Comparison of methods for imputation of missing values

D. Pejčoch

Osnova

Blok 1

- Vlastnosti dat, neúplná data
- Důsledky neúplných dat

Blok 2

- Klasifikace metod pro doplňování neúplných pozorování
- Stručný popis používaných metod

Blok 3

- Stávající benchmarky a jejich „Key Learnings“
- Návrh komplexního benchmarku

Datová kvalita a řízení dat

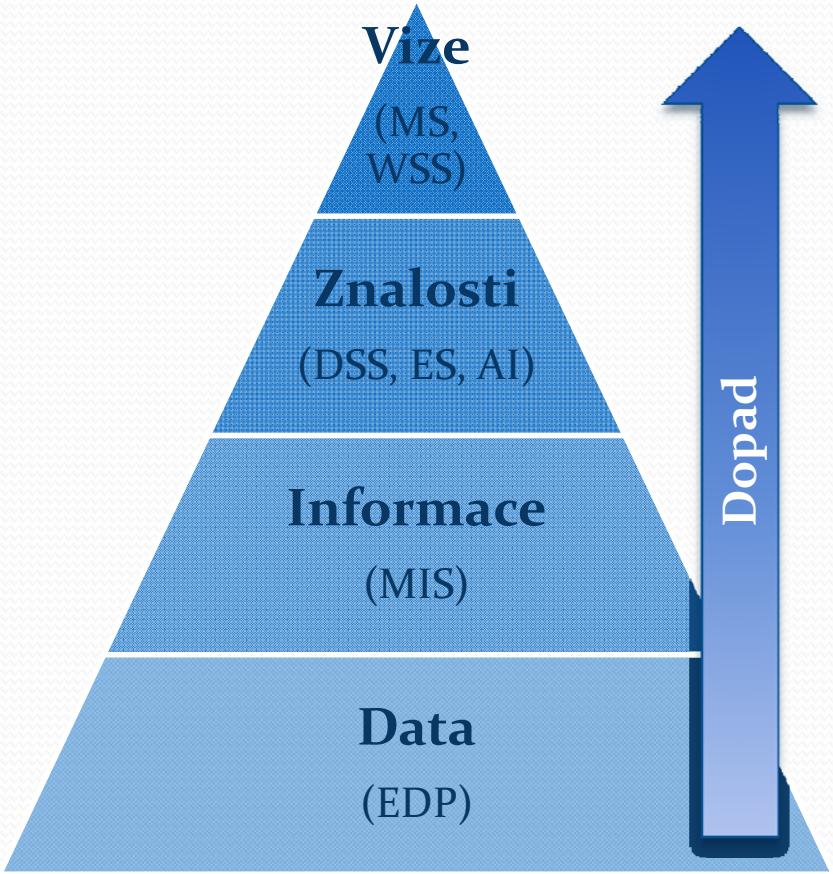
Datová kvalita: mnoho definic, obecně míra naplnění vlastností dat (objektivních, subjektivních)

Úplnost = jedna z klíčových vlastností dat, resp. metrik výkonnosti řízení dat

Správně neúplná data: nejsou pro daný subjekt k dispozici

Chybně neúplná data: hodnoty jsou reálně k dispozici

Jiný případ: Cenzorovaná pozorování



Pyramida znalostí

Úplnost jako klíčová vlastnost dat

- Světová odborná literatura: rozsáhlé monografie na toto téma
- Tuzemská odborná literatura: drobné zmínky v publikacích zabývajících se statistickou analýzou (Hebák) a data miningem (Berka)
- … znamená to, že nás toto téma netrápí?

Mechanismy výskytu chybějících hodnot

- **MCAR (Missing Completely at Random)** = chybějící hodnoty mají stejnou pravděpodobnost výskytu pro všechny záznamy. Záznamy s chybějícími hodnotami nejsou přitom nijak odlišitelné od těch bez chybějících hodnot.
- **MAR (Missing at Random)** = příčina chybějící hodnoty nezávisí na proměnné, v rámci níž se vyskytuje. **Lze je na základě ostatních proměnných predikovat.**
- **MNAR (Missing Not at Random)** = příčina výskytu závisí pouze na proměnné samotné. Konkrétní příčinou může být např. fakt, že pro daný záznam tato proměnná nebyla naměřena nebo byla data proměnné doplněna z externího zdroje pouze pro část záznamů
- **MBND (Missing By Natural Design)** = příčinou chybějící hodnoty je nemožnost jejího fyzického měření

Testování MCAR

- Mechanismus výskytu chybějících hodnot určuje použitelné techniky pro jejich odstranění (viz dále).
- Pomocí t-testů nebo speciálního Littleova MCAR testu lze testovat hypotézu, že chybějící hodnota je MCAR oproti alternativní hypotéze, že se jedná o MAR.
- Bez dalších dodatečných informací nelze testovat hypotézu, že chybějící hodnota je MAR proti alternativní hypotéze, že se jedaná o NMAR.

