Trendy v časových oknech jako rizikové faktory kardiovaskulárního onemocnění

Lenka Nováková[1], Filip Karel[1], Petr Aubrecht[1], Marie Tomečková[2], Jan Rauch[2,3], Olga Štěpánková[1]

[1]Katedra kybernetiky, FEL, ČVUT, Technická 2, Praha 6
novakova@labe.felk.cvut.cz
[2]EuroMISE centrum, Ústav informatiky AV ČR, v.v.i.,
Pod Vodárenskou věží 2, 182 07, Praha 8
tomeckova@euromise.cz
[3]Vysoká škola ekonomická v Praze,
nám. W. Churchilla 3, 130 67, Praha 3
rauch@vse.cz

Abstrakt Článek se zabývá daty získanými v rámci longitunální studie STULONG a hledá způsob, jak charakterizovat časový průběh fyziologických parametrů a jeho vliv na to, zda se klinicky manifestuje kardiovaskulární onemocnění. K tomuto účelu navrhuje nový způsob definice odpovídajících odvozených atributů pomocí kombinace metody časových oken a diskretizace trendů. Takto získané odvozené atributy jsou použity pro identifikaci zajímavých podskupin, v nichž se výrazně liší frekvence zdravých a nemocných pacientů ve srovnání s celou skupinou pacientů. Jako nástroje pro hledání podskupin jsou použity LispMiner a heuristický nástroj pro hledání ordinálních asociačních pravidel, jejichž výsledky jsou porovnány.

Klíčová slova: Časové řady, Asociační pravidla

1 Popis dat longitunální studie STULONG

Studie (STULONG - LONGitudinal STUdy) byla realizována na II. interní klinice, 1. lékařské fakulty UK a Všeobecné fakultní nemocnice, U nemocnice 2, Praha 2 - pod vedením prof. MUDr. F.Boudíka, DrSc., MUDr. M.Tomečkové, CSc. a doc. MUDr. J.Bultase, CSc. Většina dat byla převedena do elektronické podoby v rámci evropského projektu Managing Uncertainity in Medicine programu Copernicus na pracovišti EuroMISE (Evropského centra medicínské informatiky, statistiky a epidemiologie) Karlovy univerzity a Akademie věd (pod vedením prof. RNDr. J.Zvárová, DrSc.). Analýza dat vznikla za podpory grantu MŠMT ČR LN 00B 107.

Jedná se o data z rozsáhlé epidemiologické studie primární prevence aterosklerózy, nazvané Národní preventivní multifaktoriální studie srdečních infarktů a cévních mozkových příhod. Studie zahrnuje dvacetileté pozorování přibližně 1400 mužů středního věku. Cílem projektu je identifikovat rizikové faktory aterosklerózy.

Data jsou rozdělena do 4 tabulek. Tento článek zpracovává data obsažená ve dvou z těchto tabulek - jedná se o tabulku Entry se vstupními daty o sledovaných pacientech a o tabulku Control tabulka se sérií kontrolních vyšetření. Po sloučení tvoří tyto dvě tabulky časovou řadu vyšetření o jednotlivých pacientech.

Sledování pacientů probíhalo téměř pravidelně každý rok a bylo ukončeno buď po 15 - 20 letech, nebo v okamžiku, kdy bylo u pacienta diagnostikováno některé ze sledovaných kardiovaskulárních onemocnění. Díky tomuto postupu by měli pacienti s mnoha měřeními být spíš pacienti zdraví a naopak ti pacienti, pro které máme k dispozici pouze malý počet měření, jsou právě ti, kteří onemocněli některou sledovanou chorobou. Práce [2] dokládá, že pro data STULONG skutečně platí, že čím déle je pacient sledován, tím menší je pravděpodobnost, že onemocní některou ze sledovaných chorob. V okamžiku, kdy pacient vstupuje do studie a jeho sledování začíná, nemůžeme vědět, jak dlouho bude sledován, a proto ani neznáme počet měření, která absolvuje. Celkový počet měření je odvozená charakteristika studovaných dat, kterou je nutné chápat jako anachronický atribut [1] a [5]. Platí pro něj podobně jako pro řadu jiných anachronických atributů,že jeho hodnota má výrazný vliv na případnou predikci toho, zda pacient onemocní či nikoliv.

Cílem analýzy dat z této studie je nalézt rozdíly v časovém vývoji sledovaných fyziologických veličin mezi dvěma skupinami sledovaných osob: skupinou pacientů, kteří v průběhu sledování onemocněli kardiovaskulární nemocí - atribut CVD=1, a skupinou těch, kteří zůstali zdraví - CVD=0. Základní statistickou analýzu rozdílů mezi oběma skupinami lze nalézt v [4].

