Hledání zajímavých poskupin definovaných pomocí trendů v časových oknech

Lenka Nováková, Filip Karel, Petr Aubrecht,

Marie Tomečková, Jan Rauch a Olga Štěpánková

Osnova

- Motivace - shrnutí dosavadních analýz dat STULONG
- Spolupráce SumatraTT a LispMiner
- Příprava časových řad pomocí oken
- Výsledky realizovaných analýz

Stulong data

- longitudinální studie mužů
- cílem bylo zjistit výskyt rizikových faktorů kardiovaskulárních onemocnění
- sledováno bylo 1400 mužů po dobu 20 let
- web odkaz http://euromise.vse.cz/stulong

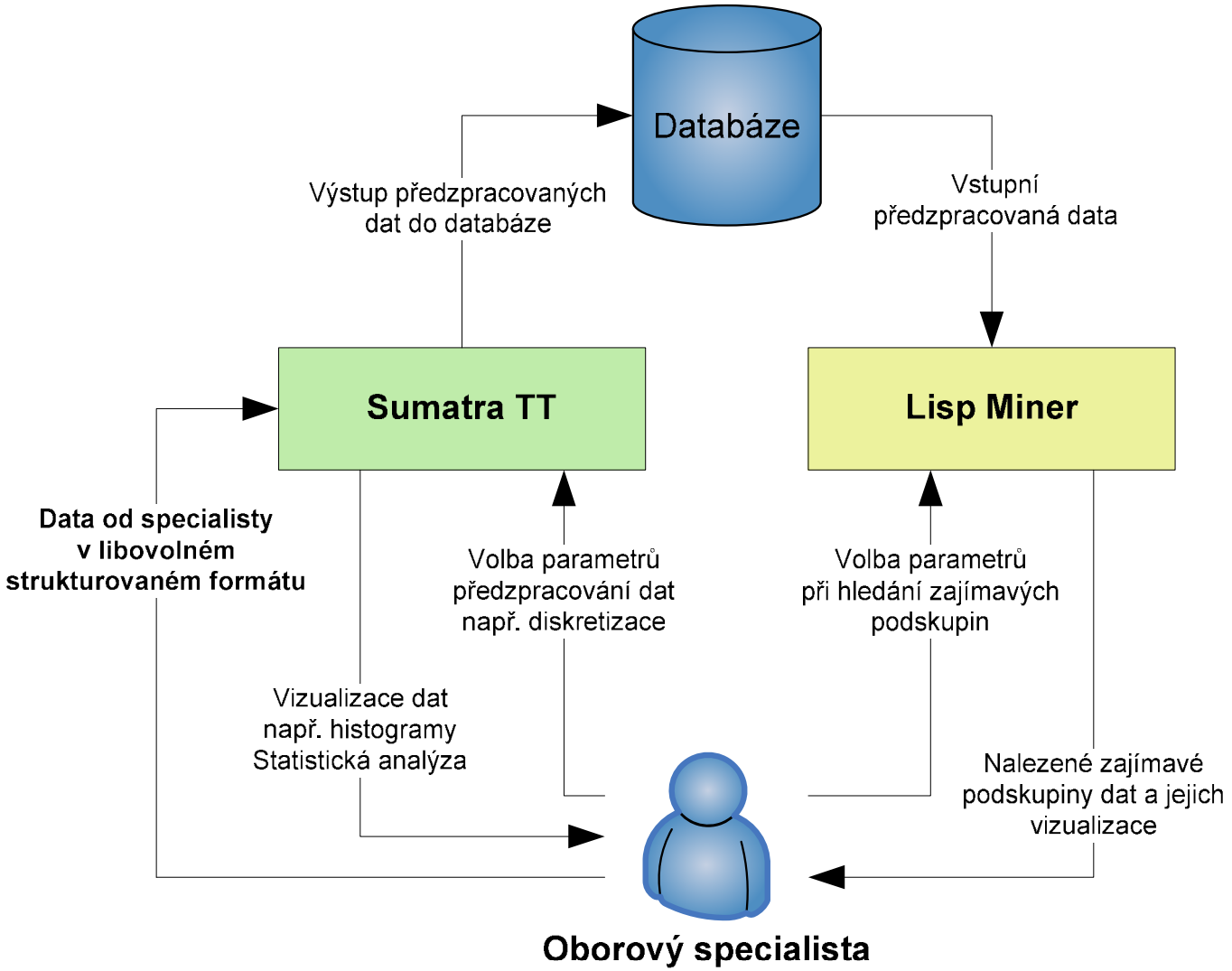
Shrnutí dosavadních analýz

- Základní statistické analýzy na základě jednotlivých atributů
  - ECML/PKDD data mining challenge - Novakova, Klema et al. - Trend analysis and risk identication, 2003
- Anachronické atributy, časová okna, analýza trendů
  - Znalosti 2004 - Nováková, Kléma, Štěpánková - Anachronické atributy a dobývání znalostí, 2004
  - Znalosti 2005 - Karel, Kléma – Ordinální asociační pravidla
- Shrnující článek hledání zajímavých vzorů chování v časových řadách
  - článek v časopisu IEEE Transactions on Systems, man and cybernetics - Kléma et al. - Sequential Data Mining: A Comparative Case Study in Development of Atherosclerosis Risk Factors, 2008

