Podle lékaře si lidé tyto informace často přilepšují. V poradně se tyto údaje s pacienty probírají hlavně proto, aby pacient dostal zpětnou vazbu. Do lékařských záznamu se proto dostávají pouze ve zkratce.

které jsou v poradně léčeny, lze životosprávou korigovat pouze částečně.

Objektivní vyšetření

Mezi objektivní vyšetření jsem zařadil body mass index a krevní tlak. Body mass index neboli index tělesné hmotnosti je podíl hmotnosti a druhé mocniny výšky. Objektivní vyšetření hmotnost a výška tak zůstaly de facto zastoupeny. Pro řešenou problematiku má význam hlavně to, jestli má člověk nadváhu nebo ne. Závislost krevního tlaku a sledované problematiky je všeobecně známá. Pro údaje do tabulky jsem vybral pouze údaj o systolickém tlaku. Pro určení hypertenze je dostačující.

Nově se v lékařských kartách také objevuje index BSA a index pas/výška. Oba dva indexy by měly doplňovat a korigovat body mass index, který je problematický například u kulturistů či lidí s nezvyklejší stavbou těla. Podle lékaře ale do poradny žádní vrcholoví sportovci nechodí a tak tyto indexy nemají velkou důležitost. Podobný smysl má mezi objektivními vyšetřeními také údaj procento tělesného tuku. Jde o doplňkový údaj.

Mezi objektivní vyšetření patří ještě vyšetření pulsu. Ten má význam z hlediska nasazovaného typu antihypertenziv. Některá antihypertenziva totiž puls zvyšují. Pokud nás ale zajímají antihypertenziva pouze jako třída léků, nemá tento údaj smysl zpracovávat.

Laboratorní vyšetření

Zdravotní karty obsahovaly u některých pacientů až 72 laboratorních vyšetření. Celá řada z nich je prováděna ale pouze při vstupním vyšetření. Jsou prováděny pro úplnost nebo jako doplňkové vyšetření. Vzhledem k zmíněnému a také vzhledem k faktu, že celá řada těchto laboratorních vyšetření již není realizována při pozdějších kontrolních vyšetřeních, jsem po konzultaci s lékařem počet zpracovávaných laboratorních vyšetření omezil na 17. Mezi těmito laboratorními vyšetřeními zůstaly především údaje sledující enzymy, sacharidový metabolismus, lipidový metabolismus a dusíkaté látky. Zcela byly například vypuštěny údaje o krevním obraze, minerálech nebo vyšetření moči.

Celkový počet údajů u jednoho záznamu, které jsem zpracovával se tak zmenšil na 25.

6.2.5 Předchozí farmakoterapie

Léky, které bral pacient před rozhodnutím nasadit nový lék, jsem nebral v potaz. V mé práci předpokládám, že předchozí farmakoterapie vedla k normalizaci hodnot vyšetření. Jiný předpoklad by udělal situaci velmi složitou.

6.3. Zpracování dat

Pří zpracování dat jsem řešil následující problémy:

- forma získaných dat,
- anonymizace dat,
- normalizace dat,
- kategorizace osobních údajů, anamnéz a objektivních vyšetření,
- kategorizace laboratorních vyšetření,
- kategorizace farmakoterapie a vzájemná závislost typu farmak.

6.3.1 Forma získaných dat

Poradna, stejně jako celá Fakultní nemocnice Plzeň, používá klinický informační systém MEDICALC. Lékař poradny má přístup pouze k jednoduchému uživatelskému rozhraní. Výstupem ze systému, který jsem získal, byly záznam ve formě tištěné lékařské karty. Údaje na těchto kartách nejsou strukturovány do tabulek, ale pouze do řádů. Karty jsem proto musel přepsat ručně do tabulkového procesoru.

Problémem při zmíněném výběru záznamů (viz 6.2.3.) byla skutečnost, že ne všechna vyšetření jsou prováděna při každé návštěvě (většinou v půlročním intervalu). Některá vyšetření stačí provést jednou za rok nebo jednou za několik let. Pokud takovéto údaje na záznamu chyběly, bylo nutné ještě zpětně výsledky dohledat v předchozích a následujících záznamech.

Z papírových záznamů jsem data nejdříve přepsal do tabulky MS Office Excel. Tuto tabulku jsem pak pomocí MS Office Access převedl do formátu mdb. Formát mdb je jedním z formátů, které podporuje systém LISp-Miner.

6.3.2 Anonymizace dat

Na každé vytištěné kartě byly uvedeny osobní údaje: jméno, adresa, datum narození, rodné číslo. Tyto údaje jsem ještě v poradně začernil fixem a každou kartu nadepsal číslem. Lékař si pak číslo napsal do seznamu svých pacientů pro případ, že by bylo nutné pacienta dohledat kvůli nejasnostem. Na kartách jsem nezačernil první dvě číslice z rodného čísla a rok narození. Počátek rodného čísla jsem použil k zjištění pohlaví, pokud nebylo v kartě vysloveně uvedeno, že jde o pacienta nebo pacientku. Rok narození jsem pak používal k zjištění věku pacienta.

6.3.3 Normalizace dat

Z důvodu změny zápisu hodnot v informačním systému v minulých letech, bylo nutné některá

laboratorní vyšetření normalizovat. Konkrétně to byla vyšetření glykohemoglobínu a C-reaktivního proteinu. U prvního zmíněného vyšetření bylo nutné všechny hodnoty menší deseti (starý zápis) vynásobit deseti. U druhého vyšetření bylo nutné všechny hodnoty větší sto naopak vydělit deseti. Jistou formou úpravy dat může být i vymazání hodnoty diastolického tlaku a ponechání pouze hodnoty systolického tlaku. Určité zpracování podstoupil také údaj věk. Od věku získaného z rodného čísla jsem ještě odečetl jeden až tři roky, podle toho, zda předepsání léku proběhlo v roce 2010, 2011 nebo 2012.

6.3.4 Kategorizace osobních údajů, anamnéz a objektivních vyšetření

Osobní údaje věk a pohlaví bylo jednoduché kategorizovat. Věk jsem kategorizoval do intervalů po deseti letech s přihlédnutím na jejich četnost (první kategorie je *0-30*, poslední *60-100* let).

O něco obtížnější bylo kategorizovat anamnézy. Zde jsem provedl po dohodě s lékařem výraznější zjednodušení problematiky. Osobní anamnéza je *riziková*, pokud pacient trpí:

- poruchou metabolismu,
- diabetem,
- hypertenzí,
- nebo prodělal kardiovaskulární onemocnění či onemocnění ledvin.

V ostatním případech byla přidělena hodnota *neriziková*.

Ještě větší zjednodušení bylo u kategorizace rodinné anamnézy. Rodinná anamnéza byla kategorizována jako *riziková*:

- pokud měl někdo z rodiny kardiovaskulární onemocnění,
- pokud se v rodině vyskytl diabetes nebo porucha metabolismu tuku a současně byly tyto choroby diagnostikovány také u pacienta.

V ostatních případech byla přidělena kategorie *neriziková*. Zcela jsem opomenul skutečnost, jestli chorobou v rodině trpěli rodiče nebo prarodiče, či kdy u nich onemocnění propuklo (rozdíl mezi diabetem 1. a 2 typu). Kategorizace anamnézy abúzus probíhala následovně:

- *riziková*: pravidelné pití alkoholu a kouření,
- *potencionálně riziková:* příležitostné kouření nebo dřívější kouření,
- *neriziková*: alkohol příležitostně nebo vůbec ne.

Objektivní vyšetření BMI jsem kategorizoval podle hodnot z referenční příručky „Memorix vademecum lékaře“ (Droste a Planta 1992). Kategorizaci objektivního vyšetření krevní tlak jsem

zjednodušil a s přihlédnutím k četnosti hodnot v souboru jsem místo pěti kategorií (*nízký, normální, zvýšený, hypertenze 1. stupně, hypertenze 2. stupně*) použil pouze tři kategorie (*normální, zvýšený, vysoký).*

6.3.5 Kategorizace laboratorních vyšetření

Jednou z klíčových otázek úspěšného získávání znalostí z dat je otázka správné kategorizace. Ne jinak to bylo u mé práce. Je ale nutné opravdu číselné hodnoty kategorizovat?

Systém LISp-Miner umožňuje použít pro číselné hodnoty[11] typ koeficientu right cut, případně i left cut. Nevýhodou koeficientu right cut a left cut je ale nemožnost provádět některé úlohy, například úlohy používající kvantifikátor dvojitá fundovaná implikace.

Jak kategorizovat laboratorní vyšetření? Při volbě vhodných kategorií jsem měl k dispozici dvě základní možnosti. První bylo zvolit kategorizaci laboratorního systému FN Plzeň (dále také jako kategorizace FN Plzeň). Druhou bylo zvolit nějakou obecnou referenční příručku, třeba „Memorix vademecum lékaře“ (Droste a Planta 1992) (dále také jako kategorizace Memorix). Jaké jsou výhody a nevýhody obou řešení?

Kategorizace FN Plzeň

Nejdříve popis kategorizace, která je přímo v laboratorním systému FN Plzeň. Její výhoda je, že je lékařům již známá, i když hlavně v podobě grafických značek ve výsledkových tabulkách systému než pod jednotlivými kategoriemi. Další výhodou této kategorizace je vysoká granularita, hodnoty *patologicky nízká, nízká, nižší, normální, vyšší, vysoká, patologicky vysoká*. Nevýhodou je opět velmi vysoká granularita (viz níže) a rozdílná kategorizace pro muže a ženy. Rozdílné normální hodnoty pro muže a ženy jsou zde u 8 z 18 laboratorních vyšetření.[12] Rozdílná kategorizace podle pohlaví je problémem hlavně proto, že LISp-Miner nenabízí možnost podmíněné kategorizace. V následujících třech odstavcích popíši tři možná řešení.

Prvním řešením je rozdělit soubor na dva malé soubory, jeden pro ženské pacienty a jeden pro mužské pacienty. Vzhledem k již tak malé velikosti souboru to ale není příliš dobré řešení.

Druhým řešením je provést kategorizaci mimo LISp-Miner například v MS Excelu (funkce IF) a do systému LISp-Miner vložit již kategorizované hodnoty. Zde je ale problém s nastavením typu koeficientu. Pokud bych používal sedmi člennou kategorizaci FN Plzeň, bylo by nutné použit typ koeficientu subset alespoň o délce 3 (*vyšší, vysoká, patologická*). To by ale znamenalo explozi mnoha zbytečných pravidel s hodnotami literálu například Cholesterol(*nižší, normální, vyšší)*.