Kdybychom uměli dobře popsat chování časové posloupnosti měření pomocí nějaké číslené charakteristicky, bylo by možné tyto hodnoty následně vyhodnotit pomocí asociačních pravidel. Ovšem zatím se nám nepodařilo nalézt žádnou dostatečně vypovídající charakteristiku, která by k tomu účelu v případě dat STULONG mohla sloužit. Při zpracování časových řad se často pracuje místo se všemi prvky řady pouze s několika charakteristickými odvozenými hodnotami, jakými jsou například aritmetický průměr, směrodatná odchylka, případně náhrada regresní křivkou, ať již lineární nebo vyššího řádu. Každá z těchto veličin je lineárně nebo nelineárně závislá na anachronickém atributu "počet měření" a lze tedy tvrdit, že má rovněž anachronický charakter, což ji činí nevhodnou pro zamýšlený účel. Při přípravě dat musíme použít takové předzpracování dat, které by časovou řadu nějakým způsobem charakterizovalo, ale zároveň odstranilo vliv počtu měření na tuto charakteristiku.

1.1 Příprava dat pomocí oken

Jednou z metod, jak odstranit vliv počtu měření, je použití časových oken [2]. Tato metoda transformuje časovou řadu o n měřeních na novou sadu časových

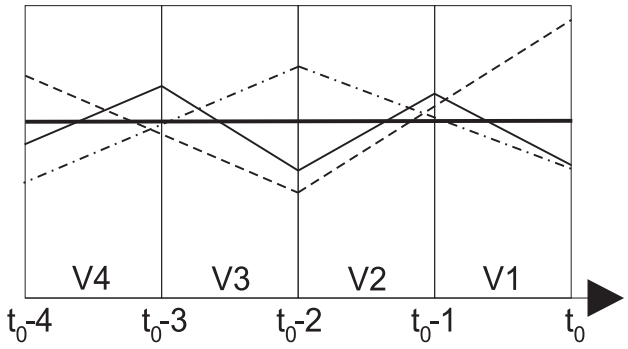
řad o konstantním počtu měření l. Predikovaný výsledek již není závislý na počtu měření, protože počet měření je konstantní.

Nejdříve je třeba stanovit délku okna. Nelze dát univerzální předpis, jak postupovat, neboť délku je nutné volit s ohledem na řešenou úlohu. Okno by nemělo být kratší než je minimální časové období, během něhož může dojít ke změnám hodnot měřených veličin, které mohou výrazně ovlivnit výsledek. Pro naši úlohu volíme délku okna 5, tj. vezmeme 5 po sobě následujících měření pacienta. Pro pacienty, kteří onemocněli, volíme 5 posledních měření před propuknutím nemoci, tj. budeme hledat změny v období před onemocněním v porovnání se skupinou zdravých pacientů. Jako reprezentanty zdravých pacientů jsme vybrali ty pacienty, kteří byli sledováni po dobu minimálně 15 let, tedy máme pro ně v databázi alespoň 15 měření, a kteří během celého sledování neonemocněli. Při výběru okna pro tyto zdravé pacienty existují dva protichůdné požadavky. Prvním požadavkem je, abychom vybrali takové okno, pro které platí, že pacient zůstal ještě po jeho uzavření nějakou dobu určitě zdráv. Protože nemůžeme zaručit, že zmíněná osoba neonemocněla krátce po skončení studie, není vhodné uvažovat poslední měření. Druhým požadavkem je vyloučení vlivu věku - skupiny zdravých a nemocných pacientů by měly mít podobnou věkovou strukturu. U pacientů, kteří onemocněli, bereme posledních 5 měření, tj. záznamy, kde dosáhli nejvyššího věku. Pokud bychom téhož efektu chtěli dosáhnout i pro skupinu pacientů, kteří neonemocněli, měli bychom pracovat s posledním měřením. Tyto dva protichůdné požadavky lze nejlépe splnit výběrem středového okna (okna uprostřed celého intervalu v průběhu kterého probíhal sběr zpracovávaných dat). Vybrali jsme proto okno mezi 8. a 12. měřením.

Tímto způsobem jsme transformovali původně naměřené časové řady do nové sady. Získali jsme 156 časových oken délky přibližně 5 let, která reprezentují pacienty s diagnózou CVD v závěrečném okamžiku časového okna. Dále pak 269 časových oken téže délky pro pacienty zdravé. Tato okna už mají všechna stejný počet měření a čas odpovídající všem těmto oknům je zhruba stejný. Data tohoto typu lze charakterizovat pomocí trendů. Nejjednodušším, i když poněkud hrubým řešením, se zdá být charakterizovat vývoj libovolné veličiny v okně pomocí parametrů její linearizace. První předběžné výsledky pro tento postup, byly zveřejněny v [4]. Problém tohoto přístupu je to, že úplně stejně reprezentuje 2 výrazně rozdílné vývojové trendy - viz obr. 1.

V tomto příspěvku se pokusíme o poněkud podrobnější analýzu vývojových trendů sledovaných atributů, která bude pracovat s kvalitativní informací o jednotlivých intervalech tvořících celé sledované okno. Pro data STULONG jsme zvolili délku okna 5, tj. pro každého pacienta máme 4 po sobě jdoucí intervaly. Zajímáme se pouze o vývoj sledovaných fyziologických veličin, tj. nebudeme se starat o absolutní hodnotu studované veličiny, ale pouze o směrnici jejího vývoje. Tuto směrnici budeme reprezentovat její kvalitativní hodnotou z množiny {"klesá", "stagnuje", "roste"}, která v upraveném souboru dat bude zaznamenána pomocí celých čísel {0,1,2}.

