Předzpracování dat pro data mining: metody a nástroje

Olga Štěpánková, Zdeněk Kouba,
P. Mikšovský, P. Aubrecht

Gerstnerova laboratoř pro inteligentní rozhodování a řízení
České vysoké učení technické v Praze

Získávání znalostí z dat (KDD)

- **Cíl:** částečná automatizace procesu získání **zajímavých vzorů chování z reálných dat:** tvorba jejich modelů - např. pomocí nástrojů strojového učení
- **Aplikace**: průmysl (diagnostika poruch), obchod (marketing, bankovnictví), věda (charakterizace karcinocenních látek), ...
- Nové slibné odvětví SW průmyslu, jehož cílem je využít existující data pro **zlepšení rozhodovacích procesů**

- **Problémy reálných dat?**
- Data nejsou sbírána jako zdroj trénovacích příkladů, ale především kvůli podnikové dokumentaci a archivaci. Z tohoto hlediska bývá sběr i uložení optimalizováno.

Problémy reálných dat?

- Data obsahují **špatné údaje** způsobené chybami měřicích přístrojů i lidské obsluhy
- **Nevyplněné údaje**

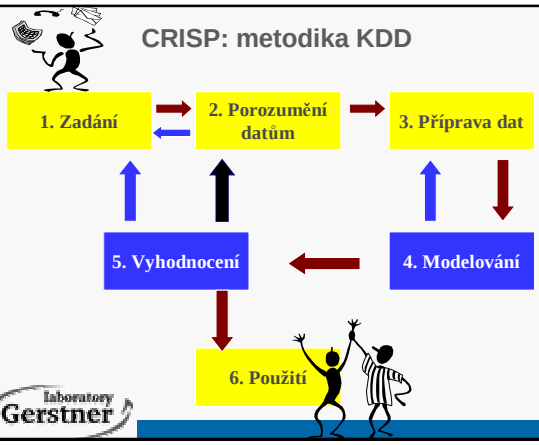
- Data jsou popsána pomocí **příliš mnoha atributů** - není zřejmé, které z nich jsou pro řešení zvolené úlohy relevantní. Úspěch modelování závisí na volbě vhodné množiny atributů (PAC učení)

- Data mají formu **složitého relačního schématu**, nikoliv jediné tabulky předpokládané atributovými metodami strojového učení

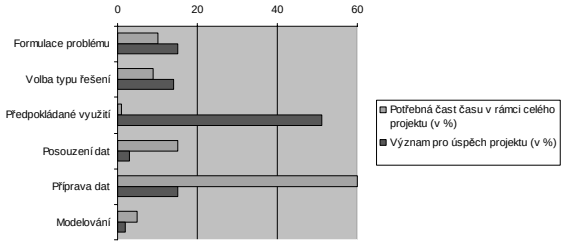
Příprava dat nabízí různé způsoby, jak se s těmito problémy vyrovnat.

První pokus o přípravu dat nebývá ten nejsprávnější - **nutné opakování**

CRISP: metodika KDD



Časové nároky na přípravu dat?



Úkoly předzpracování dat

1. Chyby v reálných datech
   - Na chyby je třeba upozornit a přijmout rozhodnutí, jak s nimi naložit (například doplnit údaje ve spolupráci s expertem z oboru aplikace nebo naopak využít statistických zákonitostí).
   - U některých atributů se stává, že vyplnění údaje je skoro výjimkou – mluvíme pak o řídce (sparse) obsazených atributech
2. Množina vlastností (atributů)
   - Bohatost dat (počet dimenzí) má zásadní vliv i pro úspěch použití technik strojového učení, neboť s dimenzí exponenciálně rostou i nároky na počet trénovacích příkladů – snaha zredukovat
   - Doplnění výchozích dat o nové informace na základě dalších datových zdrojů,
3. Velikost celého výchozího datového souboru - výběr podsouborů pro modelování
4. První pokus o přípravu dat nebývá ten nejsprávnější - nutné opakování

Analýza a úprava jednotlivých atributů I.