11 Pojem číselné hodnoty je zde relativní, ve skutečnosti jde o velmi malé intervaly pro každou hodnotu.

12 Normální hodnoty mohou být rozdílné v závislosti na pohlaví až o jednu kategorii, například *normální* kategorie apolipoproteinu A je u žen mezi 1,6-1,8 g/l, u mužů odpovídá tento interval kategorii *vyšší*.

Použití typu koeficientu right cut a left cut by nebylo možné. Systém by vložené kategorie vnímal pouze jako řetězce bez významové souvislosti.

Posledním řešením, které jsem následovně částečně použil, bylo zprůměrovat hodnoty kategorizace pro muže a ženy.

Kategorizace Memorix

Tato kategorizace používá pouze tři hodnoty: *nízká, normální, vysoká* a až na výjimky (Apo A) nerozlišuje hodnoty podle pohlaví. Výhodou je snadná a srozumitelná aplikovatelnost. Nevýhodou je malá granularita. Je pravděpodobné, že velmi vysoké patologické hodnoty některých vyšetření, mohou rozhodování ovlivnit daleko více než pouze lehce zvýšené hodnoty.

Použitá kategorizace

Kategorizace, kterou jsem nakonec použil, je směsicí obou uvedených. Základem byla kategorizace Memorix. U laboratorních vyšetření[13], kde byly významněji zastoupeny vysoké patologické hodnoty, jsem přidal čtvrtou kategorii *patologická* podle hraničních hodnot kategorizace FN Plzeň. Rozdílné normální hodnoty pro muže a ženy u laboratorního vyšetření Apolipoproteinu A a poměru Apolipoproteinu A/B jsem zprůměroval. Přehled hodnot použité kategorizace je v příloze č. 3.

6.3.6 Kategorizace farmakoterapie a vzájemná závislost typů farmak

V poradně se běžně předepisují následující druhy léků:

- *hypolipidemika (41),*
- *antihypertenziva (11),*
- *antidiabetika (16),*[14]
- *antiobezitika (14),*
- *bez léčby (33).*[15].

Čísla v závorkách představují četnost jednotlivých léků. V lékařských kartách jsou pouze konkrétní názvy léků. Přiřazení léku k jednotlivým druhům provedl lékař již při výběru záznamů.

Jednotlivá léčiva nemají vzájemnou závislost. Vysvětlím to na příkladu antihypertenziv a hypolidemik. Antihypertenzivum sníží vysoký tlak na normální hodnoty. Hodnoty laboratorních vyšetření, které indikují hyperlipidémii zůstávají stejná. Pokud je kontrolní vyšetření, při kterém

13 Jsou to vyšetření TG, Apo B, CHOL, LDL a Lpa.
14 Okrajově jsou v poradně léčeni také pacienti s poruchou metabolismu cukrů.
15 Doporučena pouze změna stravovacích návyků a životního stylu.

bylo pacientovi předepsáno více druhů léčiv, například hypolipidemika a antihypertenziva, mohu ze záznamu udělat ve výsledné datové tabulce záznamy dva. Tyto záznamy budou mít stejné hodnoty až na údaj farmakoterapie. První záznam se bude týkat hypolipidemik, druhý záznam se bude týkat antihypertenziv. Když jsem záznamy s více předepsanými léky takto rozložil, získal jsem ze 105 záznamů 115 záznamů.

6.3.7 Použité kategoriální hodnoty

V této sekci jsou všechny vstupní atributy. V úvodní fázi porozumění datům je vhodné zjistit i četnosti jednotlivých atributů (Berka, 2003) V závorkách za každou kategorií jsou proto také uvedeny počty pacientů.

osobní údaje

věk: *do 30 (9); 30-40 (22); 40-50 (26); 50-60 (31); nad 60 (27)*

pohlaví: *m - muži (49); ž – ženy (66)*

anamnézy

OA – osobní anamnéza: *n – neriziková (21); r – riziková (93); null (1)*[16]

RA – rodinná anamnéza: *n – neriziková (26); r – riziková (80); null (9)*

abusus: *n – nerizikový (61); p - potencionálně rizikový (24); r – rizikový (28); null (2)*

objektivní vyšetření

BMI – body mass index: *podváha (0), normální (23); nadváha (52); obezita (40)*

krevní tlak: *normální (77); zvýšený (22); vysoký (16)*

laboratorní vyšetření

ALT – alaninaminotransferáza: *normální (96); vysoký (19)*

GMT – gama-glutamyltransferáza: *normální (100); vysoká (15)*

CK – kreatinkináza: *normální (82); vysoká (33)*

glukóza – glukóza: *normální (75); vysoká (40)*

GHBA – glykohemoglobin: *normální (43); vysoká (70); null (2)*

TG – triglyceridy: *nízká (20); normální (54); vysoká (35); patologická (6)*

CHOL – cholesterol: *nízká (4); normální (36); vysoká (65); patologická (10)*

HDL – HDL cholesterol: *normální (43); vysoká (72)*

LDL – LDL cholesterol: *normální (52); vysoká (56); patologická (7)*

Apo-A – apolipoprotein A: *normální (104); vysoká (11)*

16 Null = chybějící hodnoty.

Apo-B – apolipoprotein B: *normální (43); vysoká (64); patologická (8)*

Apo-A/Apo-B[17] - poměr Apo-A/Apo-B: *nízká (89); normální (13); vysoká (13)*

Lpa – lipoprotein malé a: *normální (61); vysoká (50); patologická (4)*

KREA – kreatinin: *normální (111); vysoká (4)*

KMOC – kyselina močová: *nízká (10); normální (98); vysoká (7)*

CRP – C-reaktivní protein: *normální (88); vyšší (27)*

TSH – tyreotropin: *normální (97); vysoká (12)*

MALB – mikroalbuminurie: *normální (105); mikroalbuminurie(2)*

Farmakoterapie

Farmakoterapie – *antidiabetika (16); antihypertenziva (10); antiobezitika (15); bez léčby (33); hypolipidemika (41)*

6.4. Analytické otázky

Analytická otázka popisuje problém, který má být řešen metodami data miningu. Je formulována v přirozeném jazyce. Na základě získaných dat, jsem ve své práci zvolil následující analytické otázky:

1. Kdy jsou předepsána antihypertenziva?
2. Kdy jsou předepsána antidiabetika?
3. Kdy jsou předepsána antiobezitika?
4. Kdy je pacient nechán bez farmakoterapie?
5. Která z těchto pravidel nejvíce pomáhají při rozhodování?
6. Jsou nějaké laboratorní vyšetření duplicitní?
7. Kdy dochází k nějaké léčbě?
8. Která vyšetření jsou pro rozhodování klíčová?
9. Jak se mění výsledky vyšetření, jedná-li se o muže, ženu? (ukázka pravidel s podmínkami)

17 Jde o poměr. Nízké hodnoty jsou pro zdraví pacienta škodlivé, vysoké naopak v pořádku.

7. Výsledky

Tato kapitola se věnuje výsledkům mé práce. Kapitola je rozčleněna podle analytických otázek z předchozí kapitoly. Začíná otázkami využívajícími fundovaný implikační kvantifikátor (Kdy je předepsáno hypolipidemikum, antihypertenzivum atd., dále také jako implikační otázky). Následuje otázka pracující se zobecněným sukcendentem (Kdy je předepsán nějaký lék? - dále v textu spadá také pod implikační otázky) a otázka na duplicitní vyšetření využívající dvojitý fundovaný implikační kvantifikátor (dále také jako dvojitá implikační otázka). Poslední analytická otázka je ukázkou využití podmínky v pravidlech (Role pohlaví na předepisování léku - dále jako otázka s podmínkou).

7.1. Nastavení úloh

U každé otázky jsem potřeboval vyzkoušet vždy několik úloh lišících se nastavením, než jsem získal nosné výsledky.[18] Jak již bylo řečeno v sekci 2.1, v systému LISp-Miner lze provádět u každé úlohy celou řadu nastavení. V následujících sekcích popíši jednotlivá nastavení, které jsem použil.

7.1.1 Délka antecedentu a sukcedentu

Předpokládal jsem, že pro nalezení základních vztahů hledám pravidla o délce antecedentu jeden nebo dva literály. Pokud žádná dostačující pravidla nebyla nalezena, budu hledat pravidla o délce antecedentu tři literály. Délka sukcedentu byla u většiny analytických otázek rovna jeden literál. Výjimkou byla otázka hledající duplicitní vyšetření. V této otázce se antecedent rovnal sukcedentu.

7.1.2 Výběr literálů

U většiny analytických otázek byly v antecedentu osobní údaje, anamnézy, objektivní vyšetření a laboratorní vyšetření. V sukcedentu pak byla pouze hodnota farmakoterapie. Výjimkou byla otázka hledající duplicitní vyšetření. Zde se antecedent rovnal sukcedentu. Do úloh byla u této otázky zahrnuta pouze laboratorní vyšetření.

7.1.3 Typ a délka koeficientu

Po konzultaci s lékařem, jsem se rozhodl, že za významná pravidla budu považovat pouze ta, která obsahují literál s hodnotou zvýšenou. Předpokladem zde je, že normální hodnoty nějakého vyšetření nevedou k předepsání léku.

Začnu antecedentem. Literály pohlaví, osobní anamnéza, rodinná anamnéza, abúzus byly

18 V kapitole s výsledky jsem nechal pouze úlohy jejichž nastavení přinášelo nosná pravidla. Seznam všech úloh, které jsem provedl, a jejich nastavení je v příloze č. 1.

kategorizovány od začátku. Žádné číselné hodnoty zde ani z logiky věci neexistují. Typ koeficientu se v úlohách nikdy nezměnil. Vždy byl nastaven typ subset o maximální délce 1, případně o maximální délce dvě u abúzu (kategorie: *nerizikový*, *potencionálně rizikový*, *rizikový*). Z výsledných pravidel bylo pak nutné vytřídit pravidla s literálem abúzus (*normální, potencionálně rizikový*) a nechat pouze pravidla s literálem abúzus (*potencionálně rizikový, rizikový*).

Nastavení typu koeficientu u ostatních literálů v antecedentu se lišilo podle toho, zda jde o číselné hodnoty nebo kategoriální hodnoty.

U číselných hodnot jsem použil typy koeficientu right cut a left cut (dále také jako řezy). Right cut jsem použil u implikačních otázek. Nyní je možné namítnout, že i příliš nízké hodnoty jsou patologické a mohou vést k předepsání léku. Po konzultaci s odborníkem jsem se rozhodl, že tuto skutečnost budu opomíjet. V praxi nízké hodnoty ve valné většině případů k předepsání léku nevedou.[19] Výjimkou byl poměr Apo A/Apo B, kde nízké hodnoty jsou „špatné“. Zde jsem nastavil u číselných hodnot typ koeficientu left cut.

