DOLOVÁNÍ SILNÝCH VZORŮ Z LÉKAŘSKÝCH SEKVENČNÍCH DAT

J. Kléma*, T. Holas*, F. Železný* and F. Karel*

* Gerstnerova laboratoř, katedra kybernetiky,
České vysoké učení technické, Technická 2,
166 27 Praha, Česká republika

{klema,zelezny}@labe.felk.cvut.cz, {holast1,karelf1}@fel.cvut.cz

Abstract:

Sekvenční data jsou důležitým zdrojem lékařských znalostí. Tato specifická data mohou vznikat řadou různých způsobů. V tomto článku na příkladu konkrétní studie prezentujeme obecné postupy pro jejich dolování. Jde o preventivní dlouhodobou studii aterosklerózy – data jsou výsledkem dvě dekády trvajícího sledování vývoje rizikových faktorů a přidružených jevů. Hlavním cílem je identifikovat časté sekvenční vzory, tj. opakující se časové jevy, a studovat jejich možnou souvislost s objevením jedné ze sledovaných kardiovaskulárních nemocí. Z širší škály dostupných metod se soustředíme na induktivní logické programování, které potenciální vzory vyjadřuje ve formě rysů v predikátové logice prvního řádu. Rysy jsou nejprve automaticky extrahovány a následně sdružovány do pravidel, která představují výstupní formu získané znalosti. Navržený postup je porovnán s tradičnějšími metodami publikovanými dříve. Jde o metodu posuvných oken a epizodní pravidla.

Sequential data represent an important source of automatically mined and potentially new medical knowledge. They can originate in various ways. Within the presented domain they come from a longitudinal preventive study of atherosclerosis – the data consist of series of long-term observations recording the development of risk factors and associated conditions. The intention is to identify frequent sequential patterns having any relation to an onset of any of the observed cardiovascular diseases. This paper focuses on application of inductive logic programming. The prospective patterns are based on first-order features automatically extracted from the sequential data. The features are further grouped in order to reach final complex patterns expressed as rules. The presented approach is also compared with the approaches published earlier (windowing, episode rules).

Úvod

Lékařské databáze obsahují velké množství informací o pacientech a jejich klinických vyšetřeních. Komplexní vztahy a vzory skrytě obsažené v těchto datech mohou přinést dosud neznámé znalosti. Využitelnost těchto znalostí byla prokázána celou řadou úspěšných medicínských aplikací metod dolování dat. Hlavní těžiště těchto aplikací spočívalo ve využití atributových metod učení (attribute-valued learning, AVL). Tyto metody jsou však omezeny na data, ve kterých je každý objekt, v našem případě pacient, popsán pevnou množinou předem daných vlastností, tj. atributů. Někdy je tato podmínka splněna přímo, jindy postačí doménově nezávislá metoda předzpracování dat (např. selekce atributů). Stále však existuje velké množství úloh, u kterých je převod na AVL reprezentaci netriviální a specifický. Ad-hoc převod je pak časově náročný, vyžaduje současně účast lékaře a informatika, často s nejistým praktickým výsledkem. Je proto vhodné aplikovat techniky, které mohou pracovat se sekvenčně-relačními daty přímo.

Tento článek diskutuje a vytěžuje časově-sekvenční data, která obvykle vyžadují komplexní předzpracování. Pod pojmem sekvence chápeme časovou posloupnost událostí. Každá událost má přiřazený typ a je popsána hodnotou spolu s časovou známkou. Celá databáze pak může obsahovat jednu nebo v obecném případě více nezávislých sekvencí. Protože platí, že více nezávislých sekvencí jednoho typu je možné převést na sekvenci jedinou, není toto dělení úplně podstatné. Konečným cílem je nalezení silných vzorů, za které považujeme často se opakující charakteristické řetězce událostí (podsekvence) a posouzení jejich možného vztahu s cílovou událostí. Typickou cílovou událostí v lékařské aplikaci je projev nemoci nebo prokazatelná změna zdravotního stavu pacienta.

Konkrétně se zaměříme na data STULONG [1], dlouhodobou 20-ti letou preventivní studii mužů středního věku. Studie obsahuje data, jež jsou výsledkem sledování přibližně 1400 mužů. Hlavním záměrem projektu bylo odhalit rizikové faktory

atherosklerózy. Data jsou ze své podstaty multirelační, skládají se ze 4 základních relací. V čase se vyvíjející údaje jsou uloženy v tabulce vyšetření (Control), která zaznamenává u jednotlivých osob různě dlouhé série vyšetření. Každé vyšetření pak shrnuje konstantní soubor hodnot rizikových veličin spolu s nimi souvisejícími doplnkovými údaji. Příkladem rizikových faktorů jsou veličiny BMI (Body Mass Index), krevní tlak nebo biochemická vyšetření (cholesterol, triglyceridy), doplnkovými údaji jsou například fyzická aktivita v zaměstnání a její změny, způsob dopravy do práce, užívání léků apod. Data o jednom muži odpovídají z logického hlediska jedné sekvenci událostí různého typu. Někteří muži byli sledováni po dobu několika let (několik málo vyšetření), u jiných máme k dispozici až 20 vyšetření – délka sekvencí se tedy může lišit velmi výrazně. V neposlední řadě jsou zaznamenána a časově označena pozorování spojená s projevy kardiovaskulárních nemocí nebo přímo jejich diagnóza.