Vzory v časových oknech

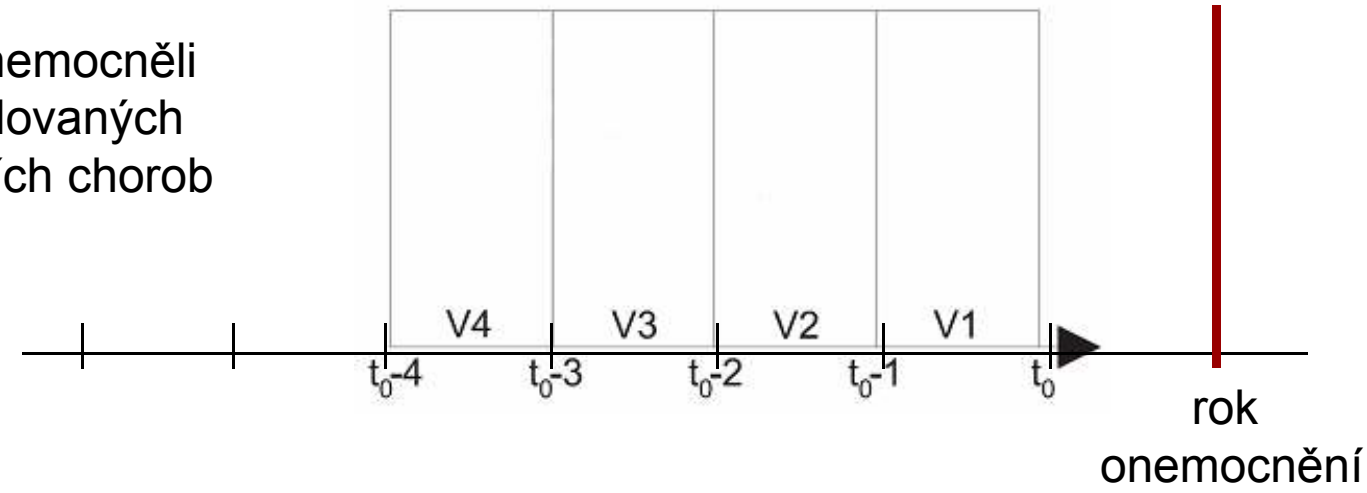
- běžný způsob popisu chování veličiny v čase pomocí trendů
  - zachycení veličiny jako celku, např. lineární regrese
  - nepopisuje chování detailně – např. klesá, roste - ??
- popis pomocí vzorů
  - využití diskrétního charakteru měření - popis trendu v období mezi po sobě jdoucími měřeními
  - nutný důraz na definici diskretizace – pásmo stagnace
- hledání asociačních pravidel
  - pomocí rozdílu podpory pro skupinu, která splňuje, a skupiny, která nesplňuje, zadanou podmínku