Left cut jsem použil pouze u otázky zkoumající, kdy není předepsaný žádný lék. Předpokladem u této otázky bylo to, že mě zajímá hranice hodnoty vyšetření, od které směrem dolů není předepsaný žádný lék. Výjimkou byl opět poměr Apo A/Apo B, kde jsem u této otázky nastavil typ koeficientu right cut.

U kategoriálních hodnot je možné použít hned tři typy koeficientu subset, cuty (left a right cut) a one category. Pokud zvolíme u literálu typ koeficientu subset, dostaneme řadu pravidel, kde bude jeden z literálů v antecedentu mít hodnotu *normální*. Pravidlo splňuje výpočetně zadání úlohy, ale nesplňuje výše zmíněný předpoklad. Řešením může být typ koeficientu one category, kdy zvolíme pro daný literál pevnou hodnotu. Za daného předpokladu to bude hodnota *vysoká* (resp. *nízká* u Apo A/Apo B). Problém ale nastává, když mám některé literály s jemnější kategorizací např. *nízká, normální, vysoká, patologická*“.[20] U takovýchto literálů je nejlépe zvolit typ koeficientu right cut o délce 2.

Jiná situace je samozřejmě u otázky zkoumající, kdy nebyla předepsána žádná léčba. Zde jsem natavil typ koeficientu left cut o délce 2 (*nízká*, *normální*). U Apo A/Apo B jsem nastavil typ koeficientu right cut o délce 2 (*normální*, *vysoká*). Nastavení u dvojité implikační otázky jsem popsal v sekci 7.5.10. U otázky s podmínkou jsem v podmínce nastavil typ koeficientu jako one category.

19 V datovém souboru se *nízké* hodnoty vyskytly u těchto vyšetření: TG (20 pacientů), CHOL (4 pacienti), Apo_A/Apo_B (89 pacientů), KMOC (10 pacientů).

20 Podobně tomu je i u literálu věk, který jsem kategorizoval do 5 intervalů po 10 letech. Zde jsem nastavil typ koeficientu right cut a délku pět.

Co se sukcedentu týče, u implikačních otázek jsem typ koeficientu nastavil na one category. U dvojitých implikačních se sukcedent rovnal antecedentu.

7.1.4 Kvantifikátory

V nastavení úloh jsem používal dva typy kvantifikátorů, funkcionální a agregační. Z funkcionálních kvantifikátorů to byly fundovaná implikace[21] a fundovaná dvojitá implikace. Z agregačních kvantifikátorů to byl kvantifikátor BASE[22].

Problémy se u kvantifikátoru BASE vyskytly hlavně v úlohách hledajících pravidla pro předepsání antihypertenziv (10 pacientů), antidiabetik (16 pacientů) a antiobezitik (15 pacientů). Počty pacientů, kterým byly tyto léky předepsány, jsou velmi malé. Pro nalezení alespoň nějakých pravidel bylo třeba někdy hodnotu BASE značně snížit. Při přílišném snížení, ale dojde k explozi pravidel, kdy pravidla obsahují prakticky všechny možné kombinace literálů. Dostatečný počet pacientů byl k dispozici při hledání pravidel pro předepsání hypolipidemik (41 pacientů) a pravidel pro nepředepsání žádného léku (33 pacientů). Pro hledání pravidel pro předepsání nějakého léku bylo k dispozici 82 pacientů.

Při použití kvantifikátoru fundovaná implikace záleželo často na délce pravidel. Pravidla s antecedentem o délce 1 nebyla často nalezena ani při nastavení hodnoty funkce 0,6. Naopak pravidla s antecedentem o délce tři měla hodnoty implikační funkce v intervalu 0.95-1.00.

Nastavování kvantifikátorů probíhalo následujícím způsobem. Nejdříve jsem nastavil kvantifikátory na následující hodnoty:

- BASE (pokrytí) => přibližně 1/4 celkového počtu pacientů s daným sukcendentem, maximálně však 20[23]
  (např. u *hypolipidemik* (41 pacientů) je BASE >= 10 (24,39%); např. u *léčby* je BASE >= 20 (17,39%); hodnotu jsem v systému zadával v absolutních počtech, v sekci 7.4 je ale v %),
- fundovaná implikace/dvojitá fundovaná implikace (spolehlivost/D-spolehlivost) => 0,6 (v sekci 7.4 je tato hodnota vyjádřená v %).

21 Kvantifikátor fundovaná implikace je shodná s charakteristikou spolehlivosti viz 1.2.
22 Kvantifikátor BASE je shodný s charakteristikou pokrytí viz 1.2.
23 Minimální statisticky relevantní počet.

Ve většině případů LISp-Miner vyhledávání pravidel nedokončil[24], protože pravidel bylo nalezeno příliš mnoho (1000 a více). Letmou analýzou charakteristik nalezených pravidel jsem provedl zpřesnění nastavení kvantifikátorů a spustil proceduru vyhledávání pravidel znovu. Tento postup jsem několikrát opakoval, než jsem nalezl takové nastavení, které umožnilo úlohu dokončit a zároveň generovalo pokud možno dostatečně kvalitní pravidla. Pod pojmem dostatečně kvalitní pravidlo, rozumím pravidlo, kdy fundovaná implikace má hodnotu 0,8-1,0;[25] a BASE má alespoň ¼ ze všech případů (dále také jako kvalitní pravidlo). Celkový přehled úloh je v příloze č. 2.

7.1.5 Nastavení parametrů

Systém LISp-Miner umožňuje vedle nastavení zmíněných v sekci 12.1, nastavovat také tzv. parametry. Mezi ně patří například nastavení režimu práce s chybějícími údaji nebo nastavení maximálního počtu pravidel v úloze.

Co se týče nastavení režimu práce s chybějícími údaji, tak v mé práci jsem vždy nechal nastaven konzervativní režim. To znamená, že všechny záznamy s chybějícími údaji se ignorují. To má pak vliv na hodnotu charakteristiky celkový počet pacientů (viz například implikační úloha s antidiabetiky).

Maximální počet pravidel v úloze určuje počet, při kterém LISp-Miner ukončí generování pravidel. Defaultně je nastaven počet 500. U úloh, kde bylo kvůli malému počtu pacientů (antidiabetika, antiobezitika) těžké dokončit úlohu kvůli nízko nastaveným kvantifikátorům, jsem tento parametr změnil na 1000, případně 10000.

Další parametry jsem v systému LISp-Miner nechal nastavené na defaultních hodnotách.

7.2. Práce s výsledky

U většiny úloh vygeneroval systém LISp-Miner desítky nebo stovky pravidel. Bylo proto nutné, abych z těchto pravidel vybral ty nejzákladnější a nejpřínosnější. Práce s výsledky by se nechala popsat jako filtrace a interpretace.

Samotný systém LISp-Miner nabízí hned několik možností práce s vygenerovanými pravidly. Jsou to řazení podle různých kritérií, vytváření skupin, filtrování výsledků. V mé práci jsem použil především vytváření skupin.

24 V systému LISp-Miner je možné nastavit maximální počet pravidel, které může systém vygenerovat. Po překročení tohoto počtu se generování zastaví.

25 Pravidla s takovými charakteristikami jsem ovšem nezískal u implikačních otázek týkajících se antidiabetik a antiobezitik.

Zde je popis mé práce s výsledky:

A) Výběr posledních pravidel z „řady“

V úlohách s číselnými hodnotami systém LISp-Miner generoval celé řady pravidel se stejnými literály. V pravidlech se lišily pouze hodnoty řezů viz obrázek č. 1. Z této řady jsem vybral vždy poslední pravidlo do zvlášť vytvořené skupiny *Výběr*. V případě znázorněném na obrázku č. 1 je to pravidlo č. 55.

B) **Odstranění duplicitních pravidel**

Protože řady pravidel se stejnými literály se ve výsledcích často objevovaly několikrát, resp. byly přerušené jinými řadami, bylo nutné ve vzniklé skupině *Výběr* odstranit duplicitní pravidla. Dvě pravidla se stejnými literály jsem nechal pouze v případě, že se lišila jejich hodnota fundované implikace o desítku procent (100%, 90%, 80%). Výběr těchto pravidel naleznete v příloze číslo 2.

C) **Odstranění pravidel s nerelevantními literály**

U každého pravidla jsem zkontroloval hodnoty řezů literálů, zda alespoň trochu (na desetinná místa) odpovídají hraničním hodnotám pro kategorii *vysoká* (viz příloha č. 3.). Pokud tomu tak nebylo, předpokládal jsem, že tento literál je v pravidle pouze „do počtu“[26] a pravidlo jsem vyřadil. Šlo většinou o pravidla s antecedentem o délce 3.

D) **Odstranění logických prodloužení**

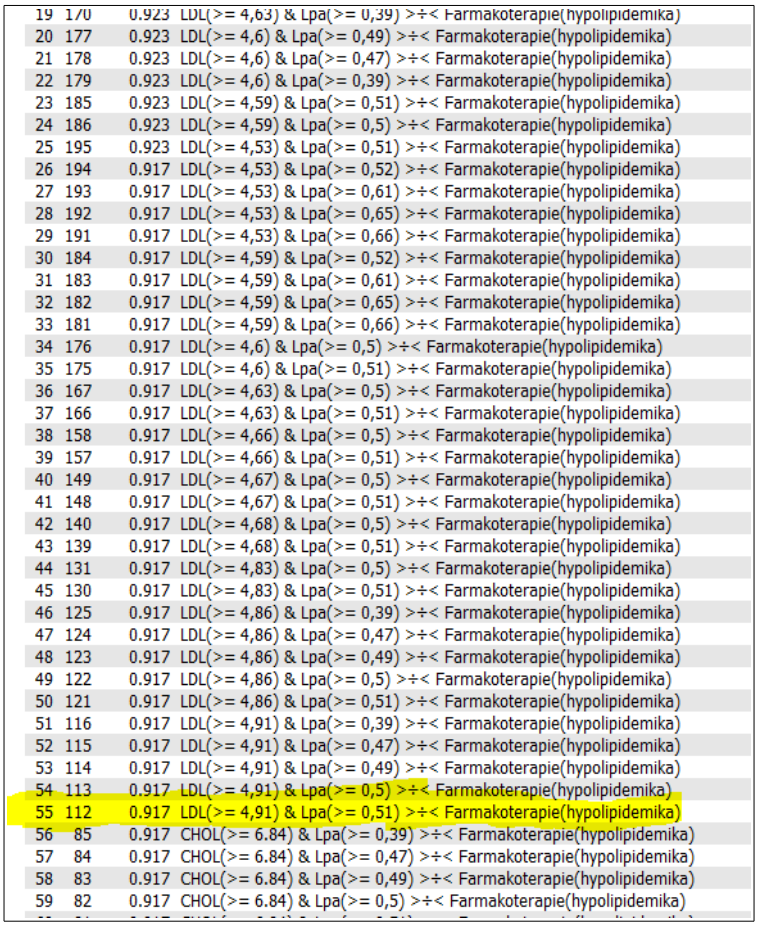
Co rozumím pod pojmem logické *prodloužení*. Pravidlo A s antecedentem o délce 2 a literály XY má hodnotu fundované implikace 100%. Pravidlo B s antecedentem o délce 3 a literály XYZ má hodnotu fundované implikace 100%. Pravidlo B je pouze logickým prodloužením pravidla A.

