Metody řešení problematiky neúplných dat

Ing. David Pejčoch, DiS.

Katedra informačního a znalostního inženýrství

Fakulta informatiky a statistiky

Vysoká škola ekonomická

Nám. W. Churchilla 4

130 00 Praha 3

Česká republika

Email: david@pejcoch.com

Klíčová slova: datová kvalita, chybějící hodnoty, nekompletní záznamy, úplnost, statistické metody, metody strojového učetní

Abstrakt:

Se stále rostoucí mírou používání dat, nejen pro účely zajištění běžného provozu firmy, ale i jako podkladů pro rozhodování, se do popředí zájmu dostala otázka kvality dat, které firma uchovává ve svých databázích. Míru kvality dat lze ztotožnit s mírou naplnění vlastností dat jako je např. jejich správnost, důvěryhodnost, úplnost, unikátnost, s ohledem na jejich potenciální využití. Tato práce se soustřeďuje na úplnost dat a to zejména z pohledu využití dat pro statistickou analýzu a získávání znalostí z databází. Klade si za cíl zmapovat možné příčiny vzniku neúplných dat, upozornit na možné negativní důsledky tohoto jevu a poskytnout stručný výčet metod použitelných pro odstranění těchto důsledků. V případě metod poukazuje na jejich přednosti a nedostatky a současně hledá náměty pro další výzkum v této oblasti.

1.1 Úvod

S enormním nárůstem dat zpracovávaných ve firmách i státním sektoru v posledních třech desetiletích se stala populární problematika řízení kvality těchto dat. Pokud bychom si položili otázku, co přesně se v tomto případě rozumí pod abstraktním pojmem „kvalita“, dospěli bychom k závěru, že se jedná o míru naplnění určitých vlastností, které od dat očekáváme. Jako příklad těchto vlastností lze uvést správnost dat (jak po syntaktické, tak po semantické stránce), důvěryhodnost, unikátnost nebo též jejich úplnost, kterou lze zjednodušeně vyjádřit jako míru vyplněnosti datových atributů. Má práce se soustřeďuje právě na tuto poslední uvedenou vlastnost dat. Jejím cílem je především zmapování přístupů pro řešení problému neúplných dat, které jsou popsány v odborné literatuře a byly uplatněny při praktickém řešení reálných problémů. Jednotlivé přístupy se tato práce pokouší porovnat zejména na základě publikovaných dílčích srovnání.

Na tomto místě považuji za nutné zmínit fakt, že problematika chybějících hodnot je sice ve světové literatuře popsána velice hojně, naproti tomu česká literatura tomuto tématu věnuje pozornost pouze malou. Zatímco v cizojazyčné literatuře lze narazit na velmi rozsáhlé monografie jako např. legendární a hojně citované [4], [6] nebo [10], v tuzemské literatuře jsem nalezl pouze dílčí zmínky (např. v publikacích [2] a [3]). Znamená to snad, že v prostředí České kotliny nepředstavují neúplná data žádný problém, anebo se jen v dnešní době obecné jazykové gramotnosti nevyplatí překládat téma precizně zpracované zahraničními autory?

Tuto práci lze z pohledu její struktury rozdělit do několika částí. V první části nejprve objasním, co přesně lze rozumět pod pojmem neúplná data. V této souvislosti zmíním mechanismy vedoucí k vzniku neúplných

dat a seznámím s jejich důsledky. V druhé části se pokusím nalézt optimální klasifikaci přístupů pro vypořádání se s neúplnými záznamy. V rámci navržené klasifikace uvedu příklady metod a jejich praktické aplikace. Poté uvedu některé příklady implementace metod v SW řešeních. Na závěr zmíním příklady studií porovnávajících některé z popsaných metod. Též se pokusím vymezit některé oblasti vhodné pro další výzkum, které dále zamýšlím rozvinout v rámci své disertační práce. Kde tedy začít? ... nejlépe od začátku.

1.1.1 Co jsou to neúplná data

V průběhu své praxe se asi každý analytik setkal s případem, kdy zdrojová data určená pro analýzu obsahovala chybějící pozorování a každý manažer přímých kampaní se setkal se situací, kdy potenciál k oslovení musel snižovat z důvodu chybějících kontaktních údajů. Zdá se tedy, že pojem „neúplná data“ lze snadno intuitivně vymezit. Zkusme ale toto zdání trochu zkomplikovat. Např. [1] uvádí dva různé typy neúplných dat: chybějící záznamy ve smyslu nevyplněných hodnot vybraných pozorování daného atributu a problematiku nekompletních záznamů související s cenzorováním dat. Druhý uvedený případ úzce souvisí s analýzou přežití. Jedná se o situaci, kdy je modelována doba do výskytu události (např. v životním pojištění úmrtí pojištěné osoby nebo v neživotním pojištění realizované storno smlouvy klientem). Může se stát, že k danému jevu u konkrétního subjektu během pozorování nedojde, přestože víme, že u něj jev v budoucnu zákonitě nastat musí. V tomto případě hovoříme o cenzorování zprava. Podobná situace nastává, pokud nemáme o subjektu k dispozici naměřená data z období před určitým datem (počátkem výzkumu, vznikem firmy, příchodem klienta od jiné firmy, ...) a k sledované události došlo před tímto datem (např. k pojistnému podvodu). Za takové situace hovoříme o cenzorování zleva. Zatímco v případě chybějících hodnot nemáme k dispozici pozorování, ačkoliv bychom jej mít měli, v případě cenzorovaných dat nemáme pozorování k dispozici, protože leží mimo časový úsek, během něhož probíhal sběr dat o daném subjektu.

1.1.2 Příčiny a mechanismy výskytu neúplných dat

Pokud se setkáme s nějakým nežádoucím jevem, je přirozené hledat jeho příčiny. Pro tyto účely je vhodné od sebe odlišit případy, kdy atribut / proměnná neobsahuje hodnotu korektně či nekorektně. K prvnímu případu dochází za situace, kdy pro danou entitu tato hodnota není k dispozici, např. ne každý člověk vlastní mobilní telefon nebo email, ne každá fyzická osoba je současně OSVČ a je u ní tudíž evidováno IČO. Jiná situace nastává v případě, kdy tato hodnota k dispozici potenciálně je. V tomto případě mohou nastat dvě varianty: Hodnota v danou chvíli není k dispozici v rámci žádného dostupného zdroje (např. v důsledku selhání senzoru měřícího data, z důvodu cenzorovaných dat, nezodpovězené nesrozumitelné otázky v rámci dotazníkového šetření, chybou při manuálním vkládání, ...) nebo v případě druhém hodnota je k dispozici a není v daném datovém zdroji z nějakého důvodu uvedena (lze ji doplnit anebo odvodit). Příčinou této druhé situace mohou být např. chyby v ETL[1] procesech přenášejících data mezi datovými zdroji. V rámci této skupiny příčin též uvažuji speciální případ, kdy hodnota dané proměnné byla v rámci přípravy dat záměrně vypuštěna vzhledem k její zjevné nekonzistenci s ostatními atributy / proměnnými (např. věk držitele vozidla = 5 let).

V odborné literatuře se často můžeme setkat s pojmem mechanismů výskutu chybějících hodnot. Tento pojem původně pochází od [26] a je zmíněn byť kuse téměř v každé publikaci na téma neúplných dat (namátkou zmiňme [4], [6], [10], [18], [20]). Z pohledu mechanismů výskytu [26] rozlišuje tři možné varianty. Prvním případem je situace, kdy chybějící hodnoty mají stejnou pravděpodobnost výskytu pro všechny záznamy. Záznamy s chybějícími hodnotami nejsou přitom nijak odlišitelné od těch bez chybějících hodnot. Za této situace hovoříme o tzv. MCAR (Missing Completely at Random) hodnotách. V případě tzv. MAR (Missing at Random) příčina chybějící hodnoty nezávisí na proměnné, v rámci níž se vyskytuje, nicméně může být závislá na jiných pozorovaných proměnných. Objevuje se též v případech, kdy se záznamy s chybějícími hodnotami liší od záznamů bez chybějících hodnot, ale tyto hodnoty lze na základě ostatních proměnných predikovat. Posledním případem, uvedeným v [26], jsou MNAR (Missing Not at Random) hodnoty, kdy 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. K těmto třem mechanismům [10] ještě doplňuje čtvrtý případ, kdy příčinou chybějící hodnoty je nemožnost jejího fyzického měření. Tento případ nazývá MBND (Missing By Natural Design).