Spolupráce SumatraTT a LispMiner

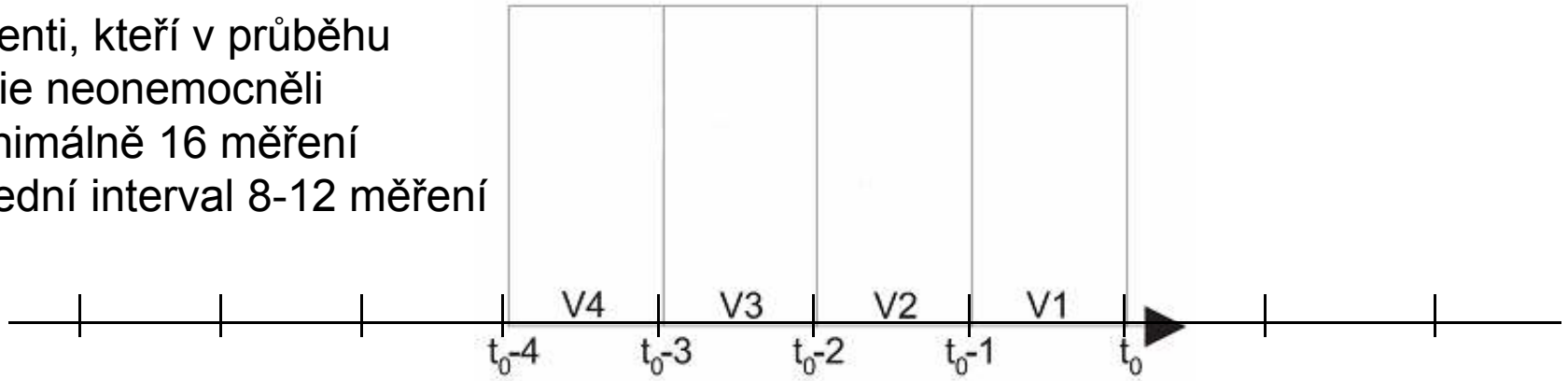


Příprava dat – volba časových oken

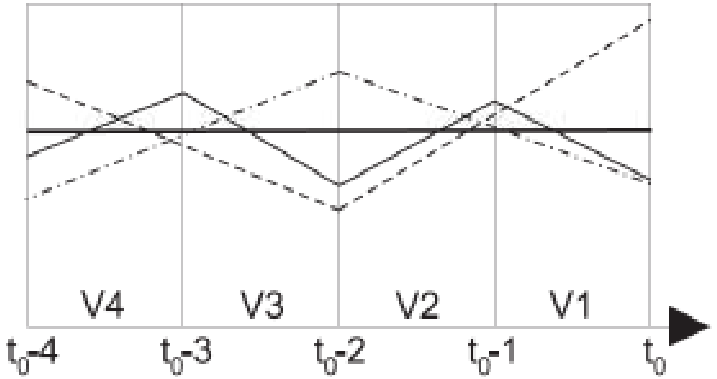
pacienti, kteří onemocněli některou ze sledovaných kardiovaskulárních chorob



pacienti, kteří v průběhu studie neonemocněli
- minimálně 16 měření
- střední interval 8-12 měření



Předzpracovaná data



- Příklad
  - **BMIStart** – startovní hodnota v čase -4
  - **BMI4** (v čase –4 až -3)
  - **BMI3** (v čase –3 až -2),
  - **BMI2** (v čase –2 až -1)
  - **BMI1** (v čase -1 až 0)
- Hodnoty
  - 0 – klesá
  - 1 – stagnuje
  - 2 – roste
- Pro každou veličinu
  - hodnoty pro jednotlivé časové úseky – **VStart, V4, V3, V2, V1**
  - **w4v** zřetězení hodnot V4,V3,V2,V1
  - **w3v** zřetězení hodnot V3,V2,V1

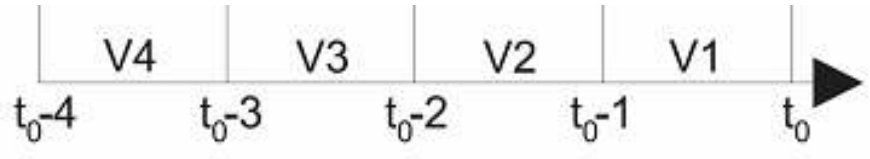
Předzpracovaná data

- Zvolené analyzované veličiny
  - systolický tlak, diastolický tlak, syst – diast tlak
  - hmotnost, body mass index
  - cholesterol, triglyceridy
- Data
  - 7 veličin * 7 hodnot = 49 hodnot pro každé okno
  - pacienti, kteří neonemocněli – 269 oken
  - pacienti, kteří onemocněli – 156 oken
  - celkem 425 oken * 49 atributů

  - frekvence CVD v datech je 36.7%

Výsledky ordinálních pravidel

- metoda popsána v článku Karel F. a Kléma J. – Ordinální asociační pravidla - Znalosti 2005
- použita na hledání pravidel s délkou podmínky 1 nebo 2
- minconf=0.7, minsupp=0.05, minlift=1.1
- Nalezená zajímavá pravidla
  - **SD4 = 1 & Hmot3 = 1** – pro toto pravidlo je frekvence CVD 21.5%, tj. o 15.2% menší – podmínku splňuje 130 pacientů
  - **w4hmot = 1111** – riziko CVD se snížilo o 27.4% - podmínku splňuje 72 pacientů



Lisp Miner

- Analytická otázka:
  - Které skupiny pacientů se od sebe hodně liší co se týče podílu pacientů s CVD?
  - Skupina_1(?) ✕ Skupina_2(?): CVD(%)