Důsledky neúplných dat

Analytické důsledky:

- Vynechání dat s chybějícími pozorováními => ztráta informace
- Chybné nahrazení => zkreslení

Finanční důsledky:

- Nemožnost oslovení klienta, zachránění v rámci retenčního programu, …
- Snížení efektivity přímých kampaní (není možnost follow-up)
- Chybné určení hodnoty klienta => chybné nastavení péče
- Chybná identifikace domácnosti => chybné nastavení péče
- Chybné údaje požadované regulátorem trhu (AML, účetnictví, …) => sankce

Imputace

- Obecně používaný termín pro doplnění chybějících záznamů o přijatelné hodnoty.
- Doplnění probíhá výběrem z jednoho nebo více kanditátů.
- V rámci **SI (Single Imputation)** každá chybějící hodnota doplňována pouze jednou hodnotou, v případě **MI (Multiple Imputation)** pro každou chybějící hodnotu generováno několik alternativních variant.
- Proces MI probíhá ve třech krocích:
  - generování množiny $m > 1$ hodnot
  - analýza m dílčích datových souborů vytvořených z původního datového souboru s využitím metod pro úplné záznamy
  - kombinace výsledků m analýz pro volbu doplňované hodnoty

Hledání klasifikace metod

Co autor, to různá klasifikace:

- 1) ignorování / vynechání záznamů, 2) odhad parametrů a doplnění chybějících hodnot, 3) imputing (imputace).
- 1) řízené daty, 2) založené na modelu a 3) založené na strojovém učení
- metody učení s učitelem (supervised learning), metody učení bez učitele (unsupervised learning). Učení s učitelem dále člení: pravděpodobnostní algoritmy, rozhodovací stromy a rozhodovací pravidla
- …
- Shoda v členění imputace na SI / MI
- Cíl: nalézt optimální klasifikaci zahrnující všechny myslitelné metody

Výsledná klasifikace metod

A: Ponechání status quo

- Ignorování / smazání pozorování
- Maximální využití dostupných dat

B: Databázové techniky

C: Procedury založené na imputaci

- Přístupy nezaložené na modelu
- Přístupy založené na modelu
  - Implicitní model
    - Faktoriální techniky
    - Metody založené na
  - Explicitní model
    - Parametrické modely
    - Neparametrické modely

syntéza přístupů uvedených v odborné literatuře + doplnění B

A: Ponechání status quo

- V odborné literatuře označovaný jako tradiční způsob
- **Listwise (LD, Listwise Deletion):** vynechání všech pozorování s chybějícími hodnotami bez ohledu na to, zda je atribut s chybějícími hodnotami v dané analýze použit.
- **Pairwise (PD, Pairwise Deletion):** vynechání pouze těch pozorování, která souvisejí s aktuální prováděnou analýzou.
- **Překódování:** Jiným způsobem řešení je překódování chybějící hodnoty neutrální kategorií „nevím“, „N/A“, „?“, apod.
- Vždy vedou ke ztrátě informace
- Aplikovatelnost pouze na MCAR.
- aximální hranici 5% relativní četnosti u dané proměnné.
- Standardní součástí statistických nástrojů.