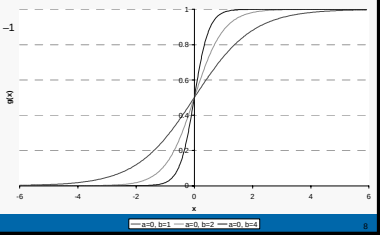
- **Zpráva o stavu proměnných**
  - typ (spojitá X diskrétní)
  - rozsah definičního oboru (počet použitých hodnot)
  - rozsah a frekvence výskytů (histogram)
  - typ rozdělení a jeho statistické charakteristiky
- **Upozornit na**
  - osamělé mimořádné hodnoty (outliers)
  - téměř konstantní atributy (možné vynechat)
  - nevyplněná datová pole
  - znečištění dat
    - data neodpovídají deklarovanému formátu
    - hodnoty neodpovídají deklarované množině

Analýza a úprava jednotlivých atributů II.

Příklady úprav

- **Náhrada chybějících údajů** - provádí se tak, aby zůstala zachována hodnota směrodatné odchylky uvažovaného atributu
- **Úprava rozsahu hodnot** atributů pomocí logistické transformace (velmi důležité v každé metodě, která počítá se vzdáleností objektů - např. CBR, nejbližší sousedi, shlukování) ,

$$g(x) = (1 + e^{-a - bx})^{-1}$$



Analýza a úprava jednotlivých atributů III.

- **Monotónní atributy** – představují obvykle jednoznačnou identifikaci pro uvažované objekty, např. pořadové číslo měření, číslo bankovního účtu. Rostou bez omezení a při tom jejich přímá hodnota jako taková nemá pro vytvoření modelu význam.
- **Řady** – tvořené hodnotami veličin, které jsou pravidelně měřeny a zaznamenávány (např. EKG, burzovní koeficienty). Vždy jsou vztaženy k jediné monotónní veličině, která slouží jako index.
  - často jako index slouží čas -> časová řada
  - Prostředky k analýze:
    - **Fourierova analýza**
    - **Vlnková (wavelet) transformace**[1] umožňuje získání časově-frekvenčního popisu signálu

[1]Louis, A., K., Maas, P., Rieder, A.: Wavelets: Theory and Applications. Wiley, 1997.

Úpravy a analýza dat ve stav.prostoru I.

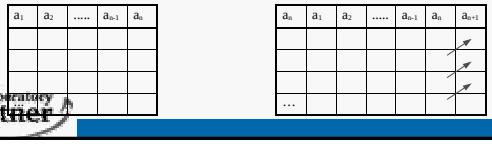
Příklady úprav

- **Snížení dimenze**
  - vynecháním
    - **konstantních** atributů
    - atributů **řídce obsazených**
    - atributů **s duplicitní informací** (rok narození X věk, apod.)
  - sloučením
    - atributů **řídce obsazených** – z několika řídce obsazených atributů je možné zřetězením vytvořit jeden nový (PVP - present value pattern)

Úpravy a analýza dat ve stav.prostoru II.

Příklady úprav

- **Zvýšení dimenze**
  - **obohacení** doplněním údajů z jiných zdrojů (např. meteorologická měření, demografické údaje, apod.)
  - **rozšíření**
    - přidání odvozených atributů (např. pohlaví z rodného čísla, apod.)
    - „otočení" dat (reverse pivoting) - nový atribut a_{n+1} přebírá údaj z objektu následujícího. Pro každý objekt *i* platí $a_{n+1}(i) = a_n(i+1)$.



Úpravy a analýza dat ve stav.prostoru III.

Příklady úprav

- **Agregace dat** - použití metod datových skladů. údaje o více objektech obsažené na několika řádcích jsou vztaženy k jedinému obecnějšímu objektu (tvoří tedy v novém souboru jedinou řádku).
- **Vizualizace** – např. umístění datového souboru ve stavovém prostoru úlohy, přirozené shluky, nepravidelné deformace,...
- **Statistické přístupy snižování dimenze**
  - podmíněná entropie
  - CHAID (Chi-square Automatic Interaction Detector)
  - hledání hlavních komponent (návrh vhodné lin. kombinace)
- **Využití neuronových sítí** - Řídce propojená autoasociativní neuronová síť (Sparcely Connected Autoasociative Neural Net: SCANN)
- **Pozor na přídání anachronických atributů !!!**

Úprava souborů pro modelování I.