Vědecký cíl výše zmíněné studie může být formulován v jazyce sekvenčního dolování dat následujícím způsobem. Cílem je identifikovat časté sekvenční vzory mající prokazatelnou souvislost s objevením některého ze sledovaných kardiovaskulárních onemocnění (KO). Pro zjednodušení uvedeme možné příklady vzorů v kontextu celých pravidel, vše v přirozeném jazyce: (1) jestliže BMI v čase klesá a poté znovu roste zatímco krevní tlak stále roste pak je jakékoli KO pravděpodobnější, (2) jestliže BMI v čase roste a hladina HDL cholesterolu je nízká pak je jakékoli KO pravděpodobnější.

Studie STULONG byla jednou z úloh hromadně řešených v rámci konference ECML/PKDD zaměřené na problémy strojového učení a dolování dat. Na dané téma byla publikována velká řada příspěvků, z nichž se ovšem pouze několik zabývalo sekvenčním dolováním dat. [3] představuje předzpracování časových dat metodou posuvného okna, tj. ad-hoc metodou vytvářející trendové atributy pomocí agregačních oken. [5] doluje epizodní pravidla univerzálním nástrojem WinMiner. Vedle problémové nezávislosti lze metodu charakterizovat tím, že automaticky vyhledává optimální velikost časového okna v jehož rozsahu se vyhledávají opakující se sekvence událostí.

V tomto textu prezentujeme alternativní přístup vhodný pro multirelační problémy a aplikovatelný i pro sekvenční data strukturovaná jako ve STULONG studii – induktivní logické programování (ILP). Představíme obecný nástroj RSD [4] pro relační vytváření rysů identifikujících významné podskupiny v datech a aplikujeme jej na uvedenou doménu. Hledané vzory budou vyjádřeny formou rysů predikátové logiky prvního řádu a budou automaticky extrahovány ze sekvenčních dat. Užitečnost těchto rysů bude vyhodnocena AVL učením, rysy budou použity k vytváření komplexních struktur (formulí) popisujících konečné vzory.

Hlavní přínos článku spočívá v prohloubení studia využití metod ILP při dolování sekvenčních dat, novém způsobu aplikace RSD a obecném porovnání výsledků s alterantivními metodami publikovanými dříve. Článek porovnává dosažené výsledky z hlediska jednoduchosti, srozumitelnosti a znovupoužitelnosti v podobných úlohách.

RSD: Relational Subgroup Discovery

Relační učení pravidel je typicky používáno při řešení klasifikačních a predikčních úloh. Předchozí výzkum STULONG dat však prokázal [3], že nalezené vzory (a zcela zřejmě ani ty potenciálně skryté) nejsou dostatečné ke spolehlivé klasifikaci sledovaných mužů do tříd. Jejich apriorní dělení na zdravé a nemocné (KO), popřípadě do jemnějších kategorií podle typu nemoci, není na základě nahromaděných dat možné. Je evidentní, že úloha by měla být formulována jako identifikace zajímavých podskupin (subgroup discovery). Na vstupu je populace objektů (individuí, mužů středního věku) popsaných hodnotou jejich cílové vlastností (KO) spolu s hodnotami veličin, které je dále charakterizují. Výstupem jsou podmnožiny dané populace, které jsou statisticky "nejzajímavější": lze je jednoznačně charakterizovat, jsou co největší (obsahují co nejvíce objektů) a mají co nejméně vyvážené rozdělení vzhledem k cílové vlastnosti. Jejich definice přitom vychází ze sekvenčních vzorů odrážejících časový vývoj rizikových faktorů a přidružených veličin.

RSD [4] umožnuje přechod mezi relačním pravidlovým učením a identifikací zajímavých podskupin. Nástroj je založen na těchto principech: úplná konstrukce rysů prvního řádu, eliminace irelevantních rysů, implementace relačního pravidlového učení, aplikace algoritmu váženého pokrytí a heuristické využití vah objektů pro stanovení relativní přesnosti algoritmu.

Proces učení může být zjednodušen do následujících kroků. Nejprve konstruujeme samotné rysy, tj. konjunkce literálů dostupných v rámci dané domény. Jejich klíčovou vlastností je schopnost definovat podskupiny charakterizované o dva odstavce výše. Poté jsou rysy sdružovány do pravidel, jejichž kritická vlastnost je velmi podobná. Dodatečným požadavkem je dostatečné pokrytí, tj. pravidlo by mělo být splněno co největším počtem dosud nepokrytých objektů (detaily lze nalézt v [4, 6]).