E) **Analýza hlavních vztahů**

Některé literály se opakují ve většině nalezených pravidel, a lze tedy předpokládat, že hrají v dané analytické otázce hlavní roli.

Práce s nalezenými pravidly se opět lišila podle toho, zda šlo o pravidla s číselnými hodnotami nebo s kategoriálními hodnotami. Krok C se u pravidel s kategoriálními hodnotami velmi zjednodušil, protože bylo nutné odstranit pouze pravidla s literálem abúzus s dvojicí hodnot *normální, potencionálně rizikový* nebo *normální, rizikový*.

26 Viz výchozí předpoklad z 7.1.3.



19	170	0.923	LDL(>= 4,63) & Lpa(>= 0,39) >÷< Farmakoterapie(hypolipidemika)
20	177	0.923	LDL(>= 4,6) & Lpa(>= 0,49) >÷< Farmakoterapie(hypolipidemika)
21	178	0.923	LDL(>= 4,6) & Lpa(>= 0,47) >÷< Farmakoterapie(hypolipidemika)
22	179	0.923	LDL(>= 4,6) & Lpa(>= 0,39) >÷< Farmakoterapie(hypolipidemika)
23	185	0.923	LDL(>= 4,59) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
24	186	0.923	LDL(>= 4,59) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
25	195	0.923	LDL(>= 4,53) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
26	194	0.917	LDL(>= 4,53) & Lpa(>= 0,52) >÷< Farmakoterapie(hypolipidemika)
27	193	0.917	LDL(>= 4,53) & Lpa(>= 0,61) >÷< Farmakoterapie(hypolipidemika)
28	192	0.917	LDL(>= 4,53) & Lpa(>= 0,65) >÷< Farmakoterapie(hypolipidemika)
29	191	0.917	LDL(>= 4,53) & Lpa(>= 0,66) >÷< Farmakoterapie(hypolipidemika)
30	184	0.917	LDL(>= 4,59) & Lpa(>= 0,52) >÷< Farmakoterapie(hypolipidemika)
31	183	0.917	LDL(>= 4,59) & Lpa(>= 0,61) >÷< Farmakoterapie(hypolipidemika)
32	182	0.917	LDL(>= 4,59) & Lpa(>= 0,65) >÷< Farmakoterapie(hypolipidemika)
33	181	0.917	LDL(>= 4,59) & Lpa(>= 0,66) >÷< Farmakoterapie(hypolipidemika)
34	176	0.917	LDL(>= 4,6) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
35	175	0.917	LDL(>= 4,6) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
36	167	0.917	LDL(>= 4,63) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
37	166	0.917	LDL(>= 4,63) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
38	158	0.917	LDL(>= 4,66) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
39	157	0.917	LDL(>= 4,66) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
40	149	0.917	LDL(>= 4,67) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
41	148	0.917	LDL(>= 4,67) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
42	140	0.917	LDL(>= 4,68) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
43	139	0.917	LDL(>= 4,68) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
44	131	0.917	LDL(>= 4,83) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
45	130	0.917	LDL(>= 4,83) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
46	125	0.917	LDL(>= 4,86) & Lpa(>= 0,39) >÷< Farmakoterapie(hypolipidemika)
47	124	0.917	LDL(>= 4,86) & Lpa(>= 0,47) >÷< Farmakoterapie(hypolipidemika)
48	123	0.917	LDL(>= 4,86) & Lpa(>= 0,49) >÷< Farmakoterapie(hypolipidemika)
49	122	0.917	LDL(>= 4,86) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
50	121	0.917	LDL(>= 4,86) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
51	116	0.917	LDL(>= 4,91) & Lpa(>= 0,39) >÷< Farmakoterapie(hypolipidemika)
52	115	0.917	LDL(>= 4,91) & Lpa(>= 0,47) >÷< Farmakoterapie(hypolipidemika)
53	114	0.917	LDL(>= 4,91) & Lpa(>= 0,49) >÷< Farmakoterapie(hypolipidemika)
54	113	0.917	LDL(>= 4,91) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)
55	112	0.917	LDL(>= 4,91) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)
56	85	0.917	CHOL(>= 6.84) & Lpa(>= 0,39) >÷< Farmakoterapie(hypolipidemika)
57	84	0.917	CHOL(>= 6.84) & Lpa(>= 0,47) >÷< Farmakoterapie(hypolipidemika)
58	83	0.917	CHOL(>= 6.84) & Lpa(>= 0,49) >÷< Farmakoterapie(hypolipidemika)
59	82	0.917	CHOL(>= 6.84) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)

Obrázek č. 1

7.3. Zhodnocení lékařem

Výsledky jednotlivých analytických otázek zhodnotil lékař v řízeném rozhovoru. Osnova tohoto rozhovoru je v příloze č. 4. V zásadě šlo o srovnání výsledků se znalostmi lékaře a zařazení výsledků do skupin (*obecně známé*, *odborně známé*, *překvapivé*, *nesouvisející*).

7.4. Výsledky úloh

Charakteristiky pravidel vycházejí z hodnot *a*, *b*, *c*, *d* ze čtyřpolní tabulky četností (tabulka č. 2)

	závěr platí	závěr neplatí
předpoklad platí	a	b
předpoklad neplatí	c	d

tabulka č. 2

Pořadí literálů v pravidlech je dáno pořadím literálů v zadání úlohy. Nemá nic společného s důležitostí literálů.

Každá následující sekce má následující strukturu:

- Číselné hodnoty,
  - Úlohy[27] (nastavení úloh, z kterých jsem vybral pravidla),
  - Pravidla (výběr pravidel),
- Kategoriální hodnoty,
  - Úlohy[28] (nastavení úloh, z kterých jsem vybral pravidla),
  - Pravidla (výběr pravidel),
- Popis výsledků (slovní popis úloh a pravidel),
- Hlavní vztahy,
- Zhodnocení lékařem ,
- Závěr.

7.4.1 Kdy jsou předepsána antihypertenziva?

Číselné hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
číslo	vyšetření, osob. údaje	2	farmakoterapie (antihypertenziva)	148	fundovaná implikace	90	20

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
Apo_B(>= 1.01) & KREA(>= 115) >÷< Farmakoterapie(antihypertenziva)	100	20,00	1,72		2	0	8	106
KREA(>= 110) & TG(>= 1.81) >÷< Farmakoterapie(antihypertenziva)	100	30,00	2,59		3	0	7	106

27 Seznam všech provedených úloh je v příloze č. 1. Zde jen o výběr těch nosných.
28 Seznam všech provedených úloh je v příloze č. 1. Zde jen o výběr těch nosných.

Kategoriální hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
kategorie	vyšetření, osob. údaje	2	farmakoterapie (antihypertenziva)	17	fundovaná implikace	60	20
kategorie	vyšetření, osob. údaje	3	farmakoterapie (antihypertenziva)	675	fundovaná implikace	60	20

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
TG(vysoka, patologicka) & KREA(vysoka) >÷< Farmakoterapie(antihypertenziva)	100	20,00	1,74		2	0	8	105

Popis výsledků

U nalezených pravidel bych chtěl zdůraznit tři fakta: velmi malá podpora, exploze počtu pravidel o délce 3 a přítomnost vyšetření **kreatininu** ve většině pravidel.

Malá podpora je dána celkově malým množstvím pacientů, kterým byla předepsána antihypertenziva. Exploze pravidel u délky antecedentu 3 a přítomnost vyšetření kreatininu u všech pravidel, ukazuje pravděpodobně na základní vztahy.

Hlavním důvodem pro předepsání antihypertenziv se podle nalezených pravidel zdá být naměřená hodnota **kreatininu** (KREA) => 110 uml/l. To samo o sobě ale nestačí. K předepsání léku musí kreatinin vstoupit do kombinace s hodnotou **triglyceridů** (TG) => 1,8 mmol/l nebo také hodnotou **Apolipoproteinu B** (Apo B) => 1 g/l. U nalezených pravidel se ještě objevu hodnota **gama-glutamyltransferázy** (GMT) => 0,4-0,8 ukat/l a diastolický **tlak** => 120, tyto hodnoty jsou ale v intervalu normálních hodnot. Všechna pravidla o délce 3 jsou kombinací vyšetření KREA + TG + libovolné vyšetření, KREA + TLAK + LIBOVOLNÉ VYŠETŘENÍ nebo KREA + Apo B + LIBOVOLNÉ VYŠETŘENÍ. Z toho lze tedy usuzovat na základní vztahy.

Podobné jsou i výsledky úloh s kategoriálními atributy. Zde ještě více vyplývá hlavní úloha vyšetření **kreatininu** a **triglyceridů.** Jako doplňkové faktory tu mohou být interpretovány ještě **mužské pohlaví, riziková rodinná anamnéza** a vyšetření **Apo B, Apo A, ALT** a **CK**.

