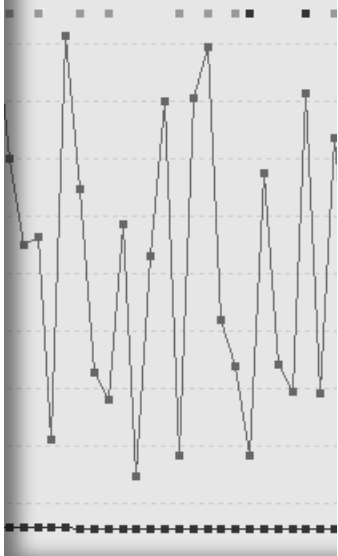
Prezentace pro vyučujícího

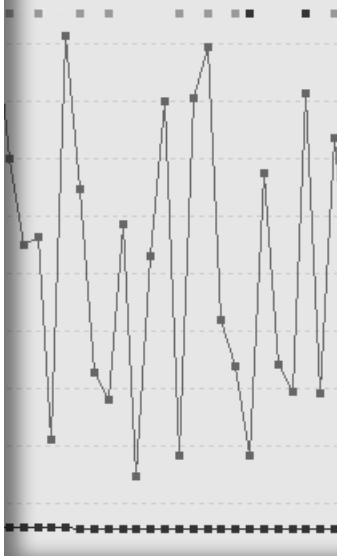
Generování umělých dat

Ludmila Svobodová



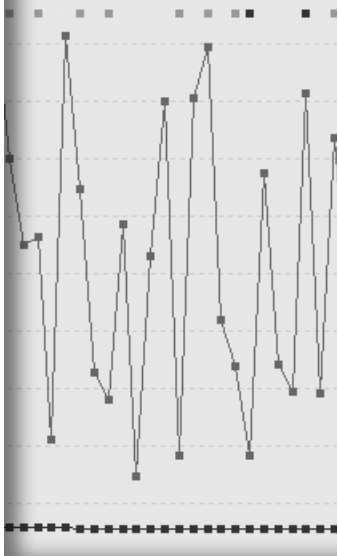
Milníky práce

- Seznámení se se základními vztahy domény.
- Příprava dat (pseudo-tabulka).
- Zadávání a generování dat pro jednotlivé vztahy.
- Generování kompletních dat se všemi vztahy.



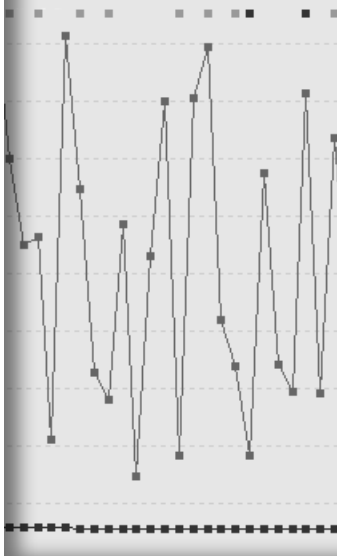
Tvorba pseudo-tabulky

- U nominálních dat jsem do pseudo-tabulky vypsala kompletní výčet přípustných hodnot např. u okresu, jsou v pseudo-tabulce vypsány názvy všech okresů v ČR.
- Zbylá data jsou ordinální, konkrétně typu Integer number, tedy do pseudo-tabulky stačilo u těchto dat napsat minimální a maximální možnou hodnotu, kterou budou nabývat.



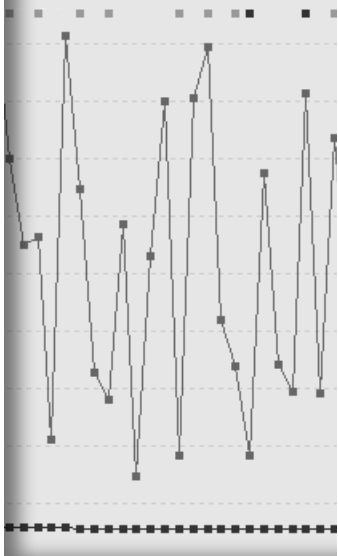
Atributy v LMWorkspace

- Po nahrání pseudo-tabulky do LMWorkspace, jsem zkontrolovala, zda u všech dat byl dobře rozeznán datový typ.
- Dále jsem vytvořila pro každý sloupeček pseudo-tabulky příslušný atribut.