Dolování STULONG dat

Proveditelnost, složitost, rozlišení

Jako nejpřirozenější přístup k dolování STULONG dat se jeví vyhledávání libovolných sekvenčních rysů, potažmo vzorů. Jeden rys by tak mohl pokrývat sekvenci libovolné délky a současně by mohl

být *vícetypový*, tj. sdružovat události odlišných typů (v případě STULONG dat různé rizikové faktory apod.). Dva příklady takových sekvencí/rysů jsou uvedeny na Obrázku 1. Časové vztahy jsou modelovány binárními predikáty $after_1$, $after_2$, ..., $after_n$ – predikáty vyjadřují, že druhá událost nastala 1, 2 nebo n vyšetření po události první – a *simultaneous* – který postihuje současné události z jednoho vyšetření. Tyto predikáty by mohly být dále doplněny o řadu zobecnění predikátu $after$, např. druhá událost nastala v libovolném vyšetřením následujícím vyšetření, v němž se objevila událost první. Podotkněme, že přestože vyšetření nejsou v čase zcela pravidelná, v rámci zjednodušení je v tomto textu považujeme za každoroční.

Abychom minimalizovali fázi předzpracování dat, spojité veličiny mohou být diskretizovány pomocnými predikáty (např. weight_cat(X, small) :- X < 64.). Tento přístup navíc přináší větší variabilitu definice událostí, protože za událost může být teoreticky považována přímo hodnota veličiny ($weight(checkup_i, 71)$) nebo kategorie ($weight(checkup_i, X)$, $weight\_cat(X, xsmall)$). Zjednodušená textová reprezentace rysu může být následující:

```
feature(ID,PAT):-checkup(PAT,Time1),
checkup(PAT,Time2), after1(Time1,Time2),
syst(Time1,V1), syst_cat(V1,low),
syst(Time2,V2), syst_cat(V2,high).
```

Rys je splněn pro všechny objekty/muže mající dvě přímo po sobě následující vyšetření, v nichž se hodnota systolického krevního tlaku mění z kategorie "low" do kategorie "high". Rys postihuje dvě události, každá z těchto událostí je popsána třemi predikáty (definujících pacienta/čas, typ události a kategorii). Obě události jsou navíc spojeny časovým predikátem.

Výše naznačená variabilita kandidátských rysů a sekvencí je jistě žádoucí z hlediska teoretického rozsahu konečné znalosti. Nicméně, naznačená variabilita výrazně zvětšuje stavový prostor sekvencí a ohrožuje praktickou proveditelnost jeho prohledávání. Počet kandidátských sekvencí může být příliš velký a znemožnit tak vytvoření konečných pravidel v únosném čase. Předpokládejme, že pracujeme s a veličinami, z nichž každá může nabývat v různých hodnot, maximální délka sekvence necht je l. Celkový počet jednotypových sekvencí pak je $O(n_s) = a\ v^{l+1}$, zatímco počet vícetypových sekvencí může být až $O(n_i) = (av)^{l+1}$. Je zřejmé, že počet sekvencí roste exponenciálně s jejich maximální délkou. Výpočet je ještě složitější pokud uvažujeme rysy. Výše uvedený příklad demonstroval, že délka rysu násobně překračuje délku sekvence, protože každá událost je reprezentována několika predikáty a události musí být vzájemně časově svázány. Technická výpočetní složitost přitom opět roste exponenciálně s maximální povolenou délkou rysu (udanou v predikátech). V jistém smyslu tedy mohutnost prohledávaného prostoru rysů překračuje mohutnost původního prostoru sekvencí, protože nelze automaticky rozlišit mezi smysluplnými rysy a těmi, které nedopovídají žádné existující sekvenci [1].

Z výše uvedeného plyne nutnost omezit délku rysů a tím i sekvencí, současně je také vhodné volit rozumný počet veličin i jejich hodnot. Vícetypové rysy jsou výpočetně velmi náročné. Odhadněme počet kandidátských sekvencí v doméně STULONG. Počet vyšetření kolísá mezi 1 a 21, přibližně u 80% mužů bylo provedeno 5 nebo více vyšetření – z tohoto důvodu se zdá rozumným omezením maximální délka sekvence 5 událostí. Nejsignifikantnějších rizikových faktorů je 5 (systolický a diastolický tlak (SYST, DIAST), hladina cholesterolu v mg%(CHLSTMG), hladina triglyceridů v mg%(TRIGLMG) a BMI), byly u nich zjištěny desítky různých hodnot. Tomu odpovídají desítky miliard kandidátských sekvencí.

V důsledku toho je třeba redukovat počet veličin (připomenme, že celkově jich jsou desítky, i když různé důležitosti), což ovlivnuje informovanost o vztahu mezi veličinami (lze uvažovat opakované běhy s různými množinami veličin). Zkrácení sekvencí omezuje rozlišení v časové oblasti. Snížení počtu hodnot jednotlivých veličin naopak redukuje rozlišení v oblasti datové. Vícetypová povaha sekvencí tak může být viděna spíše jako překážka než žádoucí vlastnost řešení. Současně ovšem platí, že přes výraznou výpočetní náročnost jde o nový způsob zpracování sekvenčních dat (viz. [2]). V souvislosti s ním jsou vyvíjeny nové a efektivnější algoritmy pro vytváření vícetypových pravidel. V našem textu je konečné řešení popsáno v následujících dvou sekcích. Představuje ekvilibrium mezi mohutností prohledávaných stavových prostorů, rozsahem jazyka vzorů a složitostí předzpracování spojeného s vytvářením apriorní znalosti o úloze. Řešení je založeno na myšlence rozdělení dat do tří disjunktních časových oken.