Základní vztahy pro předepsání antihypertenziv

KREA (=>110 uml/l), TG (=>1,8 mmol/l), Apo B (=>1 g/l),

KREA (vysoká), TG (vysoká, patologická)

Zhodnocení lékařem

Hlavním kritériem pro předepsání antihypertenziv je opakovaně zvýšený tlak => 140. Zvýšené hodnoty kreatininu jsou indikátorem pro onemocnění ledvin, které může mít vliv na zvýšený tlak. Zvýšené hodnoty trygliceridu a apolipoproteinu B a gama-glutamyltransferázy zase indikují poruchu metabolismu tuků, která většinou vede ke zvýšenému tlaku. Zmíněná vyšetření jsou pro rozhodování spíše vedlejší. Tyto souvislosti patří do skupiny *odborně známé.*

7.4.2 Kdy jsou předepsána antidiabetika?

Číselné hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
číslo	vyšetření, osob. údaje	2	farmakoterapie (antidiabetika)	385	fundovaná implikace	80	25

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
Apo_A(>= 1.45) & Glukoza(>= 6.4) >÷< Farmakoterapie(antidiabetika)	1,00	25,00	3,45		4	0	12	100
BMI(>= 30) & TSH(>= 3,14) >÷< Farmakoterapie(antidiabetika)	1,00	25,00	3,45		4	0	12	100
CHBA(>= 44) & KMOC(>= 378) >÷< Farmakoterapie(antidiabetika)	1,00	25,00	3,81		4	0	12	89
CHBA(>= 43) & TSH(>= 3,73) >÷< Farmakoterapie(antidiabetika)	1,00	28,57	3,77		4	0	10	92
Glukoza(>= 7) & KMOC(>= 377) >÷< Farmakoterapie(antidiabetika)	1,00	25,00	3,45		4	0	12	100
Glukoza(>= 6.8) & RA(r) >÷< Farmakoterapie(antidiabetika)	0,86	37,50	5,36		6	1	10	95
ALT(>= 0.62) & Glukoza(>= 5.9) >÷< Farmakoterapie(antidiabetika)	0,83	31,25	4,31		5	1	11	99
CHBA(>= 45) & Lpa(>= 0,29) >÷< Farmakoterapie(antidiabetika)	0,83	31,25	5,26		5	1	11	78
Glukoza(>= 6.5) & Lpa(>= 0,29) >÷< Farmakoterapie(antidiabetika)	0,83	31,25	4,31		5	1	11	99
Glukoza(>= 6.4) & TSH(>= 2,57) >÷< Farmakoterapie(antidiabetika)	0,83	33,33	4,35		5	1	10	99

Kategoriální hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
kategorie	vyšetření, osob. údaje	3	farmakoterapie (antidiabetika)	6	fundovaná implikace	70	18,75

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
OA(r) & ALT(vysoka) & Glukoza(vysoka) >÷< Farmakoterapie(antidiabetika)	100	25,00	3,51		4	0	12	98
RA(r) & Glukoza(vysoka) & CRP(vysoka) >÷< Farmakoterapie(antidiabetika)	75	18,75	2,65		3	1	13	96
ALT(vysoka) & Glukoza(vysoka) & GHBA(vysoka) >÷< Farmakoterapie(antidiabetika)	75	18,75	2,61		3	1	13	98
ALT(vysoka) & Glukoza(vysoka) & Apo_B(vysoka, patologicka) >÷< Farmakoterapie(antidiabetika)	75	18,75	2,61		3	1	13	98

Popis výsledků

Na první pohled jsou nápadné tyto věci: pravidel je poměrně hodně a mají poměrně dobré charakteristiky spolehlivosti. Pokrytí je opět velmi malé, což je dáno malým počtem pacientů. Pravidla s číselnými hodnotami ale vždy obsahují nějakou prahovou hodnotu, která je v rozmezí normální. Zcela chybí pravidla o délce 1, pravidel o délce 2 je málo a nejsou vzhledem k jejich malé spolehlivosti příliš použitelná.

Vyšetření, která se objevují ve valné většině pravidel, jsou: **glukóza**, **glykohemoglobin** (GHBA). Častěji se objevují atributy **věk nad 60 let**, *riziková* **osobní nebo rodinná anamnéza** (OA, RA) a vysoká hodnota **HDL cholesterolu** (HDL). Méně často se v pravidlech objevuje **mužské pohlaví, nadváha**, případně ještě lehce zvýšené hodnoty **kyseliny močové**.

Základní vztahy pro předepsání antidiabetik

GHBA (=> 42 mmol/mol), glukóza (=> 6 mmol/l)

glukóza (vysoká), GHBA (vysoká)

Zhodnocení lékařem

Hlavním kritériem pro předepsání antidiabetik je hodnota glukózy nad 7. Narůstajícím trendem je předepisovat antidiabetika pacientům s hodnotu glukózy pod 7, pokud je u nich velká pravděpodobnost, že cukrovku v budoucnu dostanou. Glykohemoglobín je s glukózou ve vztahu přímé úměry. Tyto znalosti patří do skupiny *odborně známé*. Do skupiny *všeobecně známé* bychom mohli zařadit věk, osobní anamnézu a rodinnou anamnézu. Nadváha patří mezi hlavní rizikové faktory. Kyselina močová je indikátorem pro nadvýživu a nadváhu. Přítomnost mužského pohlaví lékaře *překvapila*. Cholesterol by neměl mít s diabetem žádnou souvislost.

7.4.3 Kdy jsou předepsána antiobezitika?

Číselné hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
číslo	vyšetření, osob. údaje	3	farmakoterapie (antiobezitika)	1404	fundovaná implikace	60	33,33

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
BMI(>= 34.6) & CRP(>= 6) & Lpa(>= 0,25) >÷< Farmakoterapie(antiobezitika)	83	33,33	4,31		5	1	10	100
BMI(>= 33.5) & Glukoza(>= 5.6) & TG(>= 1.48) >÷< Farmakoterapie(antiobezitika)	83	33,33	4,31		5	1	10	100
ALT(>= 0.23) & BMI(>= 40.4) & CRP(>= 6) >÷< Farmakoterapie(antiobezitika)	71	33,33	4,31		5	2	10	99
BMI(>= 39.5) & CRP(>= 6) & TG(>= 1.13) >÷< Farmakoterapie(antiobezitika)	71	33,33	4,31		5	2	10	99

Kategoriální hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
kategorie	vyšetření, osob. údaje	3	farmakoterapie (antiobezitika)	21	fundovaná implikace	60	20

Pravidla

Pravidla	Spolehlivost(%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
Pohlavi(z) & Abusus(p) & CRP(vysoka) >÷< Farmakoterapie(antiobezitika)	60	21,43	2,63		3	2	11	98
Pohlavi(m) & Glukoza(vysoka) & CRP(vysoka) >÷< Farmakoterapie(antiobezitika)	60	20,00	2,61		3	2	12	98

Popis výsledků

Na první pohled jsou u nalezených pravidel nápadné dvě věci: malé pokrytí (absolutní hodnoty) a malá spolehlivost. Pravidel je relativně málo. Velké počty u číselných atributů jsou způsobeny typem koeficientu right cut (posloupnosti stejných pravidel). Literály u pravidel s číselnými hodnotami mají hodnoty řezů často v rozmezí normálních hodnot. Zcela chybí pravidla o délce antecedentu 1 u úloh s oběma typy hodnot a u úloh s kategoriálními hodnotami zcela chybí i pravidla o délce 2.

Z pravidel vyplývá, že hlavní pro předepsání antiobezitik je **nadváha** až **obezita** a vysoká hodnota **C-reaktivního protein**u (CRP). V pravidlech se ještě objevují vysoké až patologické hodnoty u vyšetření **lipoproteinu malé a** (Lpa)

Základní vztahy pro předepsání antiobezitik

BMI (=> 33), CRP (=> 6 mg/l)

BMI (nadváha, obezita), CRP (vysoká),

Zhodnocení lékařem

Hlavním kritériem pro předepsání antiobezitik je obezita. CRP souvisí s obezitou. CRP je ukazatelem zánětu, který se vyskytuje v tukové tkáni u obézních pacientů. Jde o skupinu *odborně známé*. Vysoké hodnoty lipoproteinu malé a s antiobezitiky nijak nesouvisí.

7.4.4 Kdy jsou předepsána hypolipidemika?

Číselné hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
číslo	vyšetření, osob. údaje	2	farmakoterapie (hypolipidemika)	204	fundovaná implikace	90	24,39
číslo	vyšetření, osob. údaje	3	farmakoterapie (hypolipidemika)	261	fundovaná implikace	95	24,39

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
LDL(>= 3,61) & Vek(>= 60) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,62		10	1	31	74
LDL(>= 5,14) & Lpa(>= 0,5) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,62		10	1	31	74
CHOL(>= 7.29) & Lpa(>= 0,51) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,62		10	1	31	74
Apo_B(>= 1.46) & Lpa(>= 0,29) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,62		10	1	31	74
Apo_B(>= 1.5) & RA(r) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,70		10	1	31	73
Abusus(p, r) & LDL(>= 4,63) >÷< Farmakoterapie(hypolipidemika)	91	25,64	8,85		10	1	29	73
LDL(>= 5,14) >÷< Farmakoterapie(hypolipidemika)	82	35,90	12,39		14	3	25	71
Apo_B(>= 1.46) >÷< Farmakoterapie(hypolipidemika)	80	29,27	10,00		12	3	29	72
CHOL(>= 7.63) >÷< Farmakoterapie(hypolipidemika)	77	24,39	8,62		10	3	31	72

Kategoriální hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
kategorie	vyšetření, osob. údaje	2	farmakoterapie (hypolipidemika)	12	fundovaná implikace	60	24,39
kategorie	vyšetření, osob. údaje	3	farmakoterapie (hypolipidemika)	600	fundovaná implikace	60	24,39

Pravidla

Pravidla	Spolehlivost(%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
Vek(60-100) & LDL(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,70		10	1	31	73
Abusus(p, r) & BMI(nadvaha) & LDL(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	83	24,39	8,70		10	2	31	72
BMI(nadvaha) & CHOL(vysoka, patologicka) & Lpa(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	81	31,71	11,30		13	3	28	71
BMI(nadvaha) & HDL(vysoka) & Lpa(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	79	26,83	9,57		11	3	30	71
Vek(60-100) & CHOL(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	71	24,39	8,70		10	4	31	70
Vek(60-100) & Apo_B(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	83	24,39	8,70		10	2	31	72

Popis výsledků

Hypolipidemika byla předepsána 41 pacientům. Jde tedy o lék, kde jsou k dispozici statisticky relevantní data. Bohužel při zadání BASE => 20, jsou vygenerována pravidla se spolehlivostí <= 0,7. Proto jsem zadal BASE => 10. Vznikly tak poměrně spolehlivá pravidla s relativně malou podporou. Ve výsledcích se objevilo i několik poměrně kvalitních pravidel o délce 1 a 2.

Základní vztahy jsou patrné při prvním pohledu. Hlavní roli hrají vyšetření **cholesterolu** (celkový a LDL) a **apolipoproteinu B**, dále pak **věk** a **body mass index**. V pravidlech se také objevuje vyšetření **lipoproteinu malé a**, **poměr apolipoproteinu A a B**, vyšetření **HDL**, **abusus**, **rodinná anamnéza** nebo **osobní anamnéza**.

Základní vztahy pro předepsání hypolipidemik

Apo B (=>1,4 g/l), LDL (=>3,11-5,14 mmol/l), Lpa (=>0,27 g/l)

věk (=>60), BMI (nadváha, obezita), CHOL(vysoká, patologická), LDL (vysoká, patologická)

Zhodnocení lékařem

Hlavním hlediskem pro předepsání hypolipidemik je vysoká hodnota cholesterolu. LDL cholesterol je „škodlivá“ frakce celkového cholesterolu a je s touto hodnotou ve vztahu přímé úměry. Apolipoprotein B je látka obsažená v cholesterolu LDL a je tak s touto hodnotou také ve vztahu přímé úměry. Poměr apolipoproteinu A a apolipoproteinu B logicky vychází z hodnot apolipoproteinu B. Lipoprotein malé a přispívá také ke sledovaným onemocněním a jeho vysoké hodnoty se stávají důležitým faktorem pro rozhodování zvlášť v kombinaci s relativně nižšími hodnotami cholesterolu. Tyto faktory patří k *odborně známým*. K *všeobecně známým* faktorům patří věk a nadváha.