Pro každou sledovanou veličinu V nechť máme 5 po sobě jdoucích hodnot $V(t_0), V(t_0 - 1), \ldots, V(t_0 - 4)$, které obsahují údaje o hodnotách této veličiny



Obrázek 1. Znázornění posloupnosti měření s vyznačením časového okna

(atributu) v posloupnosti měření. Jde vlastně o 5 po sobě následujících členů posloupnosti měření: člen číslo $t_0, t_0 - 1, \ldots, t_0 - 4$. Z těchto údajů můžeme soudit, jak se veličina V vyvíjela v průběhu celého časového okna. Jako podklad nám bude sloužit rozdíl sousedních hodnot $V(t_0 - i)$ -$V(t_0 - i - 1)$ a odpovídající kvalitativní údaj o této hodnotě budeme značit $V(i + 1)$. Pro veličinu V tedy dostáváme tyto 4 atributy $V4$, $V3$, $V2$ a $V1$ takové, že údaj $V4$ charakterizuje vývojový trend na začátku okna, zatímco $V1$ svědčí o vývojovém trendu těsně před koncem časového okna. Zbývá ještě popsat, za jakých podmínek budeme vývojový trend nějaké veličiny považovat za stagnaci.

Vzhledem k tomu, že každé měření dat může být zatíženo nějakou chybou, nemá smysl považovat za „stagnující trend“ v daném intervalu jen ty případy, kdy hodnoty veličiny na začátku i konci tohoto intervalu nezaznamenaly žádnou změnu a jsou zcela totožné. Pojem „stagnace“ zobecníme tak, aby reflektoval „nevýznamné“ změny. Přesněji, pokud je absolutní hodnota rozdílu mezi koncovou a počáteční (výchozí) hodnotou veličiny menší než předem stanovená mez (pásmo stagnace ε dané veličiny), charakterizujeme i odpovídající vývoj jako "stagnaci". Naopak, je-li tento rozdíl větší než ε, říkáme, že veličina "roste", je-li menší než $-\varepsilon$, pak "klesá".

Zavedeme ještě další odvozenou veličinu, která bude poskytovat souhrnnou informaci o vývoji veličiny V v uvažovaném oknu délky 4, totiž slovo $w4_V$, které vznikne zřetězením hodnot $V4$, $V3$, $V2$ a $V1$. Obdobně zavedeme ještě i další podobnou odvozenou veličinu $w3_V$, jejíž hodnota je dána zřetězením hodnot $V3$, $V2$ a $V1$. Toto řešení představuje abstrakci, která rozliší rozdílné vývojové situace z obr. 1 a přitom není zbytečně podrobná.

Na doporučení lékaře jsme se zaměřili na sledování systolického a diastolického tlaku (budeme značit $Syst$ a $Diast$) a jejich rozdílu, tj. systolický - diastolický tlak (budeme značit SD), dále na sledování cholesterolu ($Chlstmg$), triglyceridů ($Triglmg$) a hmotnosti ($Hmot$). Protože nás zajímá pouze směr trendu, je jedno zda použijeme hmotnost nebo body mass index. Zvolená pásma stagnace jsou uvedena v tabulce 1.

Tabulka 1. Zvolená pásma stagnace při diskretizaci trendů

veličina	**pásmo stagnace**
tlak systolický, diastolický a jejich rozdíl	±5 mmHg
hmotnost	±2 kg
cholesterol	±20 mg%
triglyceridy	±20 mg%

Výsledkem tohoto předzpracování dat jsou dvě nové tabulky dat s 425 řádky představujícími jednotlivé osoby, první tabulka T_1 má $6 * 4 = 24$ sloupců, které popisují diskrétní trendy vývoje měřených veličin za 5-leté období. Druhá tabulka T_2 obsahuje jen $2 * 6 = 12$ sloupců, z nichž každý obsahuje slova z abecedy $\{0, 1, 2\}$ tak, jak odpovídají odvozeným atributům $w4_V$ a $w3_V$. V následujících odstavcích se budeme snažit hledat pomocí asociačních pravidel takové podskupiny diskrétních trendů, pro které se významně liší frekvence zdravých a nemocných pacientů ve srovnání s celou skupinou pacientů.

2 Předběžná analýza dat

Prvním krokem, jak ověřit vypovídající schopnost atributů navržených v předchozím odstavci, bude hledání asociačních pravidel, v jejichž konsekventu se bude vyskytovat buď atribut CVD=1 nebo CVD=0. Mělo by tedy jít o pravidla, která upozorňují na ty odvozené atributy, které mají vliv na to, zda člověk onemocní (CVD=1) nebo neonemocní (CVD=0).