B: Databázové techniky

- **Join / Merge** <= existence primárního klíče
- **Lookup** do číselníku
- **Fuzzy join / Fuzzy match**: neexistence jednoznačného primárního klíče (zohlednění přibližné shody řetězců)
- Možné použít pouze u MNAR
- Úspěšnost u přibližného porovnávání dána použitím metody pro porovnávání řetězců (porovnávací kódy, míry podobnosti, …) a charakterem atributů použitých jako primární klíč

C1: Metody nezaložené na modelu

- **Nahrazení jednou hodnotou**
  - SMI (Sample Mean Imputation)
  - Medián / modální kategorie
  - Midrange (střed rozpětí)
  - Nevýhoda těchto metod: efekt „Čechové na Řípu“

- **Buckova metoda** (podmíněný průměr): doplnění více průměrných hodnot podmíněných hodnotami ostatních proměnných; konzistentní odhady u MCAR, MAR (za předpokladu otestované nezávislosti)

- Doplnění všech přípustných hodnot => náhodný výběr bez / s vracením
- Pro longitudiální data doplnění předchozího pozorování nebo na základě klouzavého průměru

C2: Metody založené na implicitním modelu

- Vychází z implicitních vztahů mezi daty, jako je např. podobnost mezi jednotlivými pozorováními.

- **HDSI (Hot Deck Single Imputation)** = doplnění shodné hodnoty, jaká se vyskytuje u podobných reprezentantů. Záznamy rozděleny do jednotlivých tříd s využitím technik jako je např. shlukování podle nejbližšího souseda. Spíše se jedná o strategii než metodu. Značně subjektivní hodnocení příslušnosti ke třídě. Kombinace HDSI s lineární regresí => lineární kombinace kandidátů.

- **CDSI (Cold Deck Single Imputation)** = výběr kandidátů z jiného datového zdroje. **Data Fusion** = CDSI z více zdrojů současně.

- **k-NNSI (k-Nearest Neighbour Single Imputation)** s využitím M-tree indexu. Spolehlivější alternativa klasického doplňování průměrem. Problematická aplikace na kategoriální proměnné =>subjektivita stanovení nejbližší kategorie.

C2: Metody založené na implicitním modelu

- **Faktoriální metody:** PCA (Principal Components Analysis) + MCA (Vícenásobná korespondenční analýza) - pouze pro optimalizaci jiných metod

- **DCI (Dynamic Clustering Imputation)**: fuzzy shluková analýza. Shluky jsou deterministicky vytvářeny na základě měr vzdálenosti okolo instancí s chybějícími hodnotami na základě jejich podobnosti, přičemž jedno chybějící pozorování může být současně obsaženo ve více shlucích. O 20% lepší výsledky než nedpodmíněný průměr a regrese.

- **Přibližné množiny (Rough Sets)**: aproximují přesnou množinu pomocí dvojice jiných množin představujících horní a dolní odhad původní množiny. Založena na očekávání, že v databázi existují stejné nebo podobné záznamy. Záznamy v doplňovaném datovém souboru jsou nejprve rozděleny podle hodnot rozhodnutí (třídy) a poté jsou mezi nimi hledány podobnosti na základě přibližných množin. Až 99% spolehlivost.

Metody založené na explicitním modelu

- **MRI (Multinomial Regression Imputation):** pomocí GLM (i LDA doporučovanou STALOG), ale zmiňována i klasická MNČ
- **MLRI (Multinomial Logistic Regression Imputation)**: Zobecnění pro q tříd předčilo GLM, MMSI (Mean Mode Single Imputation), EM a LD (Listwise)
- **Naïve Bayes** (stabilní výsledky, jeden průchod daty, vhodné pro velké soubory) + **Bayesovské sítě** (náročné na čas)
- **EMSI** (Expectation Maximization Single Imputation): počáteční nastavení parametrů -> modifikace. Lepší výsledky než k-NN.
- **EMMI** (Expectation Maximization Multiple Imputation)
- **Support Vector Regression** (SVR). V kombinaci s GA značná náročnost na strojový čas. Dobře zafungovala u spojitých proměnných, hůře u kategoriálních. Funguje tam, kde ostatní (ANN-GA, PCA-ANN-GA) selhávají!!!

Metody založené na explicitním modelu

- Tshilidzi Marwala "Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques" Information Science Reference. ISBN: 1605663360.