Příklady úprav

- **Hlavní zásada:** každý nový soubor musí s rozumnou dávkou důvěry zachovávat původní pestrost či rozložení výchozího souboru.
- **Vytvoření trénovacích a testovacích dat**
- **Vzorkování dat** (sampling)
- **Navýšení vlastnosti** v trénovacích datech (pro vzácně se vyskytující hodnoty atributu)
  - trénovací množina jako sjednocení 2 nezávislých podmnožin, z nichž v první všechny objekty mají uvažovanou vzácnou hodnotu atributu, v druhé nikoliv
  - namnožení dat s požadovanou hodnotou atributu přidáním „bílého šumu“

Předzpracování dat pro data mining: metody a **nástroje**

SumatraTT

Úvod

Čeho jsme chtěli dosáhnout? *Snížení pracnosti nezbytné k*:

- *opakovaným činnostem*
  - s využitím standardizace a automatizace → efektivní opakované použití vyvinutých komponent
- *vytváření dokumentace:* zdlouhavé, ale nezbytné dokumentování experimentu tak, aby jej bylo možné zopakovat např. na testovacích datech

→ [2]**Prepared Information Environment** (PIE)

Jak? SumatraTT

- Univerzální, metadaty řízený, transformační nástroj
- Postaven nad skriptovacím jazykem orientovaným na datové transformace, syntaxí podobným jazyce Java
- K vytváření datové trasformace využívá knihovnu šablon (templates)
- Podporuje automatickou dokumentaci datové transformace

[2]Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann, California, 1999.

Předzpracování dat pomocí SumatraTT

- Neznámá data
  - zjistit strukturu
  - prohlédnout hodnoty
- Navrhnout převod
- Dokumentovat
- Spustit (vícenásobně)

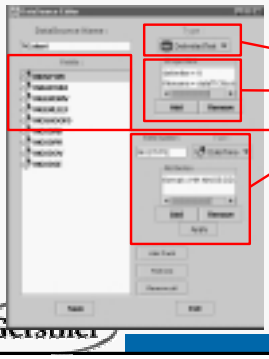
Datový zdroj

- Součástí definice je i typ zdroje (lze změnit)
  - textový soubor
  - SQL (ODBC, ADO)
  - regulární výraz (podobně jako Perl)
  - JavaDB (vzdálené JDBC, komprese, šifrování)
  - Prolog
  - WEKA
- Unifikovaný popis atributů (fieldů)
- Upřesnění atributu (délka, SQLType, formát)
- Existuje sada automatických wizardů

Laboratory Gerstner

17

Parametry datových zdrojů



- typ
- parametry DS*
- seznam atributů
- parametry atributu

* podle typu (jméno souboru a oddělovač, SQL tabulka, dotaz)

18

SumatraScript

- Univerzální interní skriptovací jazyk
- Veškeré zpracování v jazyku SumatraScript
- Synatxí podobný jazyku Java, přístup k datovým zdrojům
- SumatraTT vlastně pouze poskytuje sady funkcí a interních informací (metadat)
- Kvůli opakujícímu se kódu byl jazyk rozšířen o možnost generování části svého kódu – vznikla makra

Laboratory Gerstner

19

Šablony transformací

- Způsob zpracování dán šablonou (kostra programu)

+

- Specifikace pomocí parametrů
  - vstupní, výstupní datový zdroj
  - vnitřní kód
  - délka udržované historie
  - volitelná akce (RemoveDuplicities)
  - ...

Laboratory Gerstner

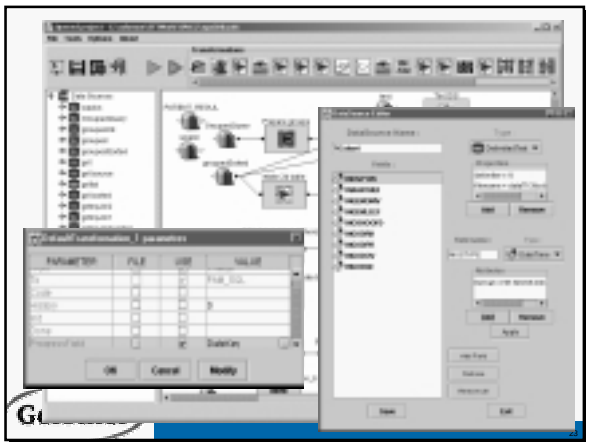
20

Dostupné šablony I.