[1] ETL = Extract Transform Load = proces pro přenos dat mezi datovými zdroji

Mechanismus výskytu chybějících hodnot určuje použitelné techniky pro jejich odstranění (viz níže). Pomocí t-testů nebo speciálního Littleova MCAR testu (viz např. v [24]) lze testovat hypotézu, že chybějící hodnota je MCAR oproti alternativní hypotéze, že se jedná o MAR. Jak ale uvádí [9] s odkazem na [6], bez dalších dodatečných informací nelze testovat hypotézu, že chybějící hodnota je MAR proti alternativní hypotéze, že se jedaná o NMAR.

1.1.3 Důsledky výskytu neúplných dat

Jaké může mít výskyt neúplných dat důsledky? Jak uvádí např. [1], součástí každé analýzy je fáze přípravy dat, v rámci níž je nutné mimo jiné vyřešit též přítomnost chybějících dat. Jak ukáži později, existují dvě hlavní stategie: chybějící data doplnit, anebo vypustit. V případě vynechání dat s chybějícími pozorováními potencitálně ztrácím informaci. Jelikož data zřídka chybějí náhodně, i v případě ponechání chybějících hodnot (a např. jejich zakódování) může dojít ke zkreslení informace. Zkreslené informace pak vedou k chybným znalostem a ty potom ke špatným rozhodnutím (viz [8]).

Na tomto místě pokládám za vhodné uvést, že v této práci se sice zabývám neúplnými daty výhradně z pohledu jejich využití pro statistickou analýzu a získávání znalostí z databází, nicméně chybějící data mohou představovat též vážný problém v běžném fungování firmy. Např. chybějící data emailů klientů mohou pro finanční instituci představovat oportunitní náklady z titulu potenciálního využití elektronické komunikace pro optimalizaci provozní korespondence. Chybějící místo narození klienta není v souladu s nařízením ČNB pro AML[2] a může vést k sankcím z její strany. Neúplné účetní záznamy nejsou v souladu s platnými právními normami a pokud mají za následek neprůkaznost účetnictví, vystavují firmu riziku vysokých pokut. Jak již ale bylo zmíněno, tato práce od těchto důsledků neúplných dat abstrahuje.

1.2 Způsoby řešení problému neúplných dat

Nejlepším způsobem, jak se vypořádat s neúplnými záznamy je bezpochyby procesně zabránit jejich vzniku, ale zůstaňme nohama na zemi. Následující odstavce poskytují stručný přehled metod, které je možné použít pro nahrazení nebo jiné vypořádání se s chybějícími hodnotami.

1.2.1 Nalezení optimální klasifikace přístupů

Pro lepší porozumění následujícím odstavcům považuji za vhodné na tomto místě definovat pojem imputace (imputing) tak, jak je zmíněn např. v [6]. Imputace je obecně používaný termín pro doplnění chybějících záznamů o přijatelné hodnoty. Doplnění probíhá výběrem z jednoho nebo více kanditátů.

Jak už tomu bývá zvykem, můžeme se v odborné literatuře setkat s mnoha dílčími klasifikacemi přístupů k neúplným datům. Např. [25] uvažuje tři typy metod / přístupů: 1) ignorování / vynechání záznamů, 2) odhad parametrů a doplnění chybějících hodnot, 3) imputing (imputace). [21] naproti tomu uvažuje v souladu s výše uvedenými mechanismy vzniku 1) metody založené na atributu obsahujícím chybějící hodnoty a 2) metody založené na ostatních atributech.

Veskrze se většina zdrojů (jako příklad uveďme [9],[17], [20] nebo [24]) shodují na dalším členěním metod pro imputaci na SI (Single Imputation) a MI (Multiple Imputation). V rámci SI je každá chybějící hodnota doplňována pouze jednou hodnotou, zatímco v případě MI je pro každou chybějící hodnotu generováno několik alternativních variant. Tento proces probíhá ve třech krocích (viz např. [9]): 1) generování množiny $m > 1$ hodnot, 2) analýza m dílčích datových souborů vytvořených z původního datového souboru s využitím metod pro úplné záznamy, 3) kombinace výsledků m analýz pro volbu doplňované hodnoty.

[19] člení jak metody SI tak MI dále na: 1) řízené daty, 2) založené na modelu a 3) založené na strojovém učení. [4] rozlišuje procedury založené na kompletních záznamech, procedury založené na imputaci, vážené procedury zohledňující rozložení v populaci a přístupy založené na modelu.

Z pohledu [14] je pro členění metod podstatné, zda se jedná o metody učení s učitelem (supervised learning), anebo o metody učení bez učitele (unsupervised learning). V rámci metod bez učitele uvažuje tzv. jednoduché metody založené na smazání / ignorování záznamů a statistické metody založené na odhadu

[2] AML = Anti money loundering. V ČR se úprava řídí Vyhláškou č. 281/2008 Sb. o některých požadavcích na systém vnitřních zásad, postupů a kontrolních opatření proti legalizaci výnosu z trestné činnosti a financování terorismu.

parametrů. V rámci metod učení s učitelem rozlišuje pravděpodobnostní algoritmy, rozhodovací stromy a rozhodovací pravidla.

Nejkomplexněji ke klasifikaci metod pro řešení neúplných dat přistupuje [18]. Na nejvyšší úrovni od sebe odlišuje následující tři přístupy: 1) ignorování / smazání pozorování, 2) maximální využití dostupných dat (viz metoda Pairwise uvedená níže), 3) procedury založené na imputaci. Poslední zmíněnou skupinu dále člení na přístupy nezalożené na modelu a přístupy založené na modelu. Model dle jeho klasifikace může být uvažován buď jako implicitní (založený na implicitních předpokladech jako je např. podobnost mezi pozorováními), anebo explicitní (založený na formálním statistickém modelu). V rámci implicitních modelů zmiňuje mimo jiné podskupinu faktoriálních technik. Explicitní modely dále člení na parametrické a neparametrické.

Pokud vezmeme jako výchozí nejpropracovanější přístup [18], na základě syntézy s ostatními přístupy můžeme dospět např. k níže uvedenému univerzálnímu členění, které jsem si pro úplnost dovolil doplnit ještě o zvláštní skupinu metod, která byla ve všech uvedených klasifikacích opomenuta.Na jednotlivých úrovních jsou již doplněny příklady konkrétních názvů metod, které budou vysvětleny a zhodnoceny v dalším textu.

1. Ponechání status quo
    a. Ignorování / smazání pozorování (stepwise)
    b. Maximální využití dostupných dat (pairwise)
2. Databázové techniky (join, lookup, fuzzy join)
3. Procedury založené na imputaci
    a. Přístupy nezalożené na modelu (nepodmíněný průměr, Buckova metoda, midrange, ...)
    b. Přístupy založené na modelu
        i. Implicitní model
            1. Faktoriální techniky (PCA, vícerorměrná korespondenční analýza)
            2. Metody založené na podobnosti (hot-deck, cold-deck, nearest neighbour)
        ii. Explicitní model
            1. Parametrické modely (jednotlivé konkrétní příklady GLM, Naïve Bayes, neuronové sítě, metody pro vícenásobnou imputaci, EM algoritmus)
            2. Neparametrické modely (neparametrická regrese, metody založené na stromech)

V odborné literatuře (např. [17]) se můžeme v souvislosti s prediktivními modely setkat s pojmem metod strojového učení. V jiných zdrojích lze nalézt pojem statistické metody. Teoreticky by se tak nabízelo členění metod pro imputaci založených na modelu na metody statistické a metody strojového učení. Toto členění, ač možná správné, mohlo by být poněkud kontroverzní, proto jsem se rozhodl od něj upustit. Byť bychom metody strojového učení chápali pouze jako metody učení s učitelem (supervised learning), pak bychom nutně dospěli k závěru, že např. obecný regresní lineární model použitý pro extrapolaci (predikci) je metodou strojového učení, což je zcela jistě tvrzení, kterým bychom řadu pravověrných statistiků příliš nepotěšili a tato klasifikace by se pro ně mohla stát nepřijatelnou.