- **MLP** (vícevrstevný perceptron): i nelineárně separabilní třídy
- **RBF** (radiální bazická funkce) – rychlejší trénování, odolnost proti nestac. vstupům, v praxi nepatrně lepší výsledky než MLP (nesignifikantní)
- Využití **genetických algoritmů** pro minimalizaci euklid. normy chybové fce => kombinace ANN-GA, RBF-GA, …
- Kombinace s **PCA** => „divoké“ kombinace typu PCA-RBF-GA (konzistentní výsledky)
- **Bayesovské neuronové sítě (BANN-GA):** tvořeny MLP formulovaném na základě Bayesovského přístupu, kdy jsou chápány jako parametrizovaný regresní model vytvářející pravděpodobnostní hypotézy o datech a trénovány s využitím hybridní metody Monte Carlo.
- **Hybridní síť MLP + RBF**: nejspolehlivější

Particle Swarm Optimalization Method

- V praxi použita v kombinaci s NN
- Navržená Kennedym a Eberhartem r. 1995
- Stochastický evoluční algoritmus používaný v široké míře pro optimalizaci
- Založena na socio-psychologických principech inspirujících se v inteligenci hejna. Inteligenci hejna tak vytváří kolektivní a individuální znalosti.
- **Konkrétní aplikace:** nejprve náhodně generováno řešení, následně členové hejna vstupují do interakce s ostatními a hledají řešení maximalizující míru vhodnosti. Zároveň si uchovávají sdílenou informaci o nejlepším dosaženém řešení, které jednotlivec dosud nalezl, ale i kterého bylo dosaženo v rámci celého hejna. Populace tak postupně konverguje k optimálnímu řešení.

MCMC (Markov Chain Monte Carlo)

- Jádro metod pro generování pseudonáhodných čísel z pravděpodobnostních rozdělení prostřednictvím Markovských řetězců.
- **Markovský řetězec** = sekvence náhodných veličin, u nichž rozdělení každého elementu závisí na hodnotě předchozího => hodnota každého náhodného vzorku závisí na hodnotě vzorku předchozího.
- **Gibbsovské samplování (Gibbs Sampling)**
- **Metropolis-Hastings algoritmus**
- **Výhoda:** nízké nároky na výpočetní kapacitu. Použitelné pro MI

Metoda propensitního skóre

- Používá řada SW řešení
- Pro imputaci spojité proměnné za předpokladu monotónního vzoru chování chybějících dat = pokud pro *i*-té pozorování *j*-tá proměnná obsahuje chybějící hodnotu, pak všechny další proměnné s vyšším indexem tohoto pozorování obsahují chybějící hodnotu též.
- Pro každou proměnnou obsahující chybějící hodnoty každému pozorování přiřazeno tzv. propensitní skóre jako odhad pravděpodobnosti, že pozorování je chybějící.
- Pozorování jsou poté sloučena podle propensitního skóre do předem daného počtu skupin (zpravidla 5).
- Následně je na ně uplatněna přibližná Bayesovská bootstrap imputace.

Metody založené na stromech

- **C4.5** (Ross Quinlan): lepší výsledky než Autoclass
- Rozhodovací strom je použit pro klasifikaci intervalů chybějících hodnot spojitých proměnných před použitím NN. **Rozšíření NN o C4.5** v obou případech vedlo k zvýšení spolehlivosti o 13%.
- Generování rozhodovacích pravidel **CLIP4** pro jednoduchou imputaci.
- **IIA (Incremental Imputation Algorithm)** jako aplikaci rozhodovacích stromů s FAST algoritmem založeném na dvoukrokovém dělení se zohledněním globální role prediktoru na lexikograficky setříděná pozorování (podle četnosti výskytu chybějících hodnot v atributech => postupováno od nejnižších četností).

Metody založené na stromech

- **CART** pro klasifikační a regresní stromy na imputaci chybějících dat ze senzorů bezdrátové sítě => označen za snadný nástroj rezistentní vůči odlehlým poz.
- **Forest Climbing** spočívající v konstrukci q různých klasifikačních stormů pro imputaci hodnot q atributů současně. Jedná se o případ, kdy jsou imputovány hodnoty proměnných v rámci dvou datových zdrojů, z nichž v prvním obsaženy jsou a v druhém chybí
- **RTII (Robust Tree-based Incremental Imputation)** umožňující doplňování chybějících hodnot pomocí klasifikačních a regresních stromů jak ze zdrojového souboru (tj. ze souboru obsahujícího chybějící data), tak z externího „dárcovského“ souboru s využitím techniky **AdaBoost** (kombinace výsledků několika jednodušších klasifikátorů)