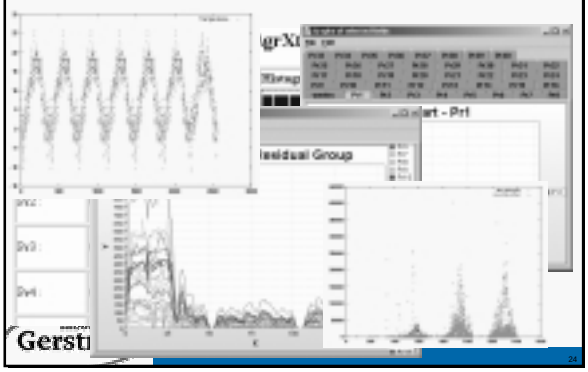
- **TableCopy** – prosté kopírování, převod mezi formáty, filtrování, výpočet nových atributů
- **Table1toN** – rozdělení záznamu na několik, spojení
- **TableSample** – náhodné rozdělení DS na dva (trénovací, testovací)
- **FairSubset** – podmnožina o přesně daném počtu prvků (nebo v procentech)

Dostupné šablony II.

- **TableReport** – textová zpráva (hledání chyb)
- **CheckDS** – kontrola čitelnosti datového zdroje
- **gnuplot** – vykreslení dat pomocí gnuplot, 2D i 3D
- **RemoveDuplicities** – zpracování duplicit v SQL databázích
- **CrossTable** – podpora (poloautomatická) pro křížové tabulky
- **ExecSQL** – spustí SQL příkaz



Vizualizace – statická



Vizualizace – interaktivní

Další nástroje

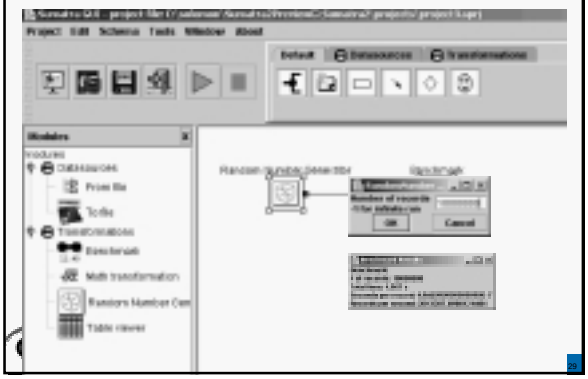
Dostupnost

- SumatraTT verze 1
  - dostupná na adrese
  - http://krizik.felk.cvut.cz/Sumatra
- Instalace obsahuje i sadu šablon pro běžnou práci i pro složitější zpracování

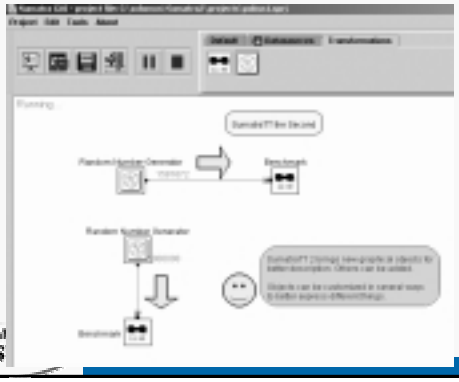
Budoucí vývoj I.

- SumatraTT verze 1
  - omezené možnosti rozšíření
  - dualita řešení: C++ a Java
- **SumatraTT verze 2**
  - pouze v Javě
  - důraz na uživatelsky přívětivé GUI
  - založeno na standardech – XML, JDBC, CORBA

Ukázka verze 2 – návrh transformace



Ukázka verze 2 - běh transformace



Dostupnost

- SumatraTT verze 1
  - dostupná na adrese
  - http://krizik.felk.cv
- Instalace … šablon pro běžnou práci … pracování

… a nyní předvedení Sumatra TT