1.2.2 Ponechání status quo

Tento přístup, v odborné literatuře označovaný jako tradiční, spočívá buď ve vynechání pozorování, jeho ignorování, anebo překódování. Přístup Listwise (LD, Listwise Deletion) uvažuje vynechání všech pozorování s chybějícími hodnotami bez ohledu na to, zda je atribut s chybějícími hodnotami v dané analýze použit. Umírněnější přístup Pairwise uvažuje vynechání pouze těch pozorování, která souvisejí s aktuální prováděnou analýzou. Maximalizuje tak použití dostupných dat. Jiným způsobem řešení je překódování chybějící hodnoty neutrální kategorií „nevím“, „N/A“, „?“, apod.

Pro tuto skupinu metod je společnou vlastností fakt, že vždy vedou ke ztrátě informace a rovněž jejich aplikovatelnost pouze na MCAR. Odborná literatura se shoduje, že aplikaci těchto metod je vhodné provádět pouze při nízkém relativním počtu chybějích hodnot. [4] na toto konto uvádí maximální hranici 5% relativní četnosti u dané proměnné. Všechny uvedené metody bývají standardní součástí statistických nástrojů.

1.2.3 Databázové techniky

Přístup využívající databázových technik jsem pro úplnost zařadil jako možnost řešení situace MNAR, obecně situace, kdy data sice v daném datovém zdroji k dispozici nejsou, ale jsou k dispozici v rámci jiného

dostupného zdroje, veřejného registru, číselníku či klasifikace. Mezi tyto techniky řadím navázání externího zdroje přes klasický JOIN / MERGE, LOOKUP v případech, kdy je k dispozici jedinečný klíč pro párování, resp. přibližného JOIN (fuzzy join) v případě, kdy jedinečný identifikátor záznamu v některém z datových zdrojů chybí. Obecnou vlastností těchto metod je fakt, že pokud jsou odpovídající externí data k dispozici, nedochází při jejich použití ke ztrátě informace, spolehlivost doplnění hodnot je maximální.

1.2.4 Metody imputace nezaložené na modelu

Do této skupiny metod patří metody doplňující jednu konkrétní hodnotu pro každou chybějící, anebo pro všechny chybějící v rámci jedné proměnné. Jako příklad tohoto typu metod lze uvést náhradu chybějících hodnot nepodmíněným průměrem hodnot proměnné (SMI, Sample Mean Imputation) (viz např. [2], [10], [17],[21], [24],[25]) , resp. vhodnějším mediánem, v případě kategoriálních proměnných doplnění modální kategorií (viz např. [21], [25]). Další možnost představuje doplnění středem rozpětí (midrange), spočteným jako aritmetický průměr maximální a minimální hodnoty dané proměnné. Nevýhodou všech dosud zmíněných metod je doplnění jediné hodnoty jako substitutu všech chybějících hodnot v rámci dané proměnné. Tím dochází k deformaci odhadu parametrů rozdělení, které se při větším počtu chybějících hodnot mimo jiné projevuje v rozdělení četností efektem, jenž by bylo možné lidově nazvat „Čechové na Řípu“.

Lepším způsobem imputace je použití podmíněného průměru, v odborné literatuře též uváděným pod názvem Buckova metoda (viz např. [4]). Spočívá v doplnění více průměrných hodnot podmíněných hodnotami ostatních proměnných. Jedná se vpodstatě o lineární regresi. Je tedy diskutabilní, zda tuto metodu neklasifikovat do přístupů založených na modelu (přestože [18] ji řadí do této skupiny). Podle [4] vede aplikace Buckovy metody ke konzistentním odhadům u MCAR a MAR. V případě MAR je ovšem nutné přijmout dodatečný předpoklad, že skutečnost výskytu chybějících hodnot nezávisí na ostatních proměnných. V takovém případě by aplikace Buckovy metody vedla k chybné imputaci. Stejně tak jako předchozí uvedené metody, doplnění podmíněným průměrem podhodnocuje variabilitu dané proměnné. Její použití má ještě jedno zřejmé omezení: proměnná s chybějícími hodnotami musí být spojitého charakteru.

Další metodou spadající do této kategorie je doplnění všech přípustných hodnot, kterých může daná proměnná nabývat ([21]), lépe na základě relativních četností jednotlivých hodnot ([9]). I tento přístup však vede ke zkreslení.

V případě longitudiálních dat, je podle [24] možné použít pro doplnění chybějícího pozorování v rámci časové řady poslední předchozí pozorování. Nicméně k lepšímu doplnění by zřejmě došlo, pokud bychom použili průměr z předchozího a následného pozorování nebo klouzavý průměr z více hodnot.

1.2.5 Metody imputace založené na modelu

1.2.5.1 Implicitní modely

Tato skupina metod vychází z implicitních vztahů mezi daty, jako je např. podobnost mezi jednotlivými pozorováními.

1.2.5.1.1 Hot Deck

Poměrně známou metodou z této oblasti je hot-deck imputace (HDSI, Hot Deck Single Imputation), uvedená např. v [10], [18], [24], založená na doplnění shodné hodnoty, jaká se vyskytuje u podobných reprezentantů. Záznamy jsou nejprve rozděleny do jednotlivých tříd s využitím technik jako je např. shlukování podle nejbližšího souseda. Chybějící hodnota je poté doplněna na základě příslušnosti pozorování k dané třídě. Výhodou této metody je to, že k odhadu nepotřebuje žádné složité předpoklady. Je též oceňována pro svou jednoduchost a tím i malé nároky na strojový čas. Na druhou stranu, její aplikace umožňuje značnou míru subjektivity při určování podobnosti mezi záznamy. Na reálných datech může dojít k situaci, kdy jedno pozorování může být podobné více třídám. V takovém případě je kandidát vybrán náhodně nebo je aplikován průměr z možných kandidátů.

[18] do této skupiny metod ještě řadí kombinované přístupy, kdy je např. metoda HDSI použita současně s lineární regresí. Tímto způsobem zřejmě uvažuje řešení výše zmíněné situace více alternativních kandidátů pro doplnění a k tomuto účelu používá jejich lineární kombinaci. Možnost využití různých metod v rámci hot-deck naznačuje, že se jedná spíše o strategii doplňování, než o jednu konkrétní metodu.

1.2.5.1.2 Cold Deck

Podobným způsobem jako u metody HDSI je postupováno u metody cold-deck (CDSI, Cold Deck Single Imputation) s tím rozdílem, že v tomto případě probíhá výběr kandidátů z jiného datového zdroje než toho obsahujícího chybějící pozorování (viz např. [10], [18], [25]). Spíše se jedná o strategii doplňování z externích zdrojů zobecňující případy jako je např. Data Fusion popisovaný [18], kdy doplňování probíhá z více jak jednoho externího zdroje.

1.2.5.1.3 k-Nearest Neighbour

V rámci přístupu hot-deck byla zmíněna jako jedna z možností hledání podobnosti mezi záznamy metoda nejbližšího souseda. Její přímou aplikaci na doplňování chybějícího pozorování (k-NNSI, k-Nearest Neighbour Single Imputation) popisuje [16] a zmiňuje ji též např. [18], [25] nebo [28]. Algoritmus spočívá v hledání pozorování nejbližšího tomu s chybějící hodnotou. Při takovém postupu dochází k velkému množství operací porovnávání, což vede ke značné náročnosti metody na strojový čas. Z toho důvodu se v praxi používají optimalizační techniky jako je např. obdoba databázového B-tree indexu nazývaná M-tree, která dobu prohledávání značně snižuje. [28] ve své studii ukazuje, že k-NNSI představuje spolehlivější alternativu oproti klasickému doplňování průměrem. Problematickým může být aplikace této metody na kategoriální proměnné, kde může činit potíže subjektivita stanovení nejbližší kategorie.