Předzpracování dat

Způsob předzpracování dat může výrazně ovlivnit efektivitu použití RSD. Nutným syntaktickým krokem je rutinní převod dat z obvyklých relačních tabulek do predikátové formy. Současně je třeba doplnit jazykové deklarace, opět v predikátové logice. Pro tento účel byl vytvořen externí konverzní program v jazyce Java. Vstupem jsou data ve formátu CSV, výstupem pak soubory .pl (data – objekty a jejich cílová vlastnost, časové údaje o vyšetřeních a hodnoty sledovaných veličin pro jednotlivá vyšetření)

[1] RSD v žádném případě negeneruje libovolné rysy, tj. libovolné konjunkce literálů. Prostor rysů je automaticky redukován tím, že každá proměnná musí být alespon jednou definována jako vstupní, rysy nesmí být rozložitelné, predikáty mohou být definované jako antisymetrické apod. Složitost výpočtu je také přímo ovlivnitelná apriorní znalostí, která může formulovat predikáty vysoké úrovně omezující prostor rysů.

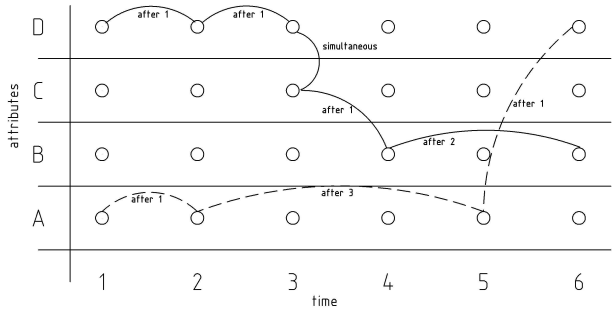

Figure 1: Vícetypové sekvence v Prologu

a .b (apriorní znalost – časové predikáty, definice časové sekvence a výčet možných elementů, z nichž se budou skládat rysy) [6].

Z logického hlediska lze předzpracování dat rozdělit do tří základních kroků: (1) převod do predikátové formy, využitelné jako kód jazyka Prolog, (2) případná diskretizace veličin, (3) konstrukce nových trendových (tj. časových) veličin. První krok byl v zásadě zmíněn v prvním odstavci, další dva již přímo ovlivnují efektivitu běhu RSD. Věnujme se nejprve diskretizaci. V předchozí sekci bylo předvedeno, jak lze diskretizaci provést přímo v predikátové logice. Jedná se pravděpodobně o nejmetodičtější a logicky elegantní řešení, které ovšem současně snižuje efektivitu generování rysů tím, že prodlužuje jejich nutnou délku. Z tohoto důvodu je vhodné veličiny diskretizovat předem (výše zmíněný Java kód). V případě dat STULONG byly generovány veličiny: NORMBMI, NORMSYST (NORMDIAST), NORMCHLSTMG a NORMTRIGLMG. Vznikly diskretizací původních veličin BMI, SYST, DIAST, CHLSTMG a TRIGLMG. Transformace byla provedena ekvidistantní diskretizací do tří intervalů označených jako "low", "medium" a "high"[2].

Konstrukce rysů může být dále zjednodušena předpřipravením krátkodobých trendových veličin. Veličiny TRENDBMI, TRENDSYST (TRENDDIAST), TRENDCHLSTMG, TRENDTRIGLMG transformují originální data do formy vyjadřující rychlost změn klíčových rizikových faktorů v čase. Možné hodnoty "trendových" veličin jsou "down2", "down", "flat", "up", a "up2", znamenající "prudký pokles", "pokles", "beze změny", "nárůst", a "prudký nárůst" příslušných originálních veličin mezi dvěma sousedními vyšetřeními. Zjednodušení rysů je zřejmé. Rys platící pro každého pacienta, jenž má dvě následná měření systolického krevního tlaku se změnou z kategorie "low" do kategorie "high", vyjádřený jiným způsobem v předchozí sekci, bude nyní vypadat takto:

[2]Způsob diskretizace je vhodné volit po dohodě s expertem. V dané úloze připadá v úvahu i větší počet kategorií, popř. jiná diskretizační metoda – frekvenční diskretizace do kategorií o stejném počtu objektů nebo lokální metody zohledňující každého z mužů odděleně).

```
feature(ID,PAC):-checkup(PAC,Time1),
trendsyst(Time1,big_increase).
```

Délka rysu poklesla ze 7 na 2. Kompaktnější základní predikáty umožní pracovat s delšími sekvencemi událostí a všeobecně vyšším datovým i časovým rozlišením při stejné mohutnosti prohledávacího prostoru. Cenou je složitější předzpracování a nutnost apriori rozhodnout o struktuře použitých základních stavebních predikátů.