7.4.5 Kdy není předepsána žádná farmakoterapie?

Číselné hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
číslo	vyšetření, osob. údaje	2	farmakoterapie (bez léčby)	77	fundovaná implikace	70	24,24
číslo	vyšetření, osob. údaje	3	farmakoterapie (bez léčby)	498	fundovaná implikace	90	33,33

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
Vek(<= 53) & RA(n) & CHOL(<= 6.36) >÷< Farmakoterapie(bez lecby)	100	36,36	10,71		12	0	21	79
Vek(<= 53) & RA(n) & Lpa(<= 0,97) >÷< Farmakoterapie(bez lecby)	100	36,36	10,91		12	0	21	77
RA(n) & Tlak(<= 125) & Lpa(<= 0,38) >÷< Farmakoterapie(bez lecby)	100	33,33	9,82		11	0	22	79
CHOL(<= 5.83) & Glukoza(<= 5.1) & KREA(<= 80) >÷< Farmakoterapie(bez lecby)	92	33,33	9,48		11	1	22	82
ALT(<= 0.41) & Lpa(<= 0,27) & MALB(<= 11) >÷< Farmakoterapie(bez lecby)	92	33,33	9,57		11	1	22	81
BMI(<= 28.8) & ALT(<= 0.31) & LDL(<= 4,59) >÷< Farmakoterapie(bez lecby)	92	33,33	9,48		11	1	22	82
BMI(<= 28.6) & CHOL(<= 6.8) & KMOC(<= 285) >÷< Farmakoterapie(bez lecby)	92	33,33	9,48		11	1	22	82
BMI(<= 27.8) & CHOL(<= 6.33) & Glukoza(<= 5.2) >÷< Farmakoterapie(bez lecby)	92	33,33	9,48		11	1	22	82
BMI(<= 26.3) & Glukoza(<= 5.3) & LDL(<= 4,37) >÷< Farmakoterapie(bez lecby)	92	33,33	9,48		11	1	22	82
Vek(<= 50) & BMI(<= 26.3) & LDL(<= 4,37) >÷< Farmakoterapie(bez lecby)	92	33,33	9,48		11	1	22	82
Vek(<= 43) & BMI(<= 28.6) & CHOL(<= 6.14) >÷< Farmakoterapie(bez lecby)	92	33,33	9,48		11	1	22	82
Vek(<= 45) & RA(n) >÷< Farmakoterapie(bez lecby)	90	27,27	7,96		9	1	24	79
RA(n) & Glukoza(<= 5.2) >÷< Farmakoterapie(bez lecby)	90	27,27	8,04		9	1	24	78
RA(n) & Tlak (<=122) >÷< Farmakoterapie(bez lecby)	75	36,36	10,91		12	4	21	73

Kategoriální hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
kategorie	vyšetření, osob. údaje	2	farmakoterapie (bez léčby)	22	fundovaná implikace	60	24,24
kategorie	vyšetření, osob. údaje	3	farmakoterapie (bez léčby)	662	fundovaná implikace	60	24,24

Pravidla

Pravidla	Spolehlivost(%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
Vek(<= 40-50) & RA(n) & LDL(normalni) >÷< Farmakoterapie(bez lecby)	100	24,24	7,08		8	0	25	80
RA(n) & Tlak(normalni) & Lpa(normalni) >÷< Farmakoterapie(bez lecby)	100	24,24	7,14		8	0	25	79
RA(n) & OA(n) & Abusus(n, r) >÷< Farmakoterapie(bez lecby)	89	24,24	7,02		8	1	25	80
Vek(<= 50-60) & RA(n) & LDL(normalni) >÷< Farmakoterapie(bez lecby)	89	24,24	7,21		8	1	25	77
Vek(<= 40-50) & RA(n) & OA(n) >÷< Farmakoterapie(bez lecby)	89	24,24	7,02		8	1	25	80
Vek(<= 40-50) & RA(n) >÷< Farmakoterapie(bez lecby)	80	36,36	10,71		12	3	21	76
Pohlavi(z) & RA(n) >÷< Farmakoterapie(bez lecby)	75	27,27	8,00		9	3	24	73

Popis výsledků

Bez léčby farmaky odešlo z poradny 33 pacientů ze 115 členného sledovaného souboru. Jde tedy o úlohu, kde jsou k dispozici statisticky relevantní data. Bohužel při zadání minimální podpory 20, jsou vygenerována pravidla se spolehlivostí <= 0,7. Proto jsem zadal minimální podporu pouze 8, tedy zhruba jednu čtvrtinu ze 33 pacientů. Vznikla tak poměrně spolehlivá pravidla, ale s malou podporou. Ve výsledcích se objevilo několik kvalitních pravidel o délce 2. V zadání úloh s numerickými atributy je na rozdíl od předchozích otázek použit left cut.

Hlavní roli pro nepředepsání léků hrají **věk pod 50 let** a příbuzní bez rizikových onemocnění (**rodinná anamnéza**). Roli také hraje dosavadní zdraví (**osobní anamnéza**), **ženské pohlaví** a dobré hodnoty **tlaku, BMI, LDL cholesterolu**, nebo **glukózy**.

Základní vztahy pro nepředepsání léků

věk (<=40, 50), RA (neriziková)

věk (<=40, 50), RA (neriziková), BMI (<=27),

Zhodnocení lékařem

Výsledky patří do skupiny *všeobecně známé.* U laboratorních vyšetření jde většinou o obrácené hodnoty k předešlým analytickým otázkám.

7.4.6 Která z těchto pravidel nejvíce pomáhají při rozhodování?

Zde jsem vytvořil výběr pravidel, která mají podporou =>7% a zároveň spolehlivost => 90%. Výběr probíhal ze všech pravidel v příloze č. 2. Výběr se týká pouze pravidel s kategoriální hodnotou.

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
RA(n) & Tlak(normalni) & Lpa(normalni) >÷< Farmakoterapie(bez lecby)	100	24,24	7,14		8	0	25	79
Vek(<= 40-50) & RA(n) & LDL(normalni) >÷< Farmakoterapie(bez lecby)	100	24,24	7,08		8	0	25	80
Vek(60-100) & LDL(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,7		10	1	31	73
Vek(60-100) & CHOL(vysoka, patologicka) & Apo_A_Apo_B(nizka) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,7		10	1	31	73
Vek(60-100) & CHOL(vysoka, patologicka) & LDL(vysoka, patologicka) >÷< Farmakoterapie(hypolipidemika)	91	24,39	8,7		10	1	31	73
Vek(<= 50-60) & RA(n) & LDL(normalni) >÷< Farmakoterapie(bez lecby)	89	24,24	7,21		8	1	25	77
Vek(<= 40-50) & RA(n) & OA(n) >÷< Farmakoterapie(bez lecby)	89	24,24	7,02		8	1	25	80

7.4.7 Kdy je předepsána nějaká farmakoterapie?

Číselné hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
číslo	vyšetření, osob. údaje	1	farmakoterapie (léčba)	29	fundovaná implikace	90	24,39
číslo	vyšetření, osob. údaje	2	farmakoterapie (léčba)	688	fundovaná implikace	10	30,49

Pravidla

Pravidla	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
Lpa(>= 0,08) & Tlak(>= 132) >÷< Farmakoterapie (01)(lecba)	100	37,80	26,72		31	0	51	34
KMOC(>= 350) & Vek(>= 42) >÷< Farmakoterapie (01) (lecba)	100	34,15	24,14		28	0	54	34
Glukoza(>= 6) >÷< Farmakoterapie (01)(lecba)	100	34,15	24,14		28	0	54	34
Tlak(>= 131) & Vek(>= 48) >÷< Farmakoterapie (01) (lecba)	100	32,93	23,28		27	0	55	34
Glukoza(>= 5.7) >÷< Farmakoterapie (01)(lecba)	94	39,02	27,59		32	2	50	32
Vek(>= 50) >÷< Farmakoterapie (01)(lecba)	91	64,63	45,69		53	5	29	29
Apo_A(>= 1.54) >÷< Farmakoterapie (01)(lecba)	91	25,61	18,10		21	2	61	32

CHBA(>= 41) >÷< Farmakoterapie (01)(lecba)	91	50,82	43,06		31	3	30	8
Tlak(>= 137) >÷< Farmakoterapie (01)(lecba)	91	24,39	17,24		20	2	62	32
Lpa(>= 0,61) >÷< Farmakoterapie (01)(lecba)	90	34,15	24,14		28	3	54	31

Kategoriální hodnoty

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
kategorie	vyšetření, osob. údaje	1	farmakoterapie (léčba)	68	fundovaná implikace	60	24,39
kategorie	vyšetření, osob. údaje	2	farmakoterapie (léčba)	85	fundovaná implikace	60	24,39

Pravidla

Pravidla	Spolehlivost(%) a/(a+b)	Pokrytí (%) a/(a+c)	Podpora (%) a/(a+b+c+d)		a	b	c	d
RA(r) & Glukoza(vysoka) >÷< Farmakoterapie(lecba)	96	30,77	21,62		24	1	54	32
Pohlavi(m) & Vek(50-60, 60-100) >÷< Farmakoterapie(lecba)	96	28,05	20,00		23	1	59	32
Tlak(zvyseny, vysoky) & Glukoza(vysoka) >÷< Farmakoterapie(lecba)	95	24,39	17,39		20	1	62	32
Vek(50-60, 60-100) >÷< Farmakoterapie(lecba)	91	64,63	46,09		53	5	29	28
RA(r) & BMI(obezita) >÷< Farmakoterapie(lecba)	91	27,03	18,69		20	2	54	31
Glukoza(vysoka) >÷< Farmakoterapie(lecba)	90	43,90	31,30		36	4	46	29
Tlak(zvyseny, vysoky) >÷< Farmakoterapie(lecba)	89	41,46	29,57		34	4	48	29

Popis výsledků

S nějakým lékem odešlo z poradny 82 pacientů. Při tomto počtu se mi podařilo nalézt pravidla s velmi dobrými hodnotami spolehlivosti i podpory. Na rozdíl od předchozích analytických otázek si zde vystačíme s pravidly o délce 1 a 2. Pravidla o délce 3 jsou v podstatě jen nevýznamná rozšíření těchto pravidel.