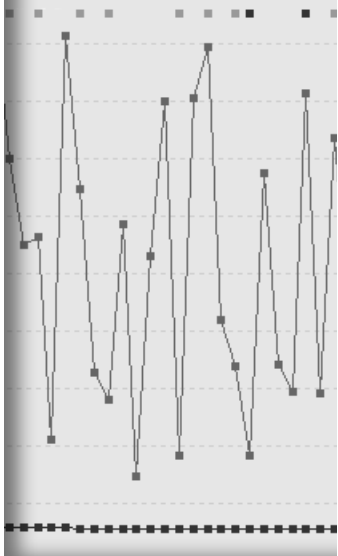
Vytvoření atributů <intervaly> 1/2

- U vytváření atributů Věk, Plat a Délka jsem musela zvolit intervaly, na které hodnoty rozdělím.
- **Věk (ordinální atribut)**
  - Rozdělení na intervaly velikosti 5 jsem zvolila proto, že plně pokrývá mé potřeby.
  - Nikde nepotřebuji využít menší intervaly a pokud někde potřebuji využít jiné rozdělení intervalů, můžu je z těchto základních intervalů vytvořit přesně jak potřebuji.
  - Např. nemocenská z důvodu porodu a šestinedělí je nejčastější mezi 25 a 35 lety. Vytvořím si tedy klon atributu Věk a v něm sloučím intervaly <25;29> a <30;34>. Tím jsem dostala interval <25;34> a mohu zadat větší četnost porodu a šestinedělí právě v tomto intervalu.



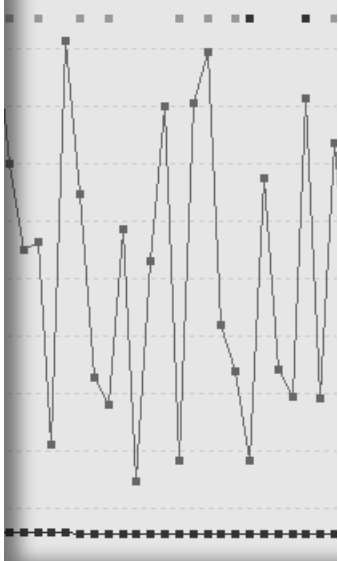
Vytvoření atributů <intervaly> 2/2

- **Plat (ordinální atribut)**
  - Plat jsem v žádném zadaném vztahu neuplatnila, ale k doméně zajisté patří, proto jsem ho do generování zanesla a proto jsem intervaly rozdělila po 10 000, což v mém případě stačí a vyhovuje.
- **Délka (ordinální atribut)**
  - Rozdělení na intervaly velikosti 10 jsem zvolila proto, že se prozatím ukázalo, jako nejlepší pro mé účely.
  - Atribut jsem nevyužila pro klonování dalších atributů a toto rozdělení se ukázalo jako dobré pro zadání všech vztahů, kde se Délka využívá.



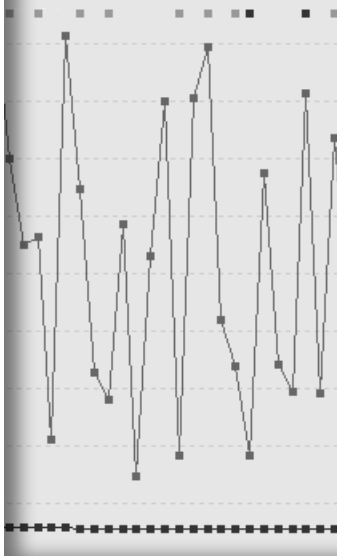
Dílčí úlohy

- Nejprve jsem zadávala všechny vztahy jednotlivě a kontrolovala, zda po rozumném množství kroků evoluce zkonverguje a dá pro požadovaný vztah výsledek.
- Nejdéle trvalo vygenerovat data pro vztah, že čím starší člověk, tím delší doba jeho nemocenské. Tento vztah jsem zadala pomocí KL-mineru a použití Kendallova TauB koeficientu. Vzhledem k náročnosti generování jsem nakonec nastavila požadovanou míru TauB pouze na 0,5.
- Vztah, že dlouhé nemocenské jsou v Šumperku jsem zadala i pomocí CF-mineru, ten ale běžel velice dlouho i jako dílčí úloha, proto jsem do finální evoluce tento vztah zadala pomocí 4FT-mineru