Pro předběžnou analýzu jsme použili rychlý heuristický algoritmus pro hledání ordinálních asociačních pravidel [3] a hledali jsme pravidla s délkou podmínky 1 nebo 2. Pro charakteristiku ($minconf = 0,7; minsupp = 0,05; minlift = 1,1$) se podařilo najít 7 pravidel s podmínkou délky 1 a 26 pravidel s podmínkou délky 2. Ve výchozí skupině pacientů pro 156 platí CVD=1 a pro zbylých 269 naopak platí CVD=0. Frekvence výskytu CVD v celé skupině je tedy $0,367$. U libovolného pravidla vyděluje jeho podmínka, z původní skupiny 425 všech sledovaných pacientů, podskupinu těch pacientů, kteří uvedenou podmínku splňují. Uveďme pro zajímavost několik pravidel, u nichž je takto vzniklá skupina dostatečně veliká. Data, která máme k dispozici, nejsou dostatečně rozsáhlá na to, aby v nich bylo možné hledat komplexní a přesná pravidla. Ovšem, i slabší pravidlo může upozornit na zajímavé jevy ve studovaných datech, pokud se pravděpodobnost výskytu CVD v podskupině pacientů odpovídající podmínce tohoto pravidla výrazně změní ve srovnání s původní skupinou.

Například poměrně velkou podskupinu pacientů, přesněji 130 pacientů, vyděluje podmínka $SD4 = 1 \wedge Hmot3 = 1$, kterou lze formulavat slovy: „Pacienti, pro které hodnota rozdílu systolického a diastolického tlaku v intervalu před 4 lety byla konstantní a pro které také hmotnost v intervalu před 3 lety byla konstantní.“ Pro tuto podskupinu platí, že frekvence výskytu CVD v této pod-

skupině je jen 21,5%. To znamená, že výskyt CVD v této podskupině poklesl o 15,2% oproti frekvenci 36,7% u všech pacientů ve studovaných datech.

V následující tabulce 2 uvedeme ještě několik nalezených pravidel, jejichž podmínku splňuje dostatečný počet pacientů a při tom v odpovídající podskupině dochází k výrazně změně frekvence výskytu CVD. Za velmi zajímavou podmínku můžeme považovat i podmínku $Diast2 > 0 \wedge Hmot2 = 2$, která vyčleňuje podskupinu pacientů, u kterých platí: „V intervalu před 4 lety jejich diastolický tlak neklesal a hmotnost v intervalu před 2 lety rostla.“. Tato skupina je relativně malá - obsahuje jen 53 pacientů, ovšem změna frekvence výskytu CVD na této skupině je značná. Frekvence CVD=1 je na této skupině 0,60, což reprezentuje nárůst 23,7% oproti původní celé skupině pacientů.

Tabulka 2. Nalezené hodnoty sledovaných veličin, pro které je významný rozdíl mezi CVD=1 a CVD=0 mezi skupinou pacientů, kteří splňují danou podmínku a celkovou skupinou pacientů.

Popis veličin	**Rozdíl mezi podskupinou a celou množinou pacientů**
$SD4 = 1$	+8,4%
$Hmot4 = 1$	+8,3%
$Syst2 < 2 \wedge SD4 = 1$	+13,2%
$SD4 = 1 \wedge Hmot3 = 1$	+15,2%
$Hmot4 = 1 \wedge Hmot1 = 1$	+10,8%
$Hmot2 = 1 \wedge Chlstmg1 = 1$	+14,4%
$Diast2 > 0 \wedge Hmot2 = 2$	-23,7%

Odvozené atributy $w4_V$ a $w3_V$ navržené v předchozím odstavci kvalitativně charakterizují průběhy jednotlivých měřených veličin ve studovaném intervalu. Jedná se o atributy, které lze chápat jako výčtové s definičním oborem mohutnosti $3 * 3 * 3 * 3$, resp. jen $3 * 3 * 3$. První pokusy potvrzují, že tyto atributy přinášejí zajímavé informace. Povšimněme si například, že v podskupině pacientů, u nichž byla hmotnost v průběhu posledních 5 let stabilizovaná, se riziko CVD snížilo o 27,4% oproti skupině, kde hmotnost stabilizována nebyla, viz tabulka 3. Hodnota rozdílu je tedy počítána jinak než v předchozí tabulce, kde se bral v potaz rozdíl dané skupiny oproti celkové skupině pacientů. Zde se počítá rozdíl mezi skupinou splňující danou podmínku a doplňkovou skupinou, která danou podmínku nesplňuje.

Předběžné výsledky potvrzují, že navržené odvozené atributy stojí za bližší zkoumání, a proto jsme předpřipravená data podrobili důkladné analýze popsané v následujícím odstavci.

Tabulka 3. Nalezené trendy, pro které je významný rozdíl mezi CVD=1 a CVD=0 mezi skupinou pacientů, kteří splňují danou podmínku a skupinou, kteří danou podmínku nesplňují.