1.2.5.1.4 Faktoriální metody

Mezi faktoriální metody [18] řadí metodu hlavních komponent (PCA) a vícenásobnou korespondenční analýzu (MCA). Metoda hlavních komponent je standardně používanou metodou v rámci přípravy dat, umožňující redukci dimenzionality úlohy tím, že některé původní proměnné nahradí jejich ortogonální lineární kombinací, nazývanou hlavní komponenty. V rámci imputování jsem se navzdory [18] nesetkal s její přímou aplikací pro doplnění chybějících hodnot. [10] ji ale zmiňuje jako techniku, která umožňuje zrychlení aplikace neuronových sítí pro imputaci svým použitím pro původní účel, tj. redukci dimenzionality a tím snížení potřebného počtu neuronů ve skryté vrstvě. Výstupem korespondenční analýzy (viz např. [31]) jsou latentní faktory vysvětlující četnosti v kontingenční tabulce popisující závislost dvou a více proměnných. Představuje alternativu k faktorové analýze a PCA, neboť latentní faktory si lze přestavit jako osy redukovaného souřadnicového systému. Pro tuto metodu platí to, co již bylo zmíněno v souvislosti s PCA, tj. navzdory referenci [18] jsem nenalezl žádnou konkrétní aplikaci MCA pro doplňování hodnot, ale lze předpokládat, že pomocí ní lze optimalizovat použití jiných metod tak, jako tomu bylou PCA.

1.2.5.1.5 Shluková analýza

[17] popisuje přístup nazvaný DCI (Dynamic Clustering Imputation) založený na fuzzy shlukové analýze. 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. V případě kategoriálních atributů je místo chybějící hodnoty doplněna hodnota ze shluku s nejvíce přirozenými sousedy. V případě spojitého atributu je hodnota odhadnuta na základě všech hodnot shluku s nejvíce přirozenými sousedy. Jedná se v podstatě o další variantu imputovací strategie hot-deck. Při empirickém srovnání spolehlivosti této metody s doplňováním pomocí nepodmíněného průměru, podmíněného průměru a regresí dosahovala tato metoda až o 20% lepších výsledků oproti svým konkurentům.

1.2.5.1.6 Přibližné množiny (Rough Sets)

[10] popisuje hot-deck techniku pracující s tzv. přibližnými množinami (Rough Sets). Přibližné množiny jsou podobné fuzzy množinám v tom smyslu, že obě teorie pracují s pojmem vágnosti. Zatím co fuzzy množiny pracují s mírou, s níž hodnota přísluší do dané množiny, přibližné množiny aproximují přesnou množinu pomocí dvojice jiných množin představujících horní a dolní odhad původní množiny. Dolní odhad je kompletní množina objektů, které mohou být positivně klasifikovány jako prvky množiny X. Horní odhad je kompletní množina objektů, které nemohou být pozitivně klasifikovány jako prvky komplementu množiny X, tzn. mohou být prvky množiny X. Algoritmus pracuje s tzv. informační tabulkou vystavěnou na předpokladu, že data mají tvar rozhodovacích pravidel ve tvaru podmínka => rozhodnutí. Záznamy v doplňovaném datovém souboru jsou nejprve rozděleny podle hodnot rozhodnutí (třídy) a poté jsou mezi nimi hledány podobnosti na základě přibližných množin. Myšlenka aplikace této metody je založena na

očekávání, že v databázi existují stejné nebo podobné záznamy. V takovém případě může být snazší nalézt podobnosti, než konstruovat složitý model k doplnění. Empirické otestování ukázalo, že použití přibližných množin vede za situace opakujících se záznamů v databázi až k 99% spolehlivosti.

1.2.5.2 Explicitní modely

Explicitní modely jsou podle [18] metody založené na formálních statistických modelech predikovaného rozdělení chybějících dat. Na základě prostudování relevantních zdrojů jsem dospěl k závěru, že problematika imputace založené na explicitním modelu je do značné míry totožná s problematikou prediktivních modelů používaných na úplné záznamy. Jak ukáži dále, kromě prediktivních modelů lze též úspěšně použít některé simulační metody. Explicitní modely lze v souladu s [18] dále členit na parametrické a neparametrické. Představme si nyní některé zástupce těchto skupin s tím, že začneme těmi parametrickými.

1.2.5.2.1 Regresní analýza

V odborné literatuře (např. [1], [2], [7], [10], [18], [21] nebo [24]) jsou hojně zmíněny příklady aplikace lineární regrese jako způsobu pro doplňování chybějících hodnot, kdy jsou chybějící hodnoty predikovány pomocí modelu vytvořeného na úplných datech na základě ostatních proměnných. Její použití se samozřejmě (stejně tak jako použití jiné prediktivní metody) odvíjí od existence závislosti mezi prediktory a doplňovanou proměnnou. Použití regrese vede k nežádoucímu snížení variability a kovariance dat. S tímto nedostatkem se pokouší vypořádat stochastická regresní imputace (viz např. [24]), v rámci níž jsou doplňované hodnoty korigovány o rezidua, tj. odchylky reálných pozorování od proložené regresní přímky. Lineární regresi je též možné použít pouze pro imputaci spojitých proměnných. Pokud bychom lpěli na předpokladech zobecněného lineárního modelu, byla by možná její aplikace též pouze v případě spojitých vysvětlujících proměnných.

Stejně tak, jako lze lineární regresi použít pro imputaci spojité proměnné, lze logistickou regresi použít pro imputaci proměnné binární (MLRI, Multinomial Logistic Regression Imputation). Zobecněnou variantu MLRI, umožňující klasifikaci do q tříd, použil [29] pro porovnání s MMSI (Mean Mode Single Imputation), EM, regresní analýzou používající klasické MNČ (OLS)[3] kriterium s využitím Stepwise[4] metody a LD (Listwise Deletion). Při větším počtu chybějících hodnot z tohoto porovnání vychází vítězně MLRI. Detaily srovnání uvádím níže v kapitole Benchmark metod. Zůstává k diskusi, do jaké míry je korektní použití MNČ regrese pro predikci hodnot kategoriální proměnné. Jedním z kritérií použití klasické MNČ je též neexistence multikolinearity[5], které nemuselo být v tomto případě splněno.

Dalším příklad použití logistické regrese představuje její kombinace s loglineárním modelem použitá pro imputaci chybějících fakt v OLAP[6] kostce. Pomocí logistické regrese [15] predikuje výskyt chybně nevyplněných hodnot v tabulce faktů a doplňuje je na základě odhadů pomocí loglineárního modelu.

Uvedl jsem zde příklad využití lineární regrese pro účely imputace s tím, že jsem zmínil omezení v podobě spojité proměnné, kterou je možné touto metodou výhradně imputovat a spojitých vysvětlujících proměnných pomocí kterých je model pro imputaci vytvořen. Toto omezení zmizí, pokud uvažujeme obecný lineární model (GLM) popsaný např. v [30]. Při jeho použití nezáleží na tom, zda jsou vysvětlující nebo vysvětlované proměnné kategoriální či spojité. Jako specifický případ GLM umožňující predikci kategoriální vysvětlované proměnné uvažuji lineární diskriminační analýzu (LDA), která byla jako klasifikační metoda doporučena projektem STATLOG porovanávajícího jednotlivé metody jako vhodná pro rozsáhlé datové soubory (viz [5]).

1.2.5.2.2 Bayesovský přístup

Použití metody Naïve Bayes jako způsobu imputace kategoriálních proměnných popisuje [14]. Metoda je založena na analýze vztahů mezi každou nezávisle proměnnou a závisle proměnnou s cílem stanovit podmíněnou pravděpodobnost pro každý vztah. [14] tuto metodu hodnotí jako velmi efektivní už z toho

[3] MNČ = Metoda nejmenších čtverců, OLS = Ordinary Least Squares

[4] Stepwise = metoda pro zařazování vysvětlujících proměnných do modelu

[5] Multikolinearita = jedno z kritérií použití MNČ, vyžadující plnou hodnot matice vysvětlujících proměnných

[6] OLAP = Online Analytical Processing

důvodu, že pro vytvoření modelu potřebuje pouze jeden průchod daty. Též [3] hodnotí tuto metodu vhodnou pro použití na velkých datových souborech. V porovnání publikovaném v [14] a podrobněji rozebraném níže v kapitole Benchmark metod byl Naïve Bayes klasifikátor přes relativně nižší průměrnou spolehlivost v porovnání s ostatními srovnávanými metodami označen za metodu podávající stabilní výsledky při různém počtu chybějících hodnot, což může být v praktickém využití též důležitý faktor.