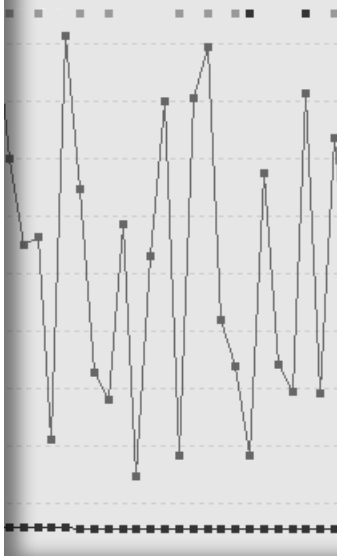
Dílčí úlohy

- Pro potřeby dílčích úloh jsem nastavila evoluci takto:
  - Pravděpodobnost křížení 30 %
  - Pravděpodobnost mutace 60 %
  - Pravděpodobnost reprodukce 10 %
  - Velikost populace 50
  - Velikost počáteční populace 100
  - Velikost turnaje 5
- S tímto nastavením došly všechny dílčí úlohy k výsledku, proto jsem jej použila i jako výchozí nastavení pro finální evoluce.



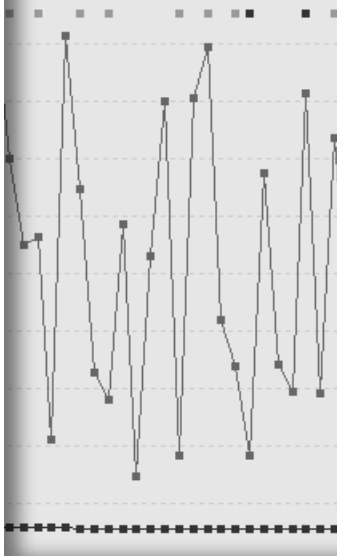
Finální úlohy

- Do konečné evoluce jsem zahrnula všechny jednotlivé vztahy a přidala i pravidla pro hlídání frekvencí atributů Důvod a Diagnóza.
  - Při zkoušení finální evoluce jsem zjistila, že držení frekvencí pro diagnózu je pro evoluci příliš omezující.
- Dále jsem se u většiny vztahů snažila evoluci pomoc. Proto jsem vztahy 4ft-mineru zadala jednou s base, potom bez base a ještě jsem u AAD (BAD) kvantifikátoru zadávala vztahy inverzní, tedy pokud by nebyl dosažen ani průměrný výskyt požadovaného jevu, tak aby měla evoluce vodítko jak k němu dojít.
- Další skupinou vztahů, které jsem použila a netýkají se vztahů samotných přímo, jsou negativní vztahy.
  - Tedy pokud chci mít v datech četnost nějakého výskytu větší, tak negativním vztahem zajišťuji, že jiný výskyt nebude četnější.



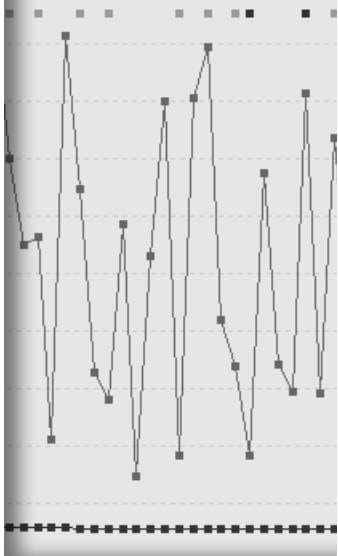
Průběh finální evoluce

- Při finální evoluci jsem ladila jak parametry evoluce, tak nastavení jednotlivých vztahů (kvantifikátorů).
  - V několika případech jsem musela na nastavení kvantifikátoru polevit, aby se vztahy vzájemně nepraly a nevylučovaly.
- Nastavení finální evoluce vycházelo z nastavení dílčích úloh.
  - Zvýšení velikosti populace na 70.
  - Zvýšení velikosti turnaje na 15.



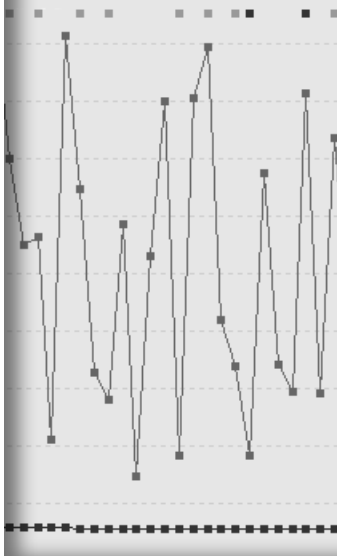
Finální evoluce

- I po mnoha pokusech a dlouhých evolucích se mi nepodařilo dosáhnout úspěšného dokončení evoluce ve všech vztazích.
- Nejaktuálnější data jsou ale zatím nejlepší, protože žádný vztah není úplně zamrznutý a nejhorší má reálnou hodnotu 0,622 a to je to vztah na hlídání frekvence atributu.
- Výsledek je proto zatím příznivý, protože i když u všech vztahů nedošla evoluce k cíli, tak všechny vztahy tam alespoň z části již jsou.
  - V předchozích evolucích byl vždy problém absolutního zamrznutí a vyloučení některých vztahů.