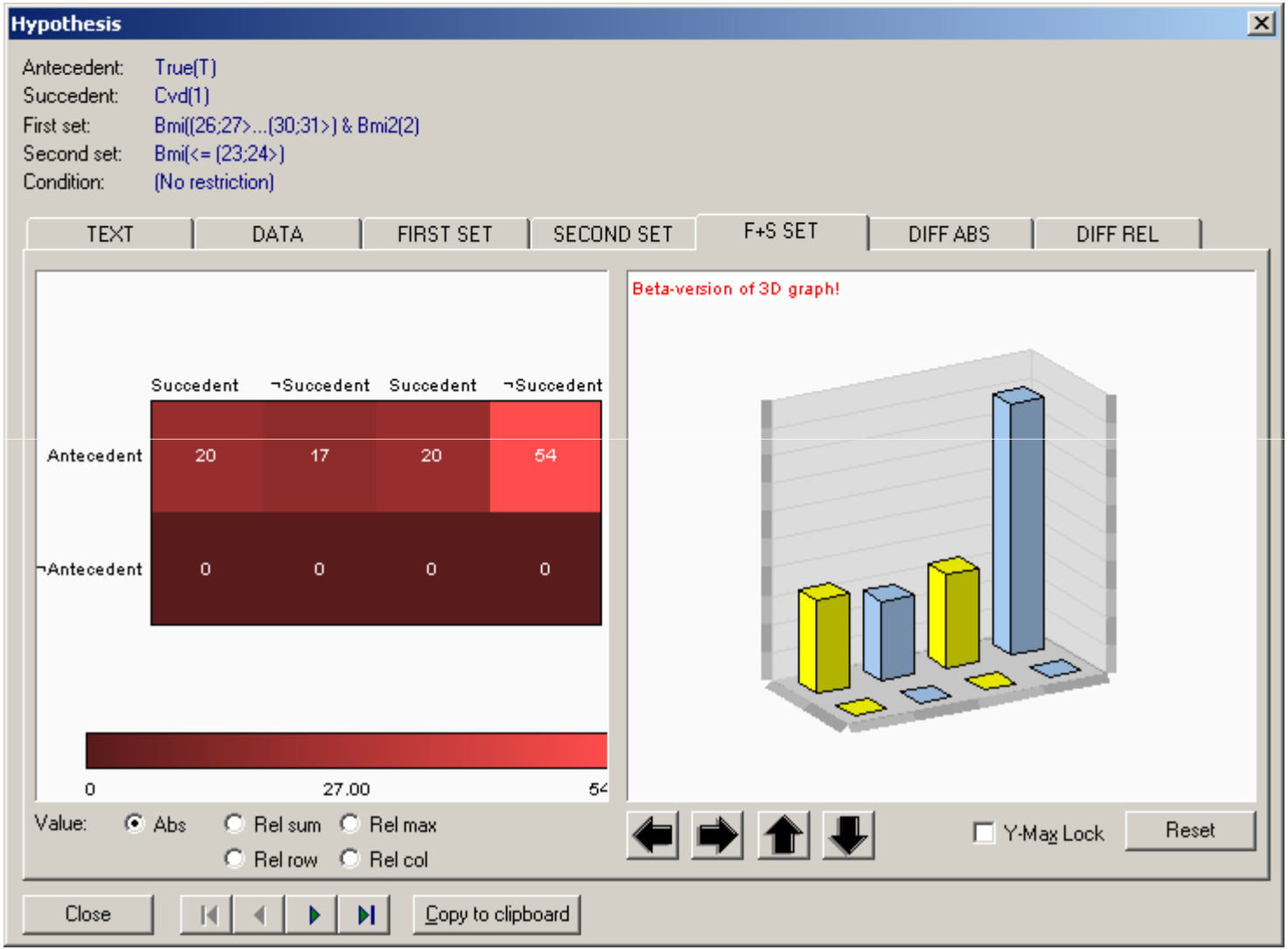
Skupina_1	CVD	¬CVD
TRUE	a_1	b_1
¬ TRUE	*0*	*0*

Skupina_2	CVD	¬CVD
TRUE	a_2	b_2
¬ TRUE	*0*	*0*

Příklad formálního vyjádření odlišnosti:

$$|a_1/(a_1+b_1) - a_2/(a_2+b_2)| \geq 0.2 \wedge a_1 \geq 20 \wedge a_2 \geq 20$$

Výsledky Lisp Miner



Závěr

- obdobu uvedených pravidel se pomocí „klasických“ postupů (např. lineární aproximace) nepodařilo objevit, což ukazuje na výhody použité nové metody
- LispMiner byl obohacen o nový vzorec pro hledání asociačních pravidel

Děkuji za pozornost