Metody založené na stromech – test vnitřních algoritmů pro imputaci

- **C4.5**: pravděpodobnostní přístup, kdy po vytvoření větvení pomocí kritéria informačního zisku aplikovaného na úplné záznamy ve smyslu metody Pairwise jsou následně chybějící záznamy partišnovány podle vah představujících pravděpodobnost příslušnosti k danému listu => v rámci všech proměnných s výjimkou třídy.
- **CN2**: triviální jednoduchou imputací nejčetnější hodnoty.
- Při porovnání s 10-NN oba vnitřní algoritmy pohořely

Implementace v nástrojích

Komerční nástroje

- SAS Enterprise Miner : node Impute (stromy, průměr, median)
- SAS/STAT: PROC MI (EM, MCMC, regrese, diskriminační analýza, logreg, propensitní skóre), PROC MIANALYZE
- SOLAS: Propensitní skóre, hot deck, podmíněný průměr, diskriminační analýza, MNČ regrese, skupinové průměry, LVCF (Last Value Carried Forward)

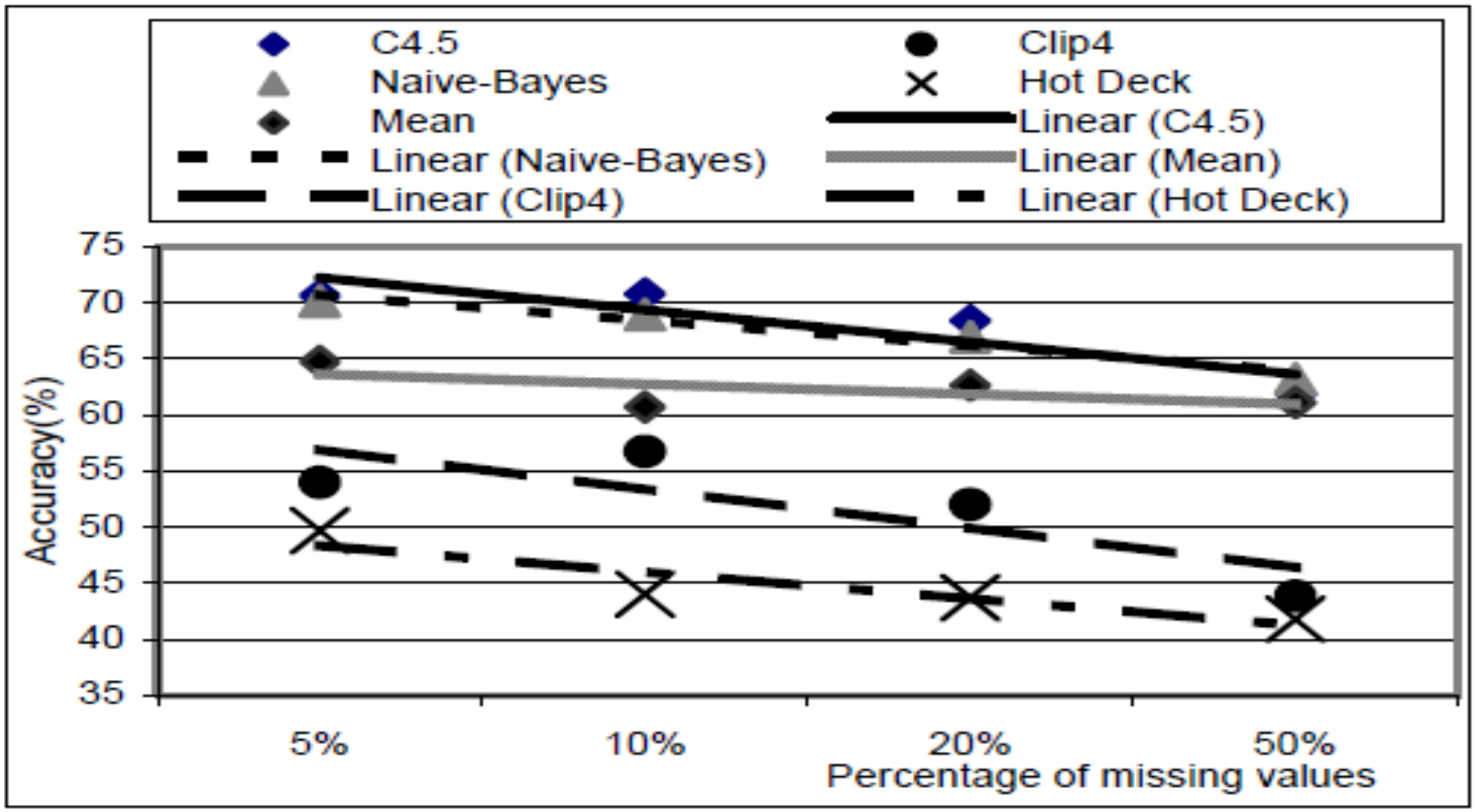
Nekomerční nástroje

- MICE (Multiple Imputation by Chained Equations)
  - knihovna pro nástroje R nebo S-Plus
  - samostatná instalace WinMICE.
  - podmíněný průměr, regrese, diskriminační analýza a MCMC, vlastní imputační funkce

Stávající benchmarky – příklad 1

- Celkem 7 datových souborů o různém počtu záznamů, různém počtu proměnných různých typů, při různé míře zastoupení booleovských atributů, náhodné generování chybějících záznamů
- **Metody ze čtyř skupin:**
  - pravděpodobnostní algoritmy (zástupcem byl zvolen Naïve Bayes)
  - rozhodovací stromy (zástupcem je C4.5)
  - rozhodovací pravidla (zástupcem je CLIP4)
  - metody učení bez učitele (doplnění průměrem, hot deck)
- **Metodika:**
  - náhodně generuje chybějící hodnoty od relativní četnosti 5% až po 50%
  - metody porovnávány na základě srovnání původních a doplněných hodnot
- **Závěry:**
  - ambiciozní záměr poskytnout strategie pro použití jednotlivých metod na konkrétní data.
  - některé závěry spíše triviálním ověření zřejmých vlastností metod

Výstupy benchmarku č. 1



Stávající benchmarky – příklad 2

- **Metody:**
  - MLR (Multinomial Logistic Regression) zobecněnou na *q* tříd