Na závěr se zmíníme o cílové veličině KO. Studie je koncipována tak, že kardiovaskulární onemocnění se může objevit pouze v posledním vyšetření. Tj. po jeho diagnóze je muž z preventivní studie vyřazen a převeden do léčebného režimu. Proto lze také obecně předpokládat, že důležitost vyšetření v čase roste. Cílová veličina KO je binárním atributem vyjadřujícím u daného muže výskyt či absenci kardiovaskulárního onemocnění na konci jeho řady vyšetření (0 – zdráv, 1 – nemocen).

Konečné parametry experimentu

Předzpracování dat navržené v předchozí sekci výrazně redukuje délku rysu při zachování komplexity a rozlišení modelovaných sekvencí. K dokončení návrhu experimentu je třeba definovat vztah rizikových faktorů a cílové KO veličiny. Délka originálních sekvencí se pohybuje od 1 do 21, průměrná délka je 8. Jednotlivé homogenní sekvence (SYST, BMI atd.) byly rozděleny do tří disjunktních oken nazvaných *begin, middle, end*. *End* okno zahrnuje vždy poslední 4 události, *middle* pokrývá 4 předchozí události a *begin* okno zahrnuje zbytek – všechny události od prvního vyšetření až k middle oknu. Každý rys postihuje události z jediného okna a reprzentuje sekvenci maximální délky 2. Časové predikáty $after_i$ uvedené v teoretickém úvodu byly nahrazeny binárními predikáty $after\_beg$, $after\_mid$ a $after\_end$. Ty definují, že druhá událost nastala v libovolném čase následujícím po čase vyšetření první události, přičemž musela nastat ve stejném okně (*beg* značí počáteční okno atd.).

Každé pravidlo může být tvořeno nejvýše 3mi rysy. Pravidlo, a tím i konečný vzor, tedy může celkově postihnout sekvenci o 6 událostech 3 různých typů. Konkrétní příklady nalezených pravidel lze nalézt v náledující sekci. Jak plyne z předchozího výkladu, počty událostí a typů se mohou lišit experimet od experimentu a závisí na způsobu formulace úlohy. V jiné doméně mohou být zcela jiné. Klíčem je pamětová a výpočetní realizovatelnost.

Výsledky

Tato sekce předkládá vybrané výsledky ve formě pravidel a jejich interpretací. Začněme následujícím pravidlem:

```
třída:0, spol:0.968, pok:0.156, zdvih:1.308
```

```
f(7369,A):-checkup(A,B), normsyst(B,medium),
trendbmi(B,flat), trendsyst(B,up).
f(3068,A):-checkup(A,B), checkup(A,C),
after_mid(C,B), trendbmi(C,flat).
f(1158,A):-checkup(A,B), checkup(A,C),
after_beg(C,B), normtriglmg(B,low),
trendtriglmg(C,up2).
```

Pravidla mají stejnou syntaxi jako klasická rozhodovací pravidla, tedy Podmínka $\Rightarrow$ Třída, kde Podmínka (premisa) má tvar "objekt současně splnuje všechny uvedené rysy" a Třída (výsledek) vyjadřuje "cílová veličina KO má pro objekt hodnotu". Rozdíl mezi identifikací zajímavých podskupin a klasifikací spočívá v tom, že pravidla nejsou použitá k disjunktnímu dělení objektů do tříd, ale k porozumění sledované doméně. Můžeme je chápat i jako asociační pravidla $Ant \Rightarrow Suc$, spojující dva jevy, antecedent a sukcedent. K hodnocení kvality pravidel používáme kvantifikátory (míry) známé právě z oblasti asociačních pravidel. Popis pravidla, uvedený vždy na první řádce, reprezentuje následující údaje. *Třída* 0 označuje pravidlo odkazující na muže bez KO, třída 1 naopak signalizuje muže s KO v posledním vyšetření. *Pokrytí (pok)* (někdy také označováno jako Podpora) proporcionálně vyjadřuje kolik objektů pravidlo splnuje, $pok = n(Ant)/n$, kde $n(Ant)$ je počet objektů splnujících podmínku, n je celkový počet objektů. Pravidla s malým pokrytím (např. 5% a méně, záleží ovšem i na celkovém počtu objektů) často nejsou brána v úvahu. To proto, že popisovaný vztah může být pouze náhodnou odchylkou ve sledovaném vzorku. *Spolehlivost (spol)* $spol = n(Ant \cap Suc)/n(Ant)$ je mírou důvěryhodnosti a přesnosti pravidla. Vyjadřuje proporcionálně kolik z objektů splnujících podmínku splnuje i závěr. *Zdvih* je definován $zdvih = spol/p_a$, kde $p_a = n(Suc)/n$ je apriorní pravděpodobnost třídy pravidla. Zdvih vyjadřuje kolikrát je dané pravidlo lepší nežli pravidlo náhodné, které zachovává apriorní rozdělení objektů do tříd. Všechny uvedené kvantifikátory jsou maximalizační. Požadujeme tedy pravidla, která adresují co nejvíce objektů. Současně platí, že se množina těchto objektů v rozdělení podle cílové vlastnosti co nejvíce liší od úplné množiny.