Pokud bychom zůstali u Bayesovského téma, nelze opomenout popis aplikace Bayesovských sítí v [22] a [18]. Bayesovské sítě (též kauzální sítě) představují pravděpodobnostní modely využívající grafickou reprezentaci formou acyklického grafu. Skládají se ze dvou typů uzlů: zdrojových uzlů s nepodmíněným rozdělením pravděpodobnosti a ostatních uzlů s podmíněným rozdělením pravděpodobnosti. Zadáním některých hodnot uzlů získáme výpočet podmíněné distribuce v dalších uzlech. Imputace s využitím této metody probíhá na základě jednoho nebo více modelů simulujících posteriorní rozdělení chybějících hodnot. [22] uvádí algoritmus IMPUTE BN(S;N) jako konkrétní metodu pro doplňování chybějících hodnot s využitím Bayesovských sítí. [18] upozorňuje na zásadní nevýhodu této metody, když ji označuje jako velice komplexní a náročnou na čas.

1.2.5.2.3 Maximalizace věrohodnosti

Další skupinou metod, které [18] řadí do metod parametrických jsou metody maximalizace spolehlivosti. Jako jejich příklad lze uvést v odborné literatuře hojně rozšířený algoritmus EM (Expectation Maximization). Při použití pro imputaci nejprve dochází k počátečnímu nastavení parametrů a doplnění hodnot takto vytvořeným modelem (E-krok), poté následuje modifikace parametrů modelu (M-krok) s cílem nalézt parametry maximalizující celkovou věrohodnost. V souvislosti s využitím EM algoritmu pro imputaci odkazuji zejména na rozsáholou (cca 300 stran) monografii [11] zpracovávající toto téma do hloubky. Z pohledu praktického použití tohoto algoritmu lze odkázat na studii[16] porovnávající empiricky kritérium spolehlivosti celkem 7 metod: LD (Listwise Deletion), EMSI (Expectation Maximization Single Imputation), EMMI (Expectation Maximization Multiple Imputation), kNNSI (k-Nearest Neighbor Simple Imputation), MMSI, FC (Fractional Cases: strategie využívaná algoritmem C4.5), SVS a MIFC[7]. Z tohoto srovnání vychází vítězně EMMI následovaná FC a EMSI.

1.2.5.2.4 Neuronové sítě

Pro imputaci lze rovněž s úspěchem využít neuronové sítě. Zcela jistě nejobsáhlejší publikací zpracovávající toto téma je [10]. Tato knižní monografie porovnává využití neuronových sítí s několika alternativními metodami a současně hledá kombinace s dalšími metodami umožňující jejich optimalizaci. V mnoha popisovaných případech jsou autoasociativní neuronové sítě za tímto účlem použity v kombinaci s genetickými algoritmy.

V prvním uváděném příkladu je porovnáno použití vícevrstevného perceptronu (MLP) s radiální bazickou funkcí (RBF). MLP užívá 3 a více vrstev neuronů s nelineární aktivační funkcí. Umí si poradit i s daty nelineárně separabilními nebo separabilními pomocí nadroviny. RBF se typicky skládá z jedné skryté vrstvy neuronů, jejichž aktivační funce je vybírána z třídy tzv. bazických funkcí. Oproti zpětné propagaci používané MLP představuje použití RBF některé výhody: rychlejší trénování sítě a menší náchylnost k problémům s nestacionárními vstupy. Genetické algoritmy byly v obou případech použity pro minimalizaci Euklidovské normy chybové funkce (MDEEF, Missing Data Estimation Error Function) spočtené jako čtverec rozdílu výstupního vektoru naučené neuronové sítě a vstupního vektoru. Výsledná kombinace metod byla označena jako AAMLP-GA, resp. AARBF-GA. Při empirickém porovnání na datech symptomů HIV bylo zjištěno, že v obou případech se zvětšujícím počtem chybějících hodnot nedochází k signifikantním změnám správnosti doplňovaných hodnot. Relativně lepších výsledků bylo dosaženo v případě AAMLP-GA, ale rozdíl nebyl signifikatní. [10] doporučuje ověření, zda výsledek experimentu nebyl ovlivněn konkrétními daty.

V dalším uváděném příkladu [10] porovnává použití Bayesovských neuronových sítí v kombinaci s genetickým algoritmem (BANN-GA), hybridní sítě složené z RBF a MLP a kombinace PCA s autoasociativní neuronovou sítí (PCA-NN). Bayesovské neuronové sítě jsou tvořeny MLP formulovaném na základě Bayesovského přístupu, kdy jsou chápány jako parametrizovaný regresní model vytvářející

[7] Bohužel [16] neuvádí dodatečný popis k metodám SVS a MIFC, ani v souvislosti s nimi neodkazuje na žádný další zdroj. Lze tudíž jen hádat, zda SVS je např. metoda založená na Support Vector Machine a zda MIFC je variantou pro vícenásobnou imputaci pomocí algoritmu FC.

pravděpodobnostní hypotézy o datech a trénovány s využitím hybridní metody Monte Carlo. GA byl opět použit pro minimalizaci chyby. U PCA-NN byla metoda hlavních komponent použita pro redukci dimenzionality před aplikací NN vedoucí ke snížení počtu neuronů. Po aplikaci neuronové sítě byla použita inverzní PCA pro rekonstrukci původních dat. Za konstrukcí hybridní autoasociativní neuronové sítě stála myšlenka kombinace toho nejlepšího z obou architektur. Tato architektura vedla k efektivnějšímu podchycení komplexních nelineárních vztahů mezi proměnnými. Sítě byly trénovány pomocí metody maximální spolehlivosti. Expertimentální srovnání proběhlo opět na datech popisujících příčiny HIV. V případě jedné odhadované chybějící hodnoty se jako nejspolehlivější ukázala hybridní neuronová síť. V případě více odhadovaných chybějících hodnot se jako nejspolehlivější ukázala kombinace PCA a neuronových sítí. Současně vedla s přibývajícím počtem imputovaných hodnot k nejkonzistentnějším odhadům.

Další experiment porovnával kombinaci smíšených Gaussovských modelů a EM algoritmu (GMM-EM) s kombinací autoasociativních neuronových sítí a evolučního algoritmu v české literatuře popisovaného jako optimalizace pomocí hejna částic (Particle Swarm Optimalization Method)(viz např. [34]). Vzniklou kombinaci AANN a evolučního algoritmu označuje [10] AANN-PSO. Metoda PSO, navržená Kennedym a Eberhartem r. 1995, je stochastický evoluční algoritmus používaný v široké míře pro optimalizaci. Je založena na socio-psychologických principech inspirujících se v inteligenci hejna, v němž se jednotlivci učí řešit problémy komunikací a interakcí s ostatními jednotlivci, tj. jednak odvozují směr svého dalšího pohybu z polohy ostatních členů hejna a jednak se vrací za potravou do míst, kde se v minulosti hojně vyskytovala. Inteligenci hejna tak vytváří kolektivní a individuální znalosti. Člen hejna balancuje mezi těmito dvěma protipóly. Konkrétní aplikace probíhá tak, že je 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 kovnerguje k optimálnímu řešení. Porovnání metod bylo realizováno celkem na třech různých datových souborech z elektrotechnického průmyslu, těžařského průmyslu a na již zmíněných datech diagnostiky HIV. GMM-EM se ukázala jako vhodnější za situace, kdy vzájemná závislost mezi vstupními proměnnými byla nízká nebo žádná, v opačném případě bodovala AANN-PSO.