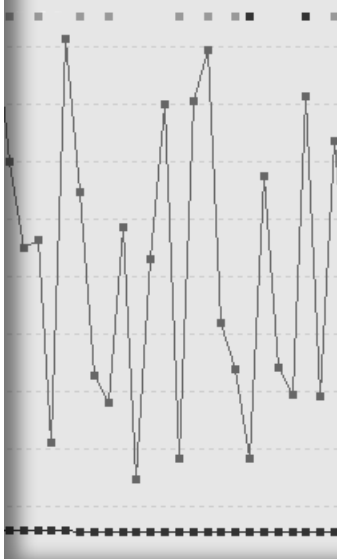
Návrhy na vylepšení ReverseMineru

- Vzhledem k dlouhému trvání evoluce, by byla zajímavá možnost evoluci ukládat a znovu pouštět třeba i na jiném PC.
- Kdyby byla evoluce v průběhu ukládána tak, aby bylo možné i po nečekaném vypnutí PC evoluci znovu obnovit (např. když někdo vyhodí v celém bytě pojistky).



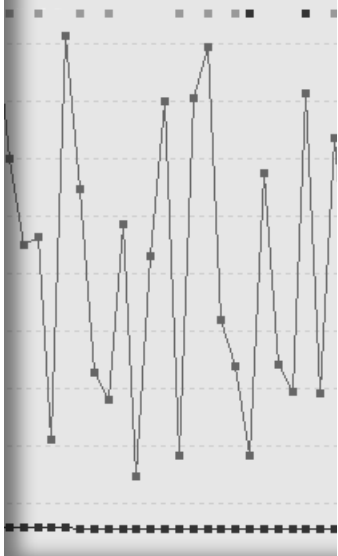
Vylepšení ReverseMineru **(již vyřešená/v řešení)**

- Hodně mocná funkce ReverseMineru by mohla být možnost upravovat již spuštěnou evoluci. Např. pokud po 1 000 iterací vidím, že se mi dva vztahy perou, nebo mám dokonce některý zadaný špatně, že to s ním nepůjde dokončit, byla by super možnost moct evoluci stopnout, vztahy upravit a pustit jí dál. Je škoda, že se pak musí zahodit dlouhý výpočet a nedojde se k cíli např. kvůli dvěma vztahům, i když už zbytek vztahů je na velmi dobré cestě.
- Vyřešení problému, proč iterace na některých počítačích zamrzávají, protože to znemožňuje běh dlouhých výpočtů např. přes noc.



Změny oproti zadání 1/2

- Za prvé jsem změnila strukturu dat.
- Vyloučení okresu ze skupiny Osoba.
  - Zjistila jsem, že zatím nemá cenu řešit, ve kterém okresu osoba pracuje a kde bydlí.
- Vyloučení kraje ze skupiny Zaměstnání.
  - Je to vlastně duplicitní informace, kraj se dá určit podle okresu.
- Vyloučení výskytu ze skupiny Nemocenská.
  - Kdybych chtěla vytvořit data, obsahující případy lidí s nemocenskou ale i bez ní, musely by být v evoluci vztahy typu, pokud Výskyt(NE) tak vlastně všechny atributy týkající se nemocenské nesmí být vyplněny.
  - Znamenalo by to obrovské zpomalení evoluce.



Změny oproti zadání 2/2

- Nepoužití externích dat.
  - Nakonec jsem se nedostala ke generování vztahů, které by využívaly externí data a proto nakonec zavedení externích dat nebylo potřeba.
- Vztah nejvíce nemocenských je v okresech Prachatice a Šumperk, nejméně se jich vyskytuje v Praze a Jeseníku.
  - Nebyl zanesen do výsledné evoluce, protože počet okresů na generovaná data je poměrně velký.
  - Pravidla, která by hlídala frekvenci daného vztahu by tedy byla dost omezující a v této fázi by znamenala značné zpomalení evoluce.