Zbývající řádky pravidla jsou výčtem rysů tvořících antecedent, tedy podmínku. Všechny uvedené rysy, v našem konkrétním příkladu 3, musí být splněny současně. První rys vyjadřuje, že daný muž měl vyšetření se středním systolickým tlakem, tento tlak mu ale vzrostl a současně nerostlo jeho BMI. Druhý rys popisuje objekt se dvěma vyšetřeními B a C ve středním okně dlouhodobého pozorování. Vyšetření C nastalo před vyšetřením B a muž při něm vykázal konstantní BMI. U tohoto rysu je zřejmé, že vyšetření B není důležité a slouží pouze k časovému určení vyšetření C. Povšimněme si, že námi definovaný jazyk kvůli omezení stavového prostoru úlohy neobsahuje predikát typu $belongs(C, mid)$, který by ukotvil vyšetření C ve středním časovém okně přímo. V konkrétním rysu pak vyšetření C nemůže být posledním vyšetřením střední části (musí být následováno B), což je však spíše důsledkem výše uvedeného jazykového omezení. Třetí rys říká, že objekt má dvě vyšetření B a C v úvodní části své sekvence. C předchází B. Nejprve tedy prudce rostou triglyceridy a následně je jejich úroven opět nízká. Pokud celý popis shrneme a interpretujeme zjednodušeně dojdeme k tomuto popisu. Náš muž měl ve vzdálené minulosti prudký nárůst triglyceridů následovaný jejich normalizací na nízké úrovni. Ve střední časti pozorování bylo jeho BMI stabilizováno. Kdykoli v jeho sledování pak došlo k tomu, že střední systolický tlak dále rostl beze změn BMI. Muž s touto charakteristikou má zhruba o 30% vyšší šanci[3], že neonemocní kardiovaskulární nemocí, než průměrný muž ze studie. Podívejme se na další pravidlo:

```
třída:1, spol:0.615, pok:0.049, zdvih:2.367
f(4380,A):-checkup(A,B), checkup(A,C),
after_end(C,B),normsyst(B,high),trendbmi(C,flat).
f(4124,A):-checkup(A,B),checkup(A,C),
after_end(C,B),normbmi(B,medium),trendchlstmg(C,up2).
f(4439,A):-checkup(A,B),checkup(A,C),
after_end(C,B),normsyst(B,high),trendchlstmg(C,up2).
```

Pravidlo má velmi dobrý zdvih, na druhou stranu nemá velké pokrytí. Jde tedy o silné pravidlo platící pro malý počet objektů. Všechny popsané události nastávají na konci sekvence vyšetření, dle definice oken nejdéle 3 vyšetření před případným objevením KO. Popsaní muži mají setrvalý stav BMI následovaný vysokou hodnotou systolického tlaku, prudce rostoucí hladinu cholesterolu následovanou střední hodnotou BMI a vysokým systolickým tlakem. Tito muži mají o 137% vyšší pravděpodobnost brzkého objevení kardiovaskulární nemoci než průměr ve studii. Pokud pravidlo porovnáme s obecně známými lékařskými znalostmi, je zřejmé, že je s nimi v souladu. Vysoký krevní tlak a rostoucí cholesterol jsou jevy přispívající k poruchám kardiovaskulárního systému.

Podívejme se podrobněji na podporu posledního pravidla z pohledu reálné velikosti popisované skupiny mužů. Pokrytí 0.049 implikuje při 800 objektech 39 mužů. Apriorní pravděpodobnost třídy 1 v datech je 26%. V náhodně vybrané skupině 39 mužů tedy nejpravděpodobněji bude 10 nemocných. Ve skupině definované pravidlem je 24 nemocných. Uvažujeme-li binomické pravděpodobnostní rozdělení, pravděpodobnost, že se v náhodně vybrané skupině 39 mužů objeví 24 a více nemocných, je pouze $2.6e^{-6}$. Tato pravděpodobnost není vysoká, musíme ale uvažovat i opakované pokusy. Při prohledávání stavového prostoru statisticky testujeme velké množství různých rysů a pravidel.

Relační učení může být využito i pro nesekvenční data. V tomoto případě je aplikace zjednodušena o časové predikáty či předzpracování trendových

[3] Určeno podle zdvihu, $p = (zdvih - 1) \cdot 100\%$.

veličin. Výsledkem aplikace pak je mj. následující pravidlo:

```
třída:0, spol:0.910, pok:0.084, zdvih:1.230
f(9745,A):-liquors(A,none).
f(9737,A):-beer(A,more_than_1_liter).
```