Popis vývoje trendu	**Rozdíl mezi disjunktními podskupinami**
$w4_{HMOT} = 1111$	-27,4%
$w4_{DIAST} = 1111$	-11,7%
$w4_{CHLSTMG} = 1111$	-8,3%
$w3_{SD} = 211$	-16%
$w3_{SYST} = 111$	-14,3%
$w3_{TRIGLMG} = 202$	-11,8%
$w3_{HMOT} = 111$	-19,8%
$w3_{CHLSTMG} = 111$	-11,7%
$w3_{DIAST} = 211$	-8,3%
$w4_{DIATS} = 1111 \wedge w4_{HMOT} = 1111$	-26,8%
$w3_{DIAST} = 111 \wedge w3_{HMOT} = 111$	-21,3%
$w3_{CHLSTMG} = 111 \wedge w3_{HMOT} = 111$	-8%

3 Aplikace procedury SD4ft-Miner

3.1 O proceduře SD4ft-Miner

Jedním z použitých přístupů k analýze trendů byla aplikace procedury SD4ft-Miner [6, 7]. Jedná se o jednu z GUHA procedur implementovaných v systému LISp-Miner, viz http://lispminer.vse.cz. Procedura je určena k řešení analytických otázek typu *Které dvě podmnožiny analyzovaných dat a za jaké podmínky se výrazně liší co se týče vztahu dvou booleovských atributů?* Formálně řečeno, procedura pracuje se vztahy tvaru

$$\alpha \bowtie \beta : \ \varphi \approx \psi \ / \ \gamma \ \ .$$

Tento vztah znamená, že podmnožiny analyzovaných dat dané booleovskými atributy α a β se liší co se týče vztahu booleovských attributů φ a ψ pokud je splněna podmínka daná booleobským attributem γ.

Procedura pracuje s maticemi dat s kategoriálními atributy, příklad takové matice je matice dat $\mathcal{M}$ v obr. 2.

Každý řádek matice odpovídá jednomu pozorovanému objektu, každý sloupec odpovídá jednomu atributu. Každý atribut má konečně mnoho možných hodnot které se nazývají *kategorie*. Booleovské atributy použité v SD4ft vztazích jsou odvozeny z literálů. Literál je jeden z následujících výrazů

- *základní booleovský atribut* $A(\alpha)$
- negace $\neg A(\alpha)$ základního booleovského atributu.

objekty t.j.	sloupce matice $\mathcal{M}$			(t.j. atributy)	příklady literálů	
řádky $\mathcal{M}$	A_1	A_2	...	A_{80}	$A_1(1,2)$	$\neg A_{80}(6)$
o_1	1	3	...	2	T	T
o_2	3	3	...	6	F	F
o_3	2	6	...	7	T	T
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
o_n	4	1	...	36	F	T

Obrázek 2. Matice dat $\mathcal{M}$

Základní booleovský atribut je výraz $A(\alpha)$ kde $\alpha \subset \{a_1, \dots a_k\}$ a $\{a_1, \dots a_k\}$ je množina všech kategorií atributu (sloupce) A. Základní booleovský atribut $A(\alpha)$ je pravdivý v řádku o matice $\mathcal{M}$ jestliže $a \in \alpha$ kde a je hodnota atributu A v řádku o. V obr. 2 jsou dva příklady literálů, $A_1(1,2)$ a $\neg A_{80}(6)$. Množina α se nazývá *koeficient literálů* $A(\alpha)$ a $\neg A(\alpha)$.

GUHA procedury systému LISp-Miner jsou vyhodnocovány pomocí různých kontingenčních tabulek. Procedura SD4ft-Miner pracuje se čtyřpolními tabulkami dvou booleovských atrtibutů, tyto tabulky se nazývají 4ft-tabulkami. *4ft-tabulka booleovských atributů* φ a ψ na matici dat $\mathcal{M}$ se značí $4ft(\varphi, \psi, \mathcal{M})$. Je to čtveřice $\langle a, b, c, d\rangle$ přirozených čísel takových, že a je počet řádků matice $\mathcal{M}$ splňujících jak φ tak i ψ, b je počet řádků splňujících φ a nesplňujících ψ, c je počet řádků nesplňujících φ a splňujících ψ, a d je počet řádků nesplňujících ani φ ani ψ.

SD4ft vztah $\alpha \bowtie \beta : \ \varphi \approx \psi \ / \ \gamma$ se v matici dat $\mathcal{M}$ vyhodnocuje pomocí dvou kontingenčních tabulek. První z nich je 4ft tabulka $4ft(\varphi, \psi, \mathcal{M}/(\alpha \wedge \gamma))$ atributů φ a ψ na matici $\mathcal{M}/(\alpha \wedge \gamma)$ viz obr. 3. Matice $\mathcal{M}/(\alpha \wedge \gamma)$ je podmaticí matice $\mathcal{M}$ která obsahuje právě všechny řádky matice $\mathcal{M}$ splňující $\alpha \wedge \gamma$. Matice $\mathcal{M}/(\alpha \wedge \gamma)$ je tedy matice, která obsahuje právě všechny řádky z podmnožiny dané atributem α a splňující podmínku γ. Tedy $a_{\alpha\wedge\gamma}$ je počet řádků matice $\mathcal{M}/(\alpha \wedge \gamma)$ splňujících jak φ tak i ψ, atd. Analogicky, $4ft(\varphi, \psi, \mathcal{M}/(\beta \wedge \gamma))$ je 4ft tabulka φ a ψ na $\mathcal{M}/(\beta \wedge \gamma)$, viz obr. 3.