Hlavní vztahy by se nechaly shrnout následujícím způsobem. Pokud je pacientovi, který přijde do ordinace, **více než 60 let** nebo má vysokou hodnotu **glukózy**, je velmi pravděpodobné, že mu bude předepsán nějaký lék, který doposud nebral. Skutečnost, že pacient již nějaké potíže má nebo měl **(OA)**, stejně jako **zvýšený a vysoký tlak,** hrají pro předepsání dalších nových léčiv také významnou roli. Dalšími faktory, které se v pravidlech objevují, je přítomnost sledovaných onemocnění v rodině pacienta **(RA)**, **nadváha až obezita, mužské pohlaví** a vysoké hodnoty **cholesterolu HDL**, **lipoproteinu malé** a (Lpa) a **glykohemoglobínu** (GHBA).

Základní vztahy pro předepsání léků

věk (=>52), glukóza (=>6 mmol/l), tlak (=>137),

věk (=>50, =>60), OA (riziková), glukóza (vysoká), tlak (zvýšení, vysoký),

Zhodnocení lékařem

Většina výsledků patří do skupin *odborně známé* a *všeobecně známé*. Věk je hlavní faktorem všech civilizačních chorob. Nejvíce léků si většinou odnesou z poradny pacienti s diabetem. *Překvapivým* byla pro lékaře přítomnost vyšetření apolipoproteinu A. Jde spíše o doplňkové vyšetření.

7.4.8 Která vyšetření jsou pro rozhodování klíčová?

Zde jsem provedl souhrn všech údajů, anamnéz a vyšetření, která se objevila v hlavních vztazích v předešlých sekcích. Cílem je zpřehlednění hlavních vztahů.

Osobní údaje

- věk

Anamnézy

- osobní, rodinná

Objektivní vyšetření

- tlak, BMI

Laboratorní vyšetření

- *sacharidový metabolismus*: glukóza, glykohemoglobin,
- *lipidový metabolismus*: triglyceridy, celkový cholesterol, LDL cholesterol, HDL cholesterol, lipoprotein malé a, apolipoprotein B, poměr apolipoproteinu A/apolipoproteinu B
- *dusíkaté látky*: kreatinin
- *bílkoviny*: C-reaktivní protein (CRP)

Z právě uvedeného by vyplývalo, že následující údaje, anamnézy a vyšetření jsou pro rozhodování spíše **vedlejší**: osobní údaje: pohlaví; anamnézy: abúzus; laboratorní vyšetření: enzymy (ALT, GMT, CK), kyselina močová; tyreotropin (TSH) a mikroalbuminurie.

Pokud bych měl učinit nějaký slovní popis, tak nápadná je významnost většiny vyšetření týkajících se metabolismu tuků a cukrů. Naopak vedlejší roli hrají vyšetření enzymů. Zajímavé může být, že mezi významnými faktory chybí abúzus.

Hodnocení lékaře

Většina výsledků patří do skupiny *odborně známé*. *Překvapivá* byla pro lékaře vyšetření kreatininu a CRP. Kreatinin se ovšem do výběru dostal spíše náhodně díky analytické otázce „Kdy jsou předepsána antihypertenziva?“

7.4.9 Jsou některá vyšetření duplicitní

Předpokládám, že vyšetření A je duplicitní k vyšetření B právě tehdy, když existuje relativně spolehlivé pravidlo, ve kterém jsou hodnoty vyšetření A a B ve vztahu dvojité fundované implikace. Když je hodnota vyšetření A vysoká, je velká pravděpodobnost, že vyšetření B má také vysokou hodnotu. Pokud má vyšetření A hodnotu normální, je pravděpodobné, že ji má i vyšetření B.

Kvantifikátorem je u následujících úloh dvojitá fundovaná implikace, pro výpočet spolehlivosti bude sloužit D-spolehlivost. Výpočet D-spolehlivosti ze čtyřpolní tabulky je a/(a+b+c).
Úlohy s číselnými atributy nebyly u této analytické otázky prováděny.

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
kategorie	anamnézy, vyšetření (pouze hodnoty neriziková, normální)	1	anamnézy, vyšetření (pouze hodnoty neriziková, normální)	222	dvojitá fundovaná implikace	60	17,39
kategorie	vyšetření a anamnézy(pouze hodnoty , riziková, vysoká)	1	vyšetření a anamnézy(pouze hodnoty , riziková, vysoká)	54	dvojitá fundovaná implikace	60	17,39
kategorie	vybraná vyšetření (pouze hodnoty nízká)	1	vybraná vyšetření (pouze hodnoty nízká)	0	dvojitá fundovaná implikace	60	17,39

Pravidla

Pravidla (vysoká-vysoká)	D-Spolehlivost (%)a/(a+b+c)	Pravidla (normální-normální)	D- Spolehlivost (%)/(a+b+c)
OA(r) >÷< RA(r)	74	není	
RA(r) >÷< OA(r)	74	není	
LDL(vysoka) >÷< Apo_B(vysoka)	71	není	
Apo_B(vysoka) >÷< LDL(vysoka)	71	není	
LDL(vysoka) >÷< CHOL(vysoka)	70	LDL(normalni) >÷< CHOL(normalni)	69
CHOL(vysoka) >÷< LDL(vysoka)	70	CHOL(normalni) >÷< LDL(normalni)	69
Apo_B(vysoka) >÷< CHOL(vysoka)	70	není	

CHOL(vysoka) >÷< Apo_B(vysoka)	70	není	
OA(r) >÷< BMI(nadvaha, obezita)	67	není	
BMI(nadvaha, obezita) >÷< OA(r)	67	není	

Popis výsledků

Předpoklad výše splňuje pouze vyšetření **celkového cholesterolu** (CHOL) a **cholesterolu LDL** (LDL). Tyto vyšetření jsou ve vztahu celek-část. Je proto logické, že mezi nimi byla nalezena závislost.

U úlohy s normálními hodnotami se ještě vyskytla celkem pevná pravidla mezi hodnotami kreatininu a mikroalbuminurie nebo kyseliny močové. Bohužel u těchto vyšetření není ve sledovaném souboru dostatečně často zastoupena vysoká hodnota vyšetření.

Zhodnocení lékařem

Celkových cholesterol a jeho frakce cholesterol LDL jsou ve vztahu přímé úměry. V tomto vztahu by ale také měly být vyšetření cholesterolu LDL s apolipoproteinem B a glukóza s glykohemoglobínem.

7.4.10 Jak se mění pravidla, jedná-li se o muže nebo ženu? (Ukázka práce s podmínkami)

Jak již bylo zmíněno v sekci o zpracování dat 6.2.3., některé vyšetření mají rozdílné rozmezí normálních hodnot pro muže a ženy. Hodnoty některých vyšetření jsou pro muže *normální* zatímco pro ženy již *zvýšené*. U některých vyšetření je to naopak. V tabulkách č. 3 a 4. uvádím vyšetření s rozdílnou kategorizací podle pohlaví, tak jak ji používá laboratorní systém FN Plzeň. U ostatních vyšetření jsou normální hodnoty pro muže i ženy stejné.

Vyšetření, která mají vyšší normální hodnoty pro muže než pro ženy

Enzymy	Pohlaví	Normální od	Normální do
ALT	Z	0,1	0,7
ALT	M	0,1	1
GMT	Z	0,1	0,8
GMT	M	0,1	1,3
CK	Z	2,4	5
CK	M	3,2	5
Dusíkaté látky			
KREA	Z	55	100
KREA	M	70	110
KMOC	Z	140	360
KMOC	M	210	450

Tabulka č. 3

Vyšetření, která mají vyšší normální hodnoty pro ženy než pro muže

Lipidový metabolismus	Pohlaví	Normální od	Normální do
HDL	Z	1,2	2,7
HDL	M	1	2,1
Apo-A-I	Z	1,6	1,8
Apo-A-I	M	1,4	1,6
Apo-A/Apo-B	Z	1,7	1,9
Apo-A/Apo-B	M	1,5	1,7

Tabulka č. 4

Předpokládám, že rozdílné normální hodnoty jsou u vyšetření A právě tehdy, když v pravidle X, které má definovanou podmínku pohlaví ženy, má vyšetření A jinou hraniční hodnotu (cuty) než v podobně spolehlivém pravidle Y, které má nastavenou podmínku pohlaví muži a stejné literály jako pravidlo X. Naopak předpokládám, že vyšetření A nemá rozdílné normální hodnoty, když hodnoty řezů vyšetření A jsou stejné v přibližně stejně spolehlivých pravidlech X, Y. Předpoklad dvou podobně spolehlivých pravidel výše vychází z přibližně stejného počtu pacientů s mužským (22) a ženským (19) pohlavím, kterým byla předepsána hypolipidemika.

Z logiky věci jsem zadal úlohy pouze s číselnými hodnotami. Protože tato úloha má za cíl práci s podmínkami pouze demonstrovat, byly úlohy zadány pouze se sukcedentem hypolipidemika. Připomeňme, že pro předepsání hypolipidemik jsou důležitá laboratorní vyšetření Apo B, CHOL, LDL, HDL a Lpa. Pří srovnání s tabulkami výše je jasné, že v úlohách půjde spíše o to najít pravidla se shodnými řezy. Jediné vyšetření, kde by se hodnoty řezů měli pro muže a ženy lišit je vyšetření HDL.

Úlohy

typ hodnot	antecedent	délka (Ant.)	sukcedent	Podmínka	počet pravidel	kvantifikátor	min. hodnota kvantifikátoru (%)	min. hodnota BASE (%)
číslo	vyšetření, osob. údaje	1	farmakoterapie (hypolipidemika)	ž	103	fundovaná implikace	60	26,32
číslo	vyšetření, osob. údaje	1	farmakoterapie (hypolipidemika)	m	51	fundovaná implikace	60	22,73
číslo	vyšetření, osob. údaje	2	farmakoterapie (hypolipidemika)	ž	9230	fundovaná implikace	80	26,32
číslo	vyšetření, osob. údaje	2	farmakoterapie (hypolipidemika)	m	11327	fundovaná implikace	70	22,73

Pravidla z úloh s délkou antecedentu 1

Pravidla (ženy)	Spolehlivost(%) a/(a+b)	Pokrytí (%) a/(a+c)	Pravidla (muži)	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)
LDL(>= 4,6) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(ž)	73	72,73	LDL(>= 4,6) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(m)	71	29,41
CHOL(>= 6.52) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(ž)	68	77,27	CHOL(>= 6.52) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(m)	67	52,63

Pravidla z úloh s délkou antecedentu 2

Pravidla (ženy)	Spolehlivost(%) a/(a+b)	Pokrytí (%) a/(a+c)	Pravidla (muži)	Spolehlivost (%) a/(a+b)	Pokrytí (%) a/(a+c)
LDL(>= 5,32) & Lpa(>= 0,27) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(ž)	100	22,73	LDL(>= 4,83) & Lpa(>= 0,38) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(m)	100	26,32
CHOL(>= 6,75) & Lpa (>=0,29)) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(ž)	83	45	CHOL(>= 0,75) & Lpa (>=038)) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(m)	83,3	26
Apo_B (>= 1,29) & Lpa (>=0,68)) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(ž)	83	23	Apo_B (>= 1,29) & Lpa (>=0,68)) >÷< Farmakoterapie(hypolipidemika) / Pohlavi(m)	83,3	26

Popis výsledků

Soubor obsahuje celkově 41 pacientů, kterým byla předepsána hypolipidemika, z toho je 19 mužů a 22 žen. Proto jsem nastavil BASE => 5, což je přibližně jedna čtvrtina ze ženských nebo mužských pacientů.