[10] neponechává stranou testování ani relativně mladou a nadějnou metodu Support Vector Machine (SVM), konkrétně její nadstavbu Support Vector Regression (SVR). SVR v kombinaci s GA porovnává s autoasociativní neuronovou sítí (ANN) optimalizovanou pomocí GA a již zmiňovanou kombinací PCA-ANN-GA. Algoritmus SVM je založen na převedení úlohy klasifikace do nelineárně separovatelných tříd na úlohu klasifikace do tříd lineárně separovatelných. V prostoru transformovaných atributů se poté snažíme nalézt dělící nadrovinu, která má co největší vzdálenost od příkladů různých tříd. Jak uvádí [3], rozhodujícími jsou pro nalezení příklady ležící nejblíže hranici mezi třídami, tzv. podpůrné vektory. V tomto smyslu je myšlenka SVR podobná: pro vytvoření modelu jsou rozhodující pozorování ležící v určitém pásu od prokládané přímky. V rámci porovnávání ANN-GA, PCA-ANN-GA a SVR-GA byla použita modifikace GA z předchozích testů, tzv. Cultural GA. Tato metoda aplikuje tzv. „prostor důvěry“ (belief space) rozdělený podle různých domén znalostí, které může populace mít o prohledávaném prostoru. Příkladem takových znalostí mohou být normativní znalosti v podobě omezení. Prostor důvěry je revidován po každé iteraci s pomocí nejlepších jedinců v populaci. CGA je aplikován pro nalezení takového vstupu modelu, který povede k nejpřesnější doplňované hodnotě. Empirické porovnání třech zmíněných metod ukázalo značnou spolehlivost ANN-GA (až 97,4%). V případě použití PCA-NNA-GA vedlo snížení dimenze z 11 na 10 proměnných ke zhoršení spolehlivosti. Aplikace SVR-GA prokázala značnou náročnost na strojový čas této metody. Metoda navíc poskytovala různorodé výsledky z pohledu spolehlivosti u jednotlivých proměnných. Dobře fungovala u doplňování spojitých proměnných avšak hůře u doplňování proměnných kategoriálních. Zajímavé bylo zjištění, že tato metoda fungovala dobře tam, kde ostatní selhávaly a naopak. Tento závěr svádí k zamyšlení nad vytvořením hybridních přístupů kombinujících různé metody.

1.2.5.2.5 Metody pro vícenásobnou imputaci - MCMC (Markov chain Monte Carlo)

Metoda MCMC tvoří jádro metod pro generování pseudonáhodných čísel z pravděpodobnostních rozdělení prostřednictvím Markovských řetězců. Jak uvátí např. [6], Markovský řetězec je sekvence náhodných veličin, u nichž rozdělení každého elementu závisí na hodnotě předchozího. Tudíž i hodnota každého náhodného vzorku závisí na hodnotě vzorku předchozího. Dvěmi nejpopulárnějšími metodami pro generování pseudonáhodných čísel pro účely MCMC jsou Gibbsovské samplování (Gibbs Sampling) a

Metropolis-Hastings algoritmus. Gibbsovské samplování generuje hodnotu z podmíněného rozdělení každé komponenty vícerozměrné náhodné veličiny, ponechávaje ostatní komponenty v cyklické modifikaci. Tím dostáváme řetězec hodnot, konvergující ke vzorku z posteriorního rozdělení. Algoritmus Metropolis-Hastings poskytuje výběr z pravděpodobnostního rozdělení za účelem aproximace rozdělení chybějící hodnoty a posléze přijetí nebo zamítnutí simulované hodnoty s konkrétní pravděpodobností. Výhodou těchto přístupů jsou především jejich nízké nároky na výpočetní kapacitu. Jak uvádí [6] existují dva hlavní důvody, proč pro imputaci uvažovat simulační metody: 1) představují přirozený doplněk stávajících nástrojů. Zatímco např. EM poskytuje pouze bodový odhad neznámých parametrů, MCMC poskytuje přímo náhodný výběr z jejich společného posteriorního rozdělení. 2) tyto metody umožňují vícenásobnou imputaci, tzn. pro každou chybějící hodnotu metoda nabízí $m > 1$ simulovaných kandidátů.

1.2.5.2.6 Metoda propensitního skóre

Metodu propensitního skóre s imputací založenou na přibližném Bayesovském bootsrapingu používá řada SW řešení (viz kapitola 1.3). Jak uvádí např. [35] nebo [7], metoda je určena pro imputaci spojité proměnné za předpokladu monotónního vzoru chování chybějících dat. Monotonním vzorem rozumí situaci, kdy pokud pro *i*-té pozorování *j*-tá proměnná obsahuje chybějící hodnotu, pak všechny další proměnné s vyšším indexem tohoto pozorování obsahují chybějící hodnotu též. V rámci této imputační metody je pro každou proměnnou obsahující chybějící hodnoty každému pozorování přiřazeno tzv. propensitní skóre jako odhad pravděpodobnosti, že pozorování je chybějící. Propensitní skóre je odvozeno z modelu logistické regrese, jehož vysvětlující proměnné jsou ve vztahu k doplňované proměnné jako „covariates“. Pozorování jsou poté sloučena podle propensitního skóre do předem daného počtu skupin (zpravidla 5). Následně je na ně uplatněna přibližná Bayesovská bootstrap imputace. Uvažujme, že ve skupině č. 1 je $n1$ pozorování bez chybějících hodnot (podskupina A) a $n0$ pozorování s chybějícími hodnotami (podskupina B). Algoritmus nejprve z první skupiny vybere náhodně s vracením $n1$ pozorování a dá je stranou. Poté z nich vybere náhodně s vracením $n0$ pozorování a jejich hodnotami nahradí chybějící pozorování v podskupině B. Tento postup se opakuje pro každou proměnnou s chybějícími hodnotami.

1.2.5.2.7 Techniky založené na stromech

Do této skupiny metod patří rozhodovací stromy a pravidla, klasifikační stromy a regresní stromy. Podstatou rozhodovacích (klasifikačních) stromů je pomocí metody rozděl a panuj rozdělit data na další podmnožiny s převažujícími příklady jedné třídy. Volbu vhodného atributu pro dělení posuzujeme z pohledu kritérií jako je entropie, informační zisk nebo chí-kvadrát. Jak popisuje [3], rozhodovací stromy je možné snadno transformovat na rozhodovací pravidla, která lze později využít v automatizovaném systému (v našem případě pro imputaci). Rozhodovací stromy lze vytvářet na základě kategoriálních atributů. V případě imputace spojitých atributů by bylo nutné nejprve data transformovat na intervaly a po aplikaci stromu pro predikci imputované hodnoty vybrat z intervalu některou z hodnot. Tím bychom se ale dopustili nežádoucího zkreslení.

Lepší metodou pro predikci spojité proměnné jsou regresní stromy. Jak uvádí např. [3], od rozhodovacích stromů se liší jednak použitým kritériem pro větvení, které v tomto případě minimalizuje variabilitu, ale též tím, že v listech stromu nejsou uvedeny kategorie třídy, ale konkrétní hodnota, odpovídající hodnotě průměrné.

Jedním z nejznámějších systémů pro vytváření rozhodovacích stromů je C4.5 spojený se jménem Johna Rosse Quinlana. Jak uvádí [27], jednoduchá imputace pomocí tohoto systému může vést k lepším výsledkům než použití Autoclass[8]. Jiné využití pro C4.5 navrhuje[10], když jej kombinuje s již uvedenými metodami PCA-NN-GA a AANN-GA použití rozhodovacích stromů, konkrétně algoritmu C4.5. Rozhodovací strom je použit pro klasifikaci intervalů chybějících hodnot spojitých proměnných před použitím v již uvedených přístupech. Rozšíření o C4.5 v obou případech vedlo k zvýšení spolehlivosti o 13%.

[14] popisuje použití algoritmu pro generování rozhodovacích pravidel CLIP4 pro jednoduchou imputaci. Algoritmus imputace se skládá z vytvoření stromu, jeho prořezání, vygenerování produkčních pravidel a validace jejich spolehlivosti. Specifickým rysem tohoto algoritmu je využití optimalizačního modelu celočíselného programování při vykonávání operací jako je např. rozdělování dat do podmnožin během první fáze algoritmu imputace.

[8] Autoclass = metoda učení bez učitele pro klasifikaci s využitím Bayesovského přístupu vyvíjená NASA.

[13] uvádí IIA (Incremental Imputation Algorithm) jako aplikaci rozhodovacích stromů s FAST algoritmem založeném na dvoukrokovém dělení se zohledněním globální role prediktoru na lexikograficky setříděná pozorování. Toto setřídění spočívá v seřazení pozorování a atributů v datové matici podle četnosti výskytu chybějících hodnot. Rozhodovací stromy jsou poté aplikovány iterativně od záznamů s nejmenším počtem výskytů chybějících hodnot v rámci atributů s nejmenším počtem vyskytujících se chybějících hodnot.