  - LD (Listwise Deletion)
  - MMSI (Mean Mode Single Imputation)
  - MNČ regrese
  - EM algoritmus.
- **Data:** International Software Benchmarking Standards Group.
- **Metodika:** podobná jako u předchozího uvedeného
- **Závěry:**
  - Efektivnost LD a MMSI při malém počtu chybějících hodnot (do 10%). Při vyšší míře neúplnosti dat byly tyto metody vyhodnoceny jako nevhodné.
  - Použití algoritmu EM se ukázalo jako velmi stabilní i při 30% míře výskytu chybějících hodnot.
  - MNČ regrese a MLR při 10% míře vykazovaly podobné výsledky jako ostatní algoritmy, při větším počtu chybějících hodnot již ostatní ve spolehlivosti předstihly. Při míře neúplnosti dat okolo 30% již vykazovala nejvyšší spolehlivost MLR.

Key Learnings

- Existuje řada metod, neexistuje však jejich obsáhlejší srovnání
- Závěry autorů dílčích benchmarků mohou být ovlivněny konkrétními daty
- Komplexní benchmark ve stylu STALOG nebo METAL zcela chybí
- Je vhodné odlišovat benchmarky pro jednotlivé typy mechanismů výskytu chybějících hodnot (MAR, MCAR, …)
- V případě MAR bude vhodnost použití metod podobná vhodnosti použití těchto metod pro predikci obecně
- Výzvou jsou další varianty strategie Hot deck / Cold deck
- Výzvou jsou kombinace různých metod (viz Marwala)

Návrh komplexního benchmarku

- K dispozici úplná datová matice obsahující kategoriální nominální, ordinální a spojité proměnné z různých předmětných oblastí (data klientů, adres, kontaktů, produktů, objektů jako např. vozidlo, … ale i ze zcela jiných domén jako jsou medicínská a meteorologická data, data Google, NASA, …)
- Vytvořena sada modelů kombinující různé typy vysvětlujících a vysvětlovaných proměnných => algoritmus pro výběr možných modelů s využitím chí-kvadrát, entropie, informačního zisku => míra vhodnosti algoritmu + potenciální spolehlivost modelu (porovnávána se spolehlivostí při následné imputaci)
- Pozorovaný vliv:
  - Vliv zvyšujícího se počtu chybějících hodnot
  - Vliv velikosti datového souboru
  - Vliv počtu chybějících hodnot v rámci více atributů současně