Pravidlo vyjadřuje, že pijáci piva, kteří současně nepijí likéry s velkým obsahem alkoholu, mají o 23% snížen výskyt KO. Porovnáme-li tuto třídu pravidel s pravidly vytvářenými statistickým či asociačním učením, dojdeme k závěru, že výsledky se výrazně neliší (stejné pravidlo již bylo nalezeno dříve). Výhodou induktivního relačního učení je však to, že sekvenční a nesekvenční rysy mohou být snadno a přirozeně kombinovány. Příklad kombinovaného pravidla je uveden zde:

```
třída:1 spol:0.568, pok:0.055, zdvih:2.185
f(9738,A):-beer(A,occasionally).
f(8453,A):-checkup(A,B),normchlstmg(B,medium),
trendchlstmg(B,flat).
f(3787,A):-checkup(A,B),checkup(A,C),after_mid(C,B),
trendtriglmg(B,down2),trendtriglmg(C,flat).
```

Pravidlo může být slovně popsáno takto. Občasní konzumenti piva s normální hladinou cholesterolu a prudkým poklesem triglyceridů v krvi mají o 118% vyšší šanci, že se u nich rozvine KO. Pokrytí pravidla opět není vysoké. Pokud spojíme znalost získanou posledními dvěma pravidly s příbuznou znalostí obecnou můžeme usoudit, že dobrou prevencí KO je nepít tvrdý alkohol a přestat kouřit (což je obecná znalost). Zajímavou informací hodnou podrobnějšího lékařského zvážení je, že pití piva není škodlivé ani ve větším množství, pokud současně neklesá hladina triglyceridů.

Table 1: Parametry nejsilnějších nalezených pravidel

Class	**Spolehlivost**	**Pokrytí**	**Zdvih**
0	0.9	0.32	1.22
0	0.95	0.2	1.28
0	0.97	0.16	1.31
0	0.90	0.15	1.22
0	0.91	0.08	1.23
0	0.97	0.13	1.31
0	0.95	0.05	1.29
0	1.0	0.07	1.35
1	0.45	0.17	1.73
1	0.47	0.13	1.81
1	0.47	0.1	1.8
1	0.57	0.06	2.19
1	0.62	0.05	2.37
1	0.7	0.03	2.68

V Tabulce 1 je uveden přehled nejsilnějších nalezených pravidel. Zaznamenány jsou pouze jejich kvalitativní charakteristiky, všechna zajímavá pravidla nelze z prostorových důvodů rozebírat podrobně. Tabulka však může být vodítkem k obecnému posouzení síly nalezených pravidel. Z obecného hlediska platí, že pravidla s větším pokrytím mají menší zdvih a naopak. Protože skupina zdravých mužů je větší (mužů bez KO jsou necelé tři čtvrtiny), pravidla na ni zaměřená mají větší pokrytí a menší zdvih. U nemocné skupiny je tomu právě naopak. Pokud pravidla porovnáme s obecnou lékařskou znalostí, zjistíme, že jsou s ní ve většině případů v souladu. Menšina pravidel je pak lékaři hodnocena jako zajímavá či překvapivá. Pouze několik pravidel bylo hodnoceno sporně.

Diskuse

Generovaná pravidla jsou dostatečně silným výrazovým prostředkem k detailnímu popisu časových vazeb mezi veličinami. Současně platí, že díky generalizaci nejsou náchylná k vyhledávání sekvencí odpovídajících náhodnému šumu. Drobné odchylky v hodnotách veličin nejsou považovány za trendy. Pro exaktní domény, kde i drobné změny mohou být významné (lze si představit například fyziku), by způsob předzpracování a apriorní znalost musely být opět přizpůsobeny charakteru dat. Z obecného hlediska je omezením umělé rozdělení časové osy do tří oken, což nemá jasné fyziologické opodstatnění. Hlavním důvodem je omezení složitosti prohledávání. Metoda je zaměřena na vyhledávání lokálních vzorů, resp. omezených podskupin. Pravidla nejsou určena ke konstrukci globálního modelu, což potvrzuje výsledek zkusmé klasifikace – klasifikační přesnost nepřekonává klasické algoritmy učení (rozhodovací stromy, bayesovské sitě apod.) nevyužívající trendových atributů, tj. sekvenční informaci.

Ačkoli sekvenční informace pro daná data nezpřesnuje globální model je zřejmé, že relační učení sekvenčních pravidel vyhledává zajímavé vzory. Tyto vzory by standardními metodami, jako jsou tradiční asociační pravidla, zůstaly opomenuty. Vzhledem k tomu, že sekvenční rysy můžeme libovolně kombinovat s rysy ne-sekvenčními (tj. volně směšovat okamžitá a časově proměnná data), jde o zobecnění tradičního asociačního učení. Jako u každé metody dolování dat jsou předzpracování, formulace apriorní znalosti i konečný výsledek problémově závislé. Nejde ale o ad-hoc postup, protože metoda definuje jasné komunikační rozhraní s uživatelem, jehož jazyk je dostatečně bohatým výrazovým prostředkem pro přizpůsobení se úloze.