$\mathcal{M}/(\alpha \wedge \gamma)$	ψ	$\neg\psi$
φ	$a_{\alpha\wedge\gamma}$	$b_{\alpha\wedge\gamma}$
$\neg\varphi$	$c_{\alpha\wedge\gamma}$	$d_{\alpha\wedge\gamma}$

$4ft(\varphi, \psi, \mathcal{M}/(\alpha \wedge \gamma))$

$\mathcal{M}/(\beta \wedge \gamma)$	ψ	$\neg\psi$
φ	$a_{\beta\wedge\gamma}$	$b_{\beta\wedge\gamma}$
$\neg\varphi$	$c_{\beta\wedge\gamma}$	$d_{\beta\wedge\gamma}$

$4ft(\varphi, \psi, \mathcal{M}/(\beta \wedge \gamma))$

Obrázek 3. 4ft tabulky $4ft(\varphi, \psi, \mathcal{M}/(\alpha \wedge \gamma))$ a $4ft(\varphi, \psi, \mathcal{M}/(\beta \wedge \gamma))$

Symbol $\approx$ se nazývá *SD4ft-kvantifikátor*. Každému SD4ft kvantifikátoru odpovídá podmínka týkající se dvojice tabulek $4ft(\varphi, \psi, \mathcal{M}/(\alpha \wedge \gamma))$ a $4ft(\varphi, \psi, \mathcal{M}/(\beta \wedge \gamma))$. Každý SD4ft-kvantifikátor je konjunkcí několika základních SD4ft-kvantifikátorů. Příkladem SD4ft-kvantifiátoru je podmínka

$$a_{\alpha\wedge\gamma} \geq BASE \quad \wedge \quad a_{\beta\wedge\gamma} \geq BASE \quad \wedge \quad |\frac{a_{\alpha\wedge\gamma}}{a_{\alpha\wedge\gamma}+b_{\alpha\wedge\gamma}} - \frac{a_{\beta\wedge\gamma}}{a_{\beta\wedge\gamma}+b_{\beta\wedge\gamma}}| \geq p \ ,$$

která je konjunkcí základních kvantifikátorů $a_{\alpha\wedge\gamma} \geq BASE$, $a_{\beta\wedge\gamma} \geq BASE$ a $|\frac{a_{\alpha\wedge\gamma}}{a_{\alpha\wedge\gamma}+b_{\alpha\wedge\gamma}} - \frac{a_{\beta\wedge\gamma}}{a_{\beta\wedge\gamma}+b_{\beta\wedge\gamma}}| \geq p$. Tento SD4ft kvantifikátor říká, že v množině dané α je alespoň $BASE$ objektů splňujících γ, v množině dané β je také alespoň $BASE$ objektů splňujících γ a že absolutní hodnota rozdílu konfidencí asociačního pravidla $\varphi \approx \psi$ na maticích $\mathcal{M}/(\alpha \wedge \gamma)$ a $\mathcal{M}/(\beta \wedge \gamma)$ je alespoň p.

Vztah $\alpha \bowtie \beta : \ \varphi \approx \psi \ / \ \gamma$ je pravdivý v matici dat $\mathcal{M}$ právě když je podmínka daná SD4ft-kvantifikátorem splněna pro dvojici čtyřpolních tabulek $4ft(\varphi, \psi, \mathcal{M}/(\alpha \wedge \gamma))$ a $4ft(\varphi, \psi, \mathcal{M}/(\beta \wedge \gamma))$, viz obr. 3.

Vstupem procedury SD4ft-Miner je matice dat a relativně jednoduché zadání velmi rozsáhlé množiny SD4ft vztahů. Je k dispozici řada nástrojů na upřesnění zadané množiny, jejich podrobný popis však přesahuje rozsah tohoto článku. Procedura vygeneruje všechny takto zadané SD4ft vztahy a jejím výstupem jsou všechny SD4ft vztahy pravdivé v zadané matici dat.