Postup benchmarku pro MAR

- Na úplných datech vytvořena sada modelů, přičemž je hledán nejvhodnější deskriptivní / prediktivní model pro daný typ dat => referenční míra spolehlivosti (horní mez)
- Pomocí náhodného výběru generován různý počet chybějících hodnot (náhodný výběr ID záznamu, náhodný výběr atributu)
- Přepočtení referenčního modelu => dolní mez spolehlivosti
- Aplikace metod pro imputaci
- Porovnání úspěšnosti dané metody na základě matice záměn
- Přepočtení referenčního modelu
- Porovnání úspěšnosti dané metody na základě přírůstku spolehlivosti referenčního modelu oproti dolní a horní mezi

Použitý software

- Příprava dat + náhodné generování missing hodnot: makro s využitím SAS BASE (funkce pro generování náhodných čísel nebo PROC SURVEYSELECT)
- SAS STAT: (regrese PROC MI pro propensitní skóre, PCA, diskriminační analýza, … obecně statistické procedury)
- SAS Enterprise Miner (stromy, NN)
- NN s využitím GA: ???

Očekávané problémy benchmarku + reálné aplikace výsledků

- Nejsou k dispozici rozsáhlá reálná data o klientech případně je nelze použít, pouze charakteristiky => nutnost vyvinout algoritmus rekonstruující populaci na základě jejich známých charakteristik
- Většina reálných datových souborů nebude MAR => online doplňování často nepřipadá v úvahu. Pozn: Pozor, datová kvalita = dodatečná informace

Postup – měl by být GANT

- Teoretická příprava – DONE
- Příprava hodnotících kritérií - DONE
- Kompletace datových zdrojů - DONE
- Příprava algoritmů
  - pro generování populace
  - pro náhodné generování chyb
  - pro určení potenciální spolehlivosti
- Příprava hodnotícího dashboardu benchmarku
- Postupné začleňování metod

Prezentace výsledků výzkumu

Data Quality CZ - portál věnující se tématu kvalitních dat

Základní informace | Topic mapa | Odkazy na zdroje | Výzkum | Články | Slovník pojmů | Nástroje | O autorovi

Základní informace o datové kvalitě

Úvod do problematiky

[29.1.2011] D. Pejčoch

Problematika kvality dat (angl. Data Quality) a na ní navazující problematika kvality informací (angl. Information Quality) se zejména s ohledem na enormní nárůst objemu zpracovávaných dat ve firmách dostala v posledním desetiletí do popředí zájmu. Firmy nejprve pochopily, že data pro ně v informační době znamenají velice cenné aktivum, účinnou zbraň v boji o udržení a získání nových zákazníků. S rostoucím uvědoměním si důležitosti vlastních dat se role kvality dat posunula od přednosti k nutnosti. Ten, kdo díky duplicitním záznamům v frustruje své zákazníky opakovaným oslovováním v marketingových kampaní dříve či později přijde. Ten, kdo zasílá svým zákazníkům permanentně chybn dopadne stejně.

Aktuality

TDWI World Conference Series - Chicago (6. - 10.6.2011)

2011-05-15 11:38:25

Pro ty, pro něž cesta do Chicaga není problém, představuje tato konference nabídku cca 40 kurzů na téma Business Intelligence, Data Warehousing, řadu prezentací (vč. Dave Wellse a Laury Reeves) mimo jiné též na téma řízení dat jako aktiv. ... (více zde)

Výzkum

Benchmarky metod

- Metody pro porovnávání řetězců
- Metody pro doplňování chybějících pozorování
- Metody pro verifikaci

http://www.dataquality.cz

Zdroje k metodám pro řešení neúplných dat

- Výpis celkem 35 zdrojů: http://www.dataquality.cz/index.php?ID=3
- PEJČOCH, D. *Metody řešení problematiky neúplných dat*[online]. 2011-01-13 Přednáška č. 4 v rámci Data Quality Tutorial. Dostupné pod odkazem: http://www.dataquality.cz/tutorial/tutorial_04.pdf.

Velmi rozsáhlé publikace:

- Tshilidzi Marwala "Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques" Information Science Reference. ISBN: 1605663360.
- Tan M, Tian GL and Ng KW (2008). Bayesian Missing Data Problems: EM, Data Augmentation and Non-iterative Computation. Chapman & Hall/CRC (Monographs on Statistics and Applied Probability), Boca Raton, USA.