Porovnejme popsanou relační metodu s jejími přímými sekvenčními alternativami aplikovanými dříve. Metoda pevných či posuvných oken je jednoduchým a často používaným postupem předzpracování sekvenčních dat. V případě pevných

oken sekvenci rozdělíme do několika disjunktních částí. Posuvné okno naopak generuje vzájemně se překrývající podsekvence. V obou případech jsou hodnoty veličin zachycených v okně převedeny na veličiny agregované a analyzovány tradičním atributových učením (AVL). V případě STULONG dat byla jako agregační funkce zvolena lineární regrese, klíčovou veličinou byl tedy lineární trend. Aplikace posuvného okna pevné délky je v [3]. Ačkoli metoda v dané úloze přinesla velmi dobré výsledky (napomohla mj. k objevení vztahu mezi počtem vyšetření a KO), projevila se současně její časová náročnost a problémová závislost. Otázky typu 'jaká je optimální délka okna?' nebo 'je linearizace vhodnou metodou generalizace při vytváření vzorů?' musí být zvažovány a experimentálně řešeny.

WinMiner [5] je naproti tomu zcela obecný nástroj pro vyhledávání *epizodních pravidel* – vzorů, které mohou být extrahovány z teoreticky libovolně dlouhé sekvence. Při jeho aplikaci je ovšem třeba řešit otázky velmi podobné otázkám řešeným v případě induktivního logického programování. Konkrétně, data musí být diskretizována, protože systém pracuje se symbolickými sekvencemi. Abychom mohli pracovat s vícetypovými vzory, konečná abeceda symbolů musí odlišit rizikové faktory, resp. události různých typů. Množina sekvencí odpovídajících jednotlivým mužům je převedena na sekvenci jedinou a to tak, aby časové známky přiřazené událostem jednoznačně oddělovaly jednotlivé muže. Tj. poslední událost předchozího muže má takovou časovou známku, která nikdy nedovolí ji zařadit do stejného okna s první událostí muže následujícího. Z důvodu výpočetní složitosti musí být poměrně výrazné omezena maximální délka okna, ve kterém vzory vyhledáváme. Opět platí, že složitost roste exponenciálně s maximální délkou okna. Sekvenční vzory nalezené pomocí RSD a WinMineru jsou významově podobné. Hlavní odlišnost spočívá v možnosti RSD předstanovit výsledný tvar vzoru pomocí apriorní znalosti. WinMiner vyhledává univerzální třídu vzorů, tj. libovolnou dostatečně často se opakující sekvenci symbolů. Přesnější definice třídy vyhledávaných vzorů přináší vedle specializace i snížení pamětové náročnosti a urychlení výpočtu. Na druhou stranu vyžaduje základní znalost predikátové logiky.

Závěr

Tento článek prezentuje induktivní logické programování jako nástroj dolování sekvenčních dat. Relační učení není prioritně použito k analýze dat rozptýlených ve více tabulkách, ale na individuálně orientovaná data. V těch jsou jednotlivé objekty popsány různým počtem transakcí charakterizovaných časovými údaji. Cílem je současné vyhledání vztahů mezi položkami (veličinami) a transakcemi. Praktické uplatnění metody je demonstrováno na příkladu identifikace rizikových faktorů atherosklerózy. Důraz je přitom kladen na efekt časových změn těch veličin, o jejichž absolutní hodnotě je známo, že ovlivnuje kardiovaskulární systém.

Poděkování

Tento výzkum vznikl v rámci programu MŠM 6840770012 "Transdisciplinary Biomedical Engineering Research II." podporovaného Ministerstvem školství a grantu 1ET101210513 "Relational Machine Learning for Analysis of Biomedical Data" podporovaného Českou akademií věd.

Studie STULONG 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 medicínksé informatiky, statistiky a epidemiologie) Karlovy univerzity a Akademie věd (pod vedením prof. RNDr. J.Zvárové, DrSc.). Analýza dat vznikla za podpory grantu MŠMT ČR LN 00B 107.

References

[1] EUROMISE, Stulong – epidemiological study of atherosclerosis. Internet site address: http://euromise.vse.cz/challenge2004/index.html.

[2] Guil, F., Bosch, A., and Marin, R. TSET: Algorithm for mining frequent temporal patterns. In *Proc. of ECML/PKDD'04 Workshop on Knowledge Discovery in Data Streams - A Collaborative Effort in Knowledge Discovery*, pages 65–74, 2004.

[3] Kléma, J., Nováková, L., Karel, F., and Štěpánková, O. Trend analysis in STULONG data. In *Proc. of ECML/PKDD'04 Discovery Challenge - A Collaborative Effort in Knowledge Discovery*. Prague: Univ. of Economics, 2004.

[4] Lavrac, N., Železný, F., and Flach, P. RSD: Relational subgroup discovery through first-order feature construction. In Matwin and Sammut, editors, *Proc. 12th Int. Conf. on Inductive Logic Programming*, 2002.

[5] Meger, N., Leschi, C., Lucas, N., and Rigotti, C. Mining episode rules in STULONG dataset. In *Proc. of ECML/PKDD'04 Discovery Challenge - A Collaborative Effort in Knowledge Discovery*. Prague: Univ. of Economics, 2004.

[6] Železný, F. RSD user's manual. Available at: http://labe.felk.cvut.cz/ zelezny/rsd/rsd.pdf.