3.2 Porovnání skupin pacientů mezi sebou

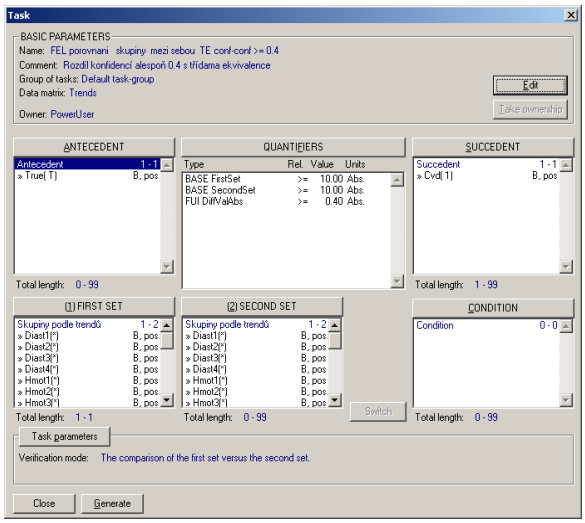
Proceduru SD4ft-Miner jsme použili pro řešení otázky: „Které dvě skupiny pacientů se od sebe hodně liší co se týká podílu pacientů s CVD?“ Skupiny pacientů jsme určovali jako množiny α (pro první skupinu) a β (pro druhou skupinu). Podíl pacientů s CVD jsme určovali jako konfidenci asociačního pravidla $\mathcal{T} \approx CVD(1)$ kde $\mathcal{T}$ je booleovský atribut pravdivý pro všechny pacienty a $CVD(1)$ je booleovský atribut který říká, že pacient má CVD. Podmínka γ nebyla použita.

Zadání obou množin α i β bylo stejné. Byly použity atributy $Diast1 \ldots Diast4$, $Hmot1 \ldots Hmot4$, $SD1 \ldots SD4$, $Chlstmg1 \ldots Chlstmg4$, $Syst1 \ldots Syst4$, $Triglmg1 \ldots Triglmg4$. Každý z těchto atributů má 3 možné hodnoty: 0, 1, 2. Pro atribut $Diast1$ byly automaticky generovány booleovské atributy $Diast1(0)$, $Diast1(0,1)$, $Diast1(1,2)$, $Diast1(2)$; analogicky i pro ostatní atributy. Dále byly automaticky použity samostatné booleovské atributy a konjunkce dvou booleovských atributů. Nebyl použity konjunkce z atributů patřících do jedné skupiny (např. $Diast1 - Diast4$).

Jako SD4ft-kvantifiátor byla použita podmínka

$$a_{\alpha\wedge\gamma} \geq 10 \quad \wedge \quad a_{\beta\wedge\gamma} \geq 10 \quad \wedge \quad |\frac{a_{\alpha\wedge\gamma}}{a_{\alpha\wedge\gamma}+b_{\alpha\wedge\gamma}} - \frac{a_{\beta\wedge\gamma}}{a_{\beta\wedge\gamma}+b_{\beta\wedge\gamma}}| \geq 0.4 \ ,$$

která říká, že v každé skupině je alespoň 10 pacientů s CVD a že relativní četnosti pacientů s CVD v obou skupinách se liší alespoň o 0.4. Zadání procedury SD4ft-Miner je ukázáno v obr. 4.



Obrázek 4. Zadání procedury SD4ft-Miner pro porovnání skupin pacientů mezi sebou

Při následném běhu procedury bylo za necelých 7 minut (PC s 1.66 GHz a 2 GB RAM) vygenerováno a verifikováno 374 tisíc SD4ft vztahů, 67 z nich bylo pravdivých a tvořilo výstup. Největší rozdíl byl mezi skupinou pacientů daných atributem $Hmot4(1,2)$ a skupinou pacientů daných atributem $Diast4(0) \wedge Hmot4(0)$, viz tab. 4.

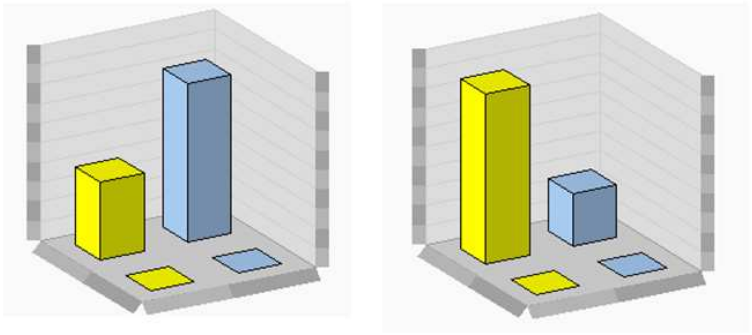
Tabulka 4. Počty pacientů ve skupinách $Hmot4(1,2)$ a $Diast4(0) \wedge Hmot4(0)$

skupina	pacientů s CVD	pacientů bez CVD	relativní četnost CVD
$Hmot4(1,2)$	119	242	0.33
$Diast4(0) \wedge Hmot4(0)$	13	4	0.76

Vidíme, že relativní četnost pacientů ve druhé skupině je o 0.43 větší než v první skupině. Grafické porovnání poměru pacientů s CVD a bez CVD v obou skupinách je v obr. 5.

3.3 Porovnání skupin pacientů s celým souborem pacientů

Proceduru SD4ft-Miner jsme použili i pro řešení otázky *Která dílčí skupina pacientů se hodně liší od skupiny všech pacientů co se týče podílu pacientů s CVD?* Dílčí skupinu pacientů jsme určovali stejně jako první skupinu v předchozí úloze, další parametry byly stejné. Ukázalo se, že nejvíce se liší skupina $Diast4(0) \wedge Hmot4(0)$ která byla nalezena i v předchozí úloze. Jako druhá nejvíce



Skupiny $Hmot4(1,2)$ a $Diast4(0) \wedge Hmot4(0)$

Obrázek 5. Porovnání poměru pacientů s CVD a bez CVD

se lišící skupina od skupiny všech pacientů byla nalezena skupina $Diast4(0) \wedge Chlstmg4(0)$. Jejich odlišnost je zachycena v tab. 5.

