Generování umělých dat
Zpoždění vlaků

Zpráva pro vyučujícího

Tomáš Pokorný

Jan Žítek

6. 12. 2015

Obsah

- Příprava dat
- Zadání požadovaných vztahů
- Dílčí generování dat
- Finální generování dat
- Randomizace
- Připomínky k ReverseMineru

Příprava dat (1)

- Zdroje dat
  - Doménové znalosti
    - Čas příjezdu
    - Datum
    - Mimořádná událost
    - Provoz na trati
    - Typ vlaku
    - Výluka
    - Zpoždění
  - Meteorologické údaje (externí data)
    - Teplota
    - Srážky
  - Kalendářní údaje (externí data)
    - Svátek

Příprava dat (2)

- Vytvoření pseudotabulky – název, hodnoty, datový typ, rozdělení
  - Sloupce definované výčtem hodnot
    - Mimořádná událost – ano, ne (nominální; 5:95)
    - Provoz na trati – malý, střední, velký (ordinální; 30:50:20)
    - Svátek – ano, ne (nominální; dle sloupce Datum)
    - Typ vlaku – Os, Sp, Ec, Rj, Sc (nominální; 83:4:2:9:2)
    - Výluka – ano, ne (nominální; 5:95)
  - Sloupce definované minimální a maximální hodnotou
    - Čas příjezdu – <0;23.99> (desetinné číslo; rovnoměrné)
    - Srážky – <0;50.5> (desetinné číslo; dle sloupce Datum)
    - Teplota – <-50.5;50.5> (desetinné číslo; dle sloupce Datum)
    - Zpoždění – <0;300> (celé číslo)
      - 20 % hodnota 0; 80 % Gaussovské rozdělení (0;85)
  - Sloupce definované kombinací
    - Datum – zadán celý první týden v roce a poslední den v roce, z důvodu odvozeného atributu Den v týdnu (datum; rovnoměrné)

Příprava dat (3)

- Vytvoření atributů
  - Nad každým sloupcem vytvořen atribut výčtem hodnot
  - U zvolených sloupců vytvoření dalších atributů (ekvidistantní, expertní, binární)
  - Specifickým sloupcem Datum – vypočtení odvozených hodnot
    - Den v týdnu – Po, Út, St, Čt, Pá, So, Ne; pracovní den/víkend
    - Den v roce – vhodným nastavením intervalu hodnot vznik ročních období

Požadované vztahy

- Vztahy pro 4ft-Miner
  - Mimořádná událost(ano) $\Rightarrow_{0.6,\ 2\ \%}$ Zpoždění(<30;300>)
  - Teplota(<=4) $\wedge$ Srážky(vysoké)) $\Rightarrow_{0.3,\ 2\ \%}$ Mimořádná událost (ano)
  - Teplota(<=4) $\wedge$ Čas příjezdu(<4;9)) $\Rightarrow_{0.6,\ 2\ \%}$ Mimořádná událost (ano)
  - Provoz(Velký) $\wedge$ Výluka(Ano) $\Rightarrow_{0.8,\ 2\ \%}$ Zpoždění(>= 5)
  - Provoz(Velký) $\Rightarrow_{0.6,\ 2\ \%}$ Zpoždění(>= 5) / Vlak(Os)
  - Provoz(Střední) $\wedge$ Výluka(Ano) $\Rightarrow_{0.4,\ 2\ \%}$ Zpoždění(>= 5) / Vlak(Os) ? Vlak(Sp)
  - Den(pracovní) $\wedge$ Svátek(ne) $\Rightarrow^{AA}_{0.3,\ 2\ \%}$ Vlak(zpožděn)
- Vztah pro CF-Miner
  - Vlaky $\approx_{Sum(Zima)>=55\ \%\ C}$ Roční období/Mimořádná událost(ano)
- Vztah pro KL-Miner
  - Vlaky $\approx_{TauB\ 0.4}$ Srážky x Zpoždění

Dílčí generování dat – úvod

- Cílový počet generovaných řádků 500
- Generování dat dle zaměření ukrytých vztahů
  - Frekvenční požadavky
  - Vztahy pro 4ft-Miner
  - Vztahy pro KL-Miner a CF-Miner