Porovnání efektivnosti algoritmů C4.5, CN2 a již zmíněného přístupu hledání k-nejbližších sousedů uvádí [25]. V tomto případě ale nebyly rozhodovací stromy použity přímo pro doplnění chybějících hodnot, ale byly otestovány jejich vnitřní mechanismy, kterými se s chybějícími hodnotami vypořádávají. V případě C4.5 se jedná o 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. Tímto způsobem je algoritmus schopen se vypořádat s chybějícími hodnotami v rámci všech proměnných s výjimkou třídy. Vnitřní alogoritmus CN2 je naprotitomu pouze triviální jednoduchou imputací nejčetnější hodnoty. Metody byly otestovány na třech datových souborech o velikosti 300 až 1500 záznamů při mírách uměle náhodně vygenerovaných chybějících hodnot 10% - 60%. 10-NN dosáhl lepších výsledků než oba vnitřní algoritmy i při velké míře zastoupení chybějících hodnot.

[12] ukazuje použití systému CART pro klasifikační a regresní stromy na imputaci chybějících dat ze senzorů bezdrátové sítě. [18] tuto metodu hodnotí pozitivně z pohledu její relativní rezistence vůči odlehlým pozorováním. Představuje podle něj snadný nástroj pro imputaci více hodnot (míněno ve smyslu jednoduché imputace v rámci jednoho atributu). K dalšímu zlepšení spolehlivosti doplňovaných hodnot došlo kombinací CART s lineární interpolací.

[18] zmiňuje techniku Forest Climbing publikovanou v [32] 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. Imputovaný vektor je průnikem hodnot koncových nodů jednotlivých stromů, do nichž sestoupí oskórované pozorování. [18] též popisuje přístup 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. AdaBoost je známý meta-algoritmus používaný pro optimalizaci klasifikace tím, že kombinuje rozhodnutí několika jednodušších klasifikátorů, v tomto případě reprezentovaných rozhodovacími stromy.

1.3 Příklady implementace metod v SW nástrojích

Nástroje pro nahrazování chybějících hodnot lze rozdělit do dvou kategorií: komerční a volně dostupné. Ze skupiny komerčních nástrojů bych zmínil zejména SAS Enterprise Miner, který v rámci nodu Impute umožňuje náhradu chybějících hodnot pomocí rozhodovacích stromů, nejčetnější hodnoty, průměru či mediánu. Imputaci na platformě SAS však není nutné provádět jen s pomocí Mineru. Modul SAS/STAT v sobě obsahuje procedury PROC MI a PROC MIANALYZE, z nichž první uvedená umožňuje provádění imputace s pomocí algoritmu EM, MCMC, regresní analýzy, diskriminační funkce, logistické regrese a tzv. propenzitního skóre, využívajícího přibližné Bayesovské bootstrap imputace. Procedura PROC MIANALYZE potom slouží ke kombinaci výsledků jednotlivých imputací. Podrobný výčet funkcionality PROC MI je dostupný na [35]. Praktické použití PROC MI na příkladech ukazuje [7].

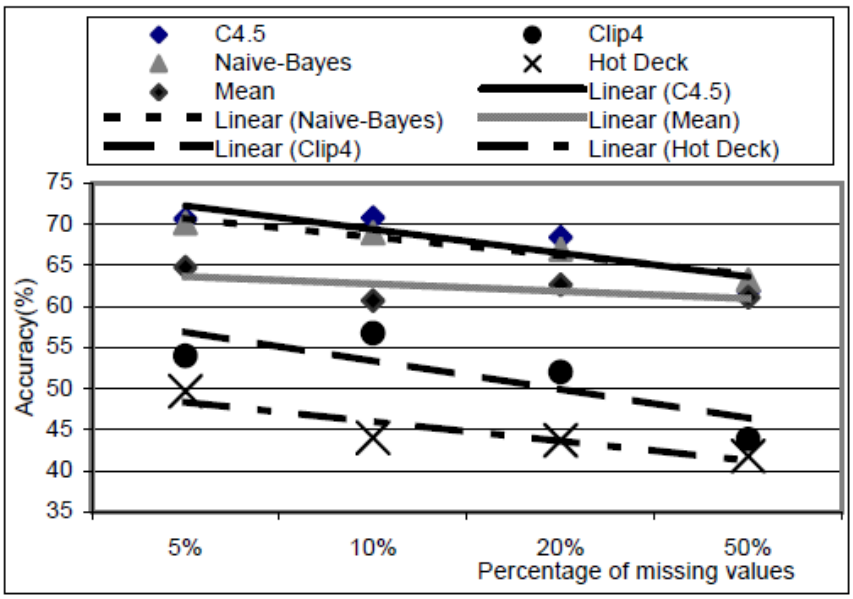
Dalším příkladem komerčního nástroje je SOLAS (viz [33]), spojený s osobou prof. Rubina z Harvardské univerzity. Nástroj poskytuje řadu metod pro jednoduchou i vícenásobnou imputaci. Z první uvedené skupiny metod je možné zmínit hot deck, podmíněný průměr, diskriminační analýzu, MNČ regresi, skupinové průměry a metodu LVCF (Last Value Carried Forward) doplňující chybějící porozování v longitudiálních datech na základě posledního předchozejícího. Z metod pro vícenásobnou imputaci lze uvést např. MI variantu regresní analýzy nebo propenzitní skóre, využívající přibližné Bayesovské bootstrap imputace.

Z nekomerčních nástrojů lze zmínit např. MICE (Multiple Imputation by Chained Equations), dostupný jednak jako knihovna pro nástroje R nebo S-Plus a jednak jako samostatná instalace pro platformu Windows dostupná pod názvem WinMICE. Nástroj poskytuje funkcionalitu pro imputaci prostřednictvím podmíněného průměru, regrese, diskriminační analýzy a MCMC. [9] vyzdvihuje u tohoto nástroje možnost naprogramovat vlastní imputační funkce.

1.4 Benchmark metod

Metodicky ucelený benchmark poskytuje [14], kdy porovnává na celkem 7 datových souborech o různém počtu záznamů, různém počtu proměnných různých typů, při různé míře zastoupení booleovských atributů metody ze čtyř skupin: 1) pravděpodobnostní algoritmy (zástupcem byl zvolen Naïve Bayes), 2) rozhodovací stromy (zástupcem je C4.5) a 3) rozhodovací pravidla (zástupcem je CLIP4), 4) metody učení bez učitele (doplnění průměrem, hot deck). V datových souborech náhodně generuje chybějící hodnoty od relativní četnosti 5% až po 50%. Metody jsou porovnávány na základě srovnání původních a doplněných hodnot.[14] uvádí svůj ambiciozní záměr poskytnout strategie pro použití jednotlivých metod na konkrétní data. Bohužel některé závěry tohoto benchmarku jsou spíše triviálním ověření zřejmých vlastností metod. Uveďme např. tvrzení, že metody bez učitele jsou stabilnější z pohledu rostoucího počtu chybějících hodnot (pochopitelně se snižuje počet záznamů, na kterém je možné metody učení s učitelem natrénovat), nebo že oproti metodám učení s učitelem jejich použití není ovlivněno počtem záznamů či atributů. Stejně tak závěr, že metody učení s učitelem zvyšují svou výkonnost se zvyšujícím se počtem pozorování (tzn. mají více dat pro naučení se), je obecnou vlastností této skupiny metod. Uvedený závěr s rostoucím počtem chybějících hodnot se snižující spolehlivosti modelů založených na učení s učitelem znázorňuje Obrázek č. 1.

Obrázek 1: Porovnání spolehlivosti srovnávaných metod při různém podílu chybějících hodnot



Zdroj: [14]

Nedostatkem tohoto benchmarku je ale především velmi malý počet porovnávaných metod, přestože pokrývají jak zástupce metod učení s učitelem, tak i metody učení bez učitele. Pro skutečně přínosné srovnání by bylo potřeba zahrnout metod více, zejména přístupy uvedené v [10].

Jiné porovnání metod, zmíněné v kapitole popisující logistickou regresi, aspirující na označení benchmark, publikoval [29]. Porovnává použití MLR (Multinomial Logistic Regression) zobecněnou pro klasifikaci do q tříd s metodou LD (Listwise Deletion), MMSI (Mean Mode Single Imputation), již výše zpochybňovanou MNČ regresí a EM algoritmem. Porovnání metod bylo provedeno na datech International Software Benchmarking Standards Group. Metodika pro porovnání jednotlivých metod byla podobná jako u předchozího uvedeného benchmarku, tj. chybějící hodnoty byly náhodně simulovány a spolehlivost metod byla určena na základě porovnání imputovaných hodnot a hodnot původních. Závěry poukazují na 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. Souboj o prvenství proběhl mezi MNČ regresí a MLR. Obě tyto metody při 10% míře vykazovaly podobné výsledky jako ostatní algoritmy, nicméně při větším počtu chybějících hodnot již ostatní ve spolehlivosti předstihly. Při míře neúplnosti dat okolo 30% již vykazovala nejvyšší spolehlivost MLR.