Tabulka 5. Počty všech pacientů a pacientů ve skupině $Diast4(0) \wedge Chlstmg4(0)$

skupina	pacientů s CVD	pacientů bez CVD	relativní četnost CVD
všichni pacienti	156	269	0.37
$Diast4(0) \wedge Chlstmg4(0)$	15	5	0.75

Vidíme, že relativní četnost pacientů s CVD mezi pacienty ve skupině $Diast4(0) \wedge Chlstmg4(0)$ je o 0.38 větší než mezi všemi pacienty.

4 Závěr

Hodnocení rizika vzniku CVD je nutno pojímat komplexně. Hraniční hodnoty rizikových faktorů (RF) byly stanoveny na základě rozboru velkých epidemiologických studií zejména s ohledem na úmrtnost na CVD, např. pro krevní tlak jsou jako rizikové určeny hodnoty vyšší než 140 mm Hg systolického a/nebo 90 mm Hg diastolického krevního tlaku. Nezáleží jen na absolutní hodnotě daného RF. U dané osoby se často vyskytuje více RF a ty se navzájem rovněž ovlivňují buď pozitivně nebo negativně. Např. obezita zhoršuje hodnoty krevního tlaku a naopak pro udržení doporučené hodnoty krevního tlaku je vhodné mj. udržet optimální hmotnost, a v případě nadváhy či obezity hmotnost snížit.

Výsledky analýzy jsou zajímavé v tom, že ukazují na některé zajímavé vztahy mezi trendem vývoje RF. Z rozboru výsledků uvedených v tabulce 2 např. vyplývá, že udržení hmotnosti v průběhu 5 let bez ohledu na trend dalších RF a

bez ohledu na absolutní hodnotu hmotnosti nebo BMI samo o sobě snižuje riziko CVD o 27,4%. Pokud po uvedenou dobu 5 let stagnovaly i hodnoty diastolického krevního tlaku, k dalšímu snížení rizika vzniku CVD již nedošlo. Udržení hmotnosti a hodnot diastolického krevního tlaku pouze po dobu čtyř let sice riziko vzniku CVD rovněž významně sníží, toto snížení (o 21,3 %) je však menší než při stagnaci obou hodnot po dobu pěti let.

Poděkování: Tato práce vznikla díky podpoře grantu GAČR 201/05/0325, grantu 1ET101210513 (Relační strojové učení pro průzkum biomedicínských dat) a grantu 1ET200300413 (Informační technologie pro rozvoj kontinuální sdílené péče o zdraví).

Reference

1. Mařík V., Štěpánková O., Lažanský J. a kol.: *Umělá inteligence 4.* Academia, Praha, s. 355-407, 2003.
2. Nováková L., Kléma J., Štěpánková O.:*Anachronické atributy a dobývání znalostí* Znalosti 2004, Brno, s.202-209, 2004.
3. Karel F., Kléma J.:*Dolování ordinálních asociačních pravidel* Znalosti 2005, Ostrava, s.226-233, 2005.
4. Nováková L., Kléma J., Jakob M., Štěpánková O., Rawles S.:*Trend analysis and risk identification* Workshop Proceedings and Tutorial notes ECML/PKDD 2003 [CD-ROM]. Stuttgart: IRB Verlag, 2003, s. 95-107.
5. Pyle D.: *Data Preparation For Data Mining.* Morgan Kaufmann, California, 1999.
6. Rauch J., Šimůnek M: GUHA Method and Granular Computing. In: Hu, X et al (ed.). Proceedings of IEEE conference Granular Computing. 2005, pp. 630–635.
7. Rauch J., Šimůnek M.: Dealing with Background Knowledge in the SEWEBAR Project. In Berendt, B. et al (ed.) Proceedings of the ECML/PKDD 2007 workshop Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery, Warsaw, Poland. 97 – 108
8. Projekt STULONG, WWW page, http://euromise.vse.cz/stulong.
9. SumatraTT, WWW homepage, http://screwdriver.felk.cvut.cz/sumatra/

Annotation:

Trends in time windows as risk factors of cardiovascular disease

This paper analyzes the STULONG (LONGitudinal STUdy) data and searches for methods how to represent time development of physiological attributes and their influence on development of cardiovascular diseases (CVD). There is suggested a new approach to definition of a derived attribute which provides information on time development of a single physiological attribute obtained through combination of windowing and discretization. The resulting derived attributes are further analyzed with intention to discover well described interesting subgroups of observed persons. A subgroup is denoted as interesting if it exhibits significantly different relation between number of subjects who remain healthy and those who develop CVD compared to the set of all patients. Two different tools for association rule mining (namely LispMiner and a new heuristic tool) are utilized for the search of interesting subgroups and their results are compared.