U pravidel s antecedentem délky 1 nalezl LISp-Miner pravidla, která splňují výše zmíněné předpoklady, pouze pro vyšetření **LDL** a **CHOL**. Hodnoty řezů jsou u těchto vyšetření stejné jak pro muže i ženy. Tato vyšetření mají normální rozmezí hodnot pro muže a ženy i v kategorizaci FN Plzeň.

U pravidel s antecedentem délky 2 byl problém s velkým počtem pravidel. Proto jsem se rozhodl výsledky vyfiltrovat. Nechal jsem pouze pravidla se spolehlivostí větší než 0,8 a pouze taková, která obsahují vyšetření Apo_B, LDL, HDL, CHOL a Lpa, to znamená vyšetření, která se pro předepsání hypolipidemik ukázala jako rozhodující (viz sekce 7.4.4). U žen jsem tak počet pravidel snížil na 3402 pravidel, u mužů na 1058 pravidel. V těchto pravidlech jsem pak hledal shodná pravidla.

U shodných pravidel s antecedentem délky 2, které jsem našel, jsou hodnoty řezů stejné pro muže i ženy. Hodnoty těchto vyšetření jsou stejné u žen i mužů i v kategorizaci FN Plzeň. Bohužel se nepodařilo najít žádné pravidlo s vyšetřením HDL. Vyšetření HDL by podle kategorizace FN Plzeň mělo být pro může a ženy rozdílné.

Zhodnocení lékařem

Hodnoty u laboratorních vyšetření by měly být stejné. Pohlaví při rozhodování nicméně hraje významnou roli. Míru rizikovosti arterosklerózy mají ženy obecně posunutou o deset let. Pokud by se v pravidlech objevil věk, měla by být tedy hodnota řezu pro muže a ženy rozdílná.

7.5. Souhrnné zhodnocení

Požitá metoda data miningu funguje zvláště u těch otázek, kde byl dostatečný počet pacientů. U těchto otázek se podařilo dobře odhalit základní souvislosti. U otázek s méně pacienty, se vyskytovalo více chyb.

Většinu nalezených pravidel zařadil lékař do skupiny *odborně známé* nebo *všeobecně známé*. Především u léků s malým počtem pacientů se objevila pravidla zařazené do skupiny *překvapivé* nebo *chybné*.

Pro přesnější výsledky by bylo třeba více pacientů a lepší znalost vzorku. Lepší znalost vzorku se týká především přesnější kategorizace anamnéz nebo lepší znalosti životního stylu pacientů. Do dat by bylo dobré zahrnout dvě nebo tři kontrolní vyšetření jdoucí za sebou a naopak vyřadit vstupní vyšetření. Problém vstupních vyšetření spočívá v tom, že pacient přichází v okamžiku, kdy je mu nejhůř. Tím jsou ovlivněné výsledky laboratorních vyšetření. Po změně životosprávy může dojít k zásadní změně hodnot.

Závěr

Cíl této práce, ukázka aplikace metod data miningu na reálných datech, byl úspěšně splněn. Myslím si, že to bylo hlavně díky pečlivému výběru řešené problematiky. Její přiměřená komplexita (počet atributů 25) a přiměřená míra vágnosti (snadné vymezení základních otázek) vedla k poměrně dobrým výsledkům, a to i když jsem měl k dispozici jen relativně malý vzorek (115 pacientů).

Cesta k cíli mně ukázala to, co ostatně říká i odborná literatura. Většinu času zabere získání správných dat a jejich správné zpracování stejně jako filtrování a interpretace výsledků. Také jsem získal zkušenost, že data mining je velmi iterativní proces, když jsem se ve všech fázích častokrát vracel. Konečně z této práce je jasné, že v data miningu jde spíše o ukázání základních vztahů než o získání exaktně přesných zákonitostí a pravidel.

Podle hodnocení lékaře postihla nalezená pravidla u většiny analytických otázek základní odborně známé vztahy. Problémy byly částečně u analytických otázek, kde jsem měl k dispozici menší vzorek pacientů. U těchto otázek se občas objevila ve výsledcích pravidla, která lékař označil jako překvapivá.

Co se týče možného rozšíření nebo doplnění mé práce, nabízí se zde například vytvoření zobecněných kategorií (např. anamnézy celkově, metabolismus tuků, metabolismus cukrů apod.), použití dalších kvantifikátorů nebo použití jiného způsobu kategorizace. Je ovšem otázka, kde leží hranice, jejíž překročení vyžaduje velmi velké úsilí a jen minimální přínos ve formě nových výsledků. I vzhledem k hodnocení lékaře mám za to, že v rámci zvolené data miningové metody bylo to podstatné řečeno.

Použitá literatura

BERKA, P., 2003. *Dobývání znalostí z databází.* Vyd. 1. Praha : Academia. ISBN 80-200-1062-9.

BERKA, P., 2003b. *Aplikace systémů dobývání znalostí pro analýzu medicínských dat*. *EuroMISE centrum - Kardio* [online]. Praha: Vysoká škola ekonomická, 30.5.2003, [cit. 2014-03-16]. Dostupné z: [http://euromise.vse.cz/kdd/index.php].

BURDA, M., 2004. *Získávání znalostí z databází - Asociační pravidla (studijní materiál)*. [online]. 27.1. 2004. 34. s. [cit. 2014-03-16] Dostupné z: [http://www.fit.vutbr.cz/study/courses/ZZD/public/seminar0304/GUHA.].

BURIAN, J. a RAUCH, J., 2003. Analysis of Death Causes in the STULONG Data Set. **In**: BERKA, P. (ed.) *ECML/PKDD 2003 Discovery Challenge Workshop Notes: Proceedings of 14th European Conference on Machine Learning and 7th European Conference on Principles and Practice of Knowledge Discovery in Databases.* Zagreb: Ruder Boskovic Institute. s. 47-58. Dostupné také z: [http://euromise.vse.cz/stulong/publikace/] ISBN 953-6690-38-1.

CIOS, K. J. et al., 2007. *Data mining: a knowledge discovery approach*. New York: Springer. ISBN 978-0-387-33333-5.

ČERNÝ, Z., DOLEJŠÍ, P., RAUCH, J. a ŠEBEK, M., 2003. Dobývání znalostí v medicínských datech – případová studie. **In**: SVÁTEK, V. (ed.). *Znalosti 2003*. Ostrava: Technická Univerzita Ostrava, Fakulta Elektrotechniky a Informatiky. s. 182–191. ISBN 80-248-0229-5.

DROSTE, C. a PLANTA, M. von, 1992. *Memorix: Vademecum lékaře.* 1. vyd. Praha: Scientia Medica. ISBN 80-85526-04-2.

HÁJEK, P., HAVRÁNEK, T. a CHYTIL, M. K., 1983. *GUHA : automatická tvorba hypotéz.* Praha: Academia.

HÁJEK, P., HOLEŇA, M. a RAUCH, J., 2010. [online] The GUHA method and its meaning for data mining. *Journal of Computer and System Sciences.* San Diego: Academic Press, vol. 76(1), s. 34–48 [cit. 2014-02-10] ISSN 0022-0000. Dostupné také z: [http://cla2008.inf.upol.cz/download/cla2008-hajek.pdf].

MIKUT, R. a REISCHL, M., 2011. [online] Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* New York: John Wiley & Sons, vol. 1(5). s. 431–443. ISSN: 1942-4795. DOI: 10.1002/widm.24. Dostupné také z: [http://onlinelibrary.wiley.com/doi/10.1002/widm.24/abstract]

PROJEKT STULONG, *2003. Projekt STULONG : Přehledný popis* [online]. Praha: Vysoká škola ekonomická, [cit. 2014-03-16]. Dostupné z: [http://euromise.vse.cz/stulong/].

RAUCH, J., 2011. *Systém LISp-Miner : Stručný popis určený pro posluchače kurzů Zpracování informací a znalostí*, [online] Praha: Vysoká škola ekonomická. Dostupné z: [http://www.pejcoch.com/vyuka/4IZ210/ukol3/LM_SKRPT_11.pdf.]

RAUCH, J., 2013. Metoda GUHA a dobývání znalostí z databází. **In:** MAŘIK, V. (ed.) et al. *Umělá inteligence* (6), Vyd. 1. Praha: Academia, s. 348-389. ISBN 978-80200-2276-9.

RUD, O. P. a MAGERA, I., 2001. *Data mining: praktický průvodce dolováním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. Vyd. 1. Praha: Computer Press, ISBN 80-7226-577-6.

KOVÁČ, S., 2012. *Suitability analysis of data mining tools and methods.* Brno, Dostupné také z: [http://is.muni.cz/thesis/.] Bakalářská práce. Masarykova Univerzita v Brně, Fakulta Informatiky, Centrum výpočetní techniky. Vedoucí práce RNDr. Jaroslav Bayer.

SKALSKÁ, H., 2010. *Data Mining a klasifikační modely*. 1. vyd. Hradec Králové: Gaudeamus. ISBN 978-80-7435-088-7.

ŠARMANOVÁ, J., 2002, Metody dolování znalostí z dat. **In:** CHLÁPEK, D. (ed). *Datakon 2002*., Brno: Vysoké učení technické, s.165-184. ISBN 80-210-2958-7.

ŠIMŮNEK, M., 2010. *Systém LISp-Miner: akademický systém pro dobývání znalostí z databází : historie vývoje a popis ovládání.* 1. vyd. Praha : Oeconomica. ISBN 978-80-245-1699-8.

ŠTOCHL, J., 2003. Data mining v databázi katetrizací **In**: SVÁTEK, V. (ed). *Znalosti* 2003. Ostrava: Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky. s. 192–201. ISBN 80-248-0229-5.

WU, X. et al., 2008. [online] Top 10 algorithms in data mining. *Knowledge and Information Systems.* London: Imperial College Press, vol. 14(1), s. 1–37. ISSN: 0219-3116. Dostupný také z databáze ProQuest. DOI 10.1007/s10115-007-0114-2.

Seznam příloh

1. Nastavení úloh
2. Mezi-výběry pravidel
3. Prahové hodnoty kategorizace
4. Osnova řízeného rozhovoru