Dílčí generování dat – frekvenční požadavky

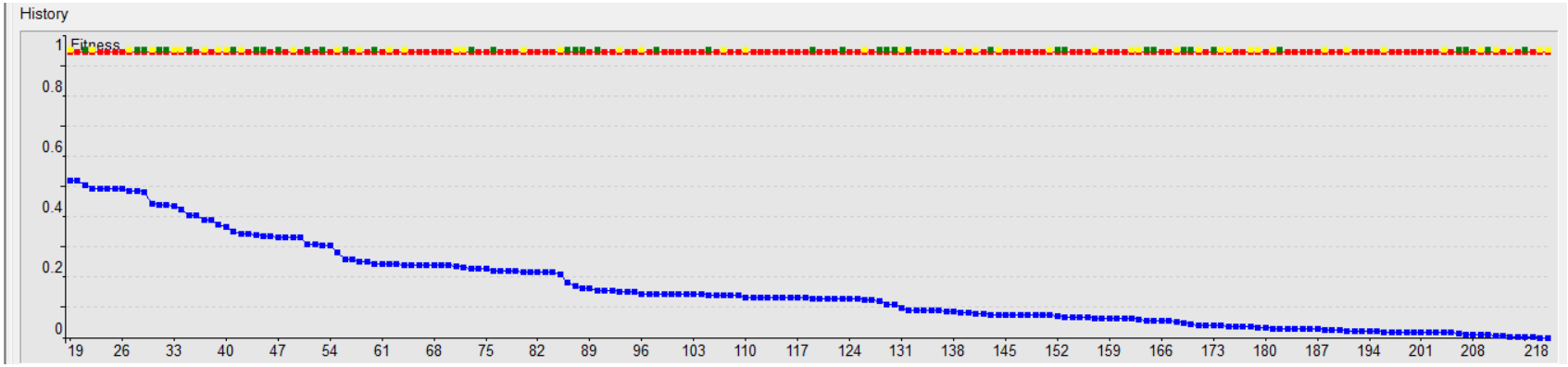
- Zadání frekvenčních požadavků v dílčích úlohách
  - Požadavky týkající se mimořádných událostí, provozu na trati, typu vlaku a výluk
  - Požadavky týkající se času příjezdu
  - Požadavky týkající se zpoždění
- Následně vytvoření úlohy obsahující všechny frekvenční požadavky, v které jsou jako Data presets zadány tři výše zmíněné úlohy
  - 98 iterací, trvání evoluce 2h 34m

Dílčí generování dat – 4ft-Miner

- Zadání vztahů pro 4ft-Miner po jednom a spuštění evoluce
  - 6 vztahů typu FUI + 6 vztahů typu BASE
  - 1 vztah typu AAD, 1 vztah typu BAD, 1 vztah typu BASE
- Následně vytvoření úlohy obsahující všechny vztahy pro 4ft-Miner a frekvenční požadavky (s váhami odpovídajícími důležitosti konkrétní úlohy)
  - Po 2h 30m a 56 iteracích evoluce ručně zastavena, vztahy pro 4ft-Miner splněny, pouze některé frekvenční požadavky nesplněny

Dílčí generování dat – KL-Miner, CF-Miner (1)

- Nejprve zadání vztahů pro KL-Miner
  - Průběh evoluce viz obrázek níže



Dílčí generování dat – KL-Miner, CF-Miner (2)

- Po úspěšné evoluci vytvoření klonu úlohy a připojení frekvenčních požadavků
  - U problémových frekvenčních požadavků použitá tolerance
  - Jako Data presets použita úloha pro všechny frekvenční požadavky a výše uvedená úloha pro KL-Miner
  - Konec za 10h 14m, 217 iterací

- Vztah pro CF-Miner zadán ke klonu předešlé úlohy
  - Jako Data presets zadána předešlá úloha

Finální generování (1)

- Vytvoření úlohy obsahující všechny požadované vztahy
- Jako Data presets použity dvě úlohy
  - Úloha spojující 4ft-Miner a frekvenční požadavky
  - Úloha spojující KL-Miner, CF-Miner a frekvenční požadavky
- Počátek generování dat především ve znamení hledání vhodných parametrů evoluce
- Po několika pokusech generování se jako problémový ukázal vztah pro 4ft-Miner
  - Teplota(<=4) $\wedge$ Srážky(vysoké)) $\Rightarrow_{0.3,\ 2\ \%}$ Mimořádná událost (ano)
  - Přisuzujeme to skutečnosti, že v roce 2014 nebylo příliš dní, kdy klesla teplota pod 4°C a zároveň byly vysoké srážky
  - Tento vztah tedy vyřazujeme z evoluce a již nepožadujeme jeho splnění

Finální generování (2)

- Poslední nastavení parametrů evoluce

Fitness Tolerance:	0.000	Crossover prob:	50 %	Row Random	50 %	Mut. Swap prob:	10 %
Max Iterations:	1000	Mutation prob:	40 %	Row Half prob:	50 %	Mut. Modify prob:	0 %
Max Time:	12 hour(s) + 0 mins	Breed prob:	5 %	Cross Col prob:	50 %	Mut Fill prob:	90 %
Population size:	50	Reprod. prob:	5 %	Cross Row prob:	50 %	Swap Length Max:	4
Initial Population:	50					Swap Count Max:	5
Tournament size:	15			Breed Length Max:	3	Modify Len Max:	4
Noise % min:	-			Breed Copy Max:	15	Modify Count Max:	5
Description:						Fill Length Max:	4
						Fill Count Max:	5

- Evoluce zastavena automaticky po 12 hodinách
  - Nesplněny dva méně důležité frekvenční požadavky, nicméně pouze v řádu jednotek tisícin
  - Takový výsledek evoluce jsme označili za **uspokojivý**

Randomizace

- Zvětšení rozsahu dat z 500 na 2000 řádků
- 5 kroků randomizace
  1. Zvětšení rozsahu dat a randomizace sloupce Datum (Best noise 8.960 %)
  2. Randomizace sloupce Výluka (Best noise 1.780 %)
  3. Randomizace sloupce Mimořádná událost (Best noise 0.745 %)
  4. Randomizace sloupců Provoz na trati a Typ vlaku (Best noise 2.415 %)
  5. Randomizace sloupců Čas příjezdu a Zpoždění (Best noise 4.820 %)

Připomínky k ReverseMineru

- Po vlastních zkušenostech bychom uvítali automatické ukládání průběhů evolucí
  - Alespoň předem zaškrtnuté políčko *Save fitness tab in each iteration*