V kapitolách popisujících jednotlivé metody jsem sice uvedl některá další dílčí srovnání, nelze však hovořit o komplexním benchmarku. Vytvoření komplexního benchmarku všech známých metod zůstává zcela jistě velkou výzvou.

1.5 Závěr

V rámci předchozích kapitol jsem stručně seznámil s problematikou neúplných dat, uvedl jsem známé mechanismy vzniku tohoto jevu, od nichž se odvíjí způsob, jakým je možné se s neúplnými daty vypořádat. Ukázal jsem, že na základě různých klasifikací metod uvedených v odborné literatuře, je možné vytvořit klasifikaci univerzální. V rámci této klasifikace jsem uvedl příklady použitelných metod se stručným popisem praktické aplikace tam, kde jsem jej v odborné literatuře nalezl.

V prostudovaných zdrojích se lze setkat s dílčími porovnáními jednotlivých metod. V tomto směru je relativně „nejobsáhlejší“ studie [16] srovnávající efektivnost celkem sedmi metod, dále publikace [29] poskytující srovnání, které se asi nejvíce podobá metodicky ucelenému benchmarku, resp. podobné srovnání publikované v [14], příp. seznam dílčích srovnání publikovaný v [10], hledající též způsoby kooperace jednotlivých metod. Nesetkal jsem se však s komplexním benchmarkem, který by na reálných datech porovnával všechny přístupy a pro jednotlivé typy atributů doporučoval konkrétní metody. Je sice zřejmé, že u pravděpodobnostních metod lze částečně aplikovat závěry komparativních studií efektivnosti využití jednotlivých metod jako jsou např. STATLOG a METAL, přesto zde chybí ucelené porovnání s ostatními uvedenými nepravděpodobnostními metodami. V souladu s připomínkou [10] k testování spolehlivosti MLP-GA proti RBF-GA považuji též za vhodné takový benchmark konstruovat na základě dat z různých předmětných oblastí současně, aby byla eliminována možnost ovlivnění výsledků konkrétním data setem.

Kromě vytoření komplexního benchmarku je námětem pro další výzkum též optimalizace formou kombinace různých metod jako následování cesty, kterou vytyčil [10]. Jako výzvu označuje [14] a [21] též další výzkum a rozvoj metod strategie hot deck.

1.6 Použitá literatura

[1] Dasu T., Johnson T. Exploratory Data Mining and Data Cleansing. Wiley & sons, New Jersey 2003.

[2] Hebák, P. – Hustopecký, J. – Jarošová, E. – Pecáková, I.: Vícerozměrné statistické metody I. Praha, Informatorium 2004.

[3] Berka P. Dobývání znalostí z databází. Academia. Praha, 2003.

[4] R. J. A. Little and D.B. Rubin. 'Statistical Analysis with Missing Data. Wiley, New York, 1987.

[5] Mitchie D., Spiegelhalter D. J., Taylor C. C. (eds): Machine Learning, Neural and Statistical Classification, Ellis Horwood, 1994.

[6] J. L. Schafer. Analysis of Incomplete Multivariate Data. Chapman & Hall, London, 1997.

[7] Y.C. Yang. Multiple imputation for missing data: Concepts and new development. SAS Institute, Inc. 2000.

[8] Zelený M.: „Management Support Systems: Towards Integrated Knowledge Management“, Human Systems Management 7, no 1, 1987, str. 59 – 70.

[9] Horton, N.J. and Lipsitz, S.R. (2001), “Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables,” Journal of the American Statistical Association, 55, 244–254.

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

[11] 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.

[12] Yuka Higashijima, Atsushi Yamamoto, Takayuki Nakamura, Motonori Nakamura, Masato Matsuo, "Missing Data Imputation Using Regression Tree Model for Sparse Data Collected via Wide Area Ubiquitous Network," Applications and the Internet, IEEE/IPSJ International Symposium on, pp. 189-192, 2010 10th IEEE/IPSJ International Symposium on Applications and the Internet, 2010.

[13] Conversano, C.,Siciliano, R. Incremental Tree-Based Missing Data Imputation with Lexicographic Ordering, Journal of Classification 2009-12-01.Springer New York. Computer Science p. 361 - 379.

[14] A. Farhangfar, L. Kurgan, W. Pedrycz, Experimental analysis of methods for imputation of missing values in databases, in: Intelligent Computing: Theory and Applications II Conference, in conjunction with the SPIE Defense and Security Symposium (formerly AeroSense), Orlando, FL, 2004, pp. 172-182.
[15] Wu, X. Barbará, D. Modeling and Imputation of Large Incomplete Multidimensional Datasets. Data Warehousing and Knowledge Discovery. Lecture Notes in Computer Science 2002 Springer Berlin / Heidelberg p. 365-374.
[16] Cartwright, M., Shepperd, M. and Twala, B., Ensemble of missing data techniques to improve software prediction accuracy, , 28th International Conference on Software Engineering (ICSE 2006), 20-28 May 2006, Shanghai, China 4.
[17] Ayuyev, V.V., Jupin, J., Harris, P.W., & Obradovic, Z. Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data. In T.B. Pedersen, M.K. Mohania, and A M. Tjoa (Eds.), Data Warehousing and Knowledge Discovery. Book Series: Lecture Notes in Computer Science. Heidelberg: Springer Berlin. (pp. 366-377).
[18] D'Ambrosio A. Boosted Incremental Tree-based Imputation of Missing Data, PhD. thesis, Universitá degli Studi di Napoli Federico II. 2007.
[19] Shukla D., Singhai R., Thakur N. S., Dembla N. Some Imputation Methods to Treat Missing Values in Knowledge Discovery in Data warehouse. International Journal of Data Engineering (IJDE), Vol. 1, Issue 2. p. 1 – 13.
[20] Dipak V Patil and R S Bichkar. Multiple Imputation of Missing Data with Genetic Algorithm based Techniques. *IJCA Special Issue on Evolutionary Computation* (2):74–78, 2010. Published by Foundation of Computer Science.
[21] Magnani, M.: Techniques for dealing with missing data in knowledge discovery tasks (2004) (Version of June 2004), http://magnanim.web.cs.unibo.it/data/ pdf/missingdata.pdf.
[22] Romero, Vanessa; Salmerón, Antonio. Multivariate imputation of qualitative missing data using Bayesian networks. (English), Soft methodology and random information systems. Collection of papers presented at the 2nd international conference on soft methods in probability and statistics (SMPS'2004), Oviedo, Spain, September 2–4, 2004. Berlin: Springer (ISBN 3-540-22264-2/pbk). Advances in Soft Computing, 605-612 (2004).
[23] Yufeng Ding and Jeffrey S. Simonoff. 2010. An Investigation of Missing Data Methods for Classification Trees Applied to Binary Response Data. *J. Mach. Learn. Res.* 11 (March 2010), 131-170.
[24] Enders, Craig K. Applied missing data analysis (Methodology in the social sciences). New York: Guilford Press, 2010. 375 p. ISBN : 9781606236390.
[25] G.E. Batista, M.C. Monard, A study of k-nearest neighbour as an imputation method, in: Second International Conference on Hybrid Intelligent Systems, vol. 87, , Santiago, Chile, 2002, pp. 251-260.
[26] Rubin, D.B (1976). Inference and Missing Data (with Discusion). Biometrika 63, pp.581-592
[27] Lakshminarayan, K., Harp, S.A., Samad, T. (1999). Imputation of Missing Data in Industrial Databases", Applied Intelligence, 11, 259-275.
[28] Song, Q. and Shepperd, M. (2004). "A Short Note on Safest Default Missingness Mechanism Assumptions", In Empirical Software Engineering. (accepted in 2004)
[29] Sentas, P., Lefteris, A., and Stamelos, I. (2004). "Multiple Logistic Regression as Imputation method Applied on Software Effort prediction", In Proc. of the 10th Int. Symp. on Software Metrics, Chicago, 14-16 September 2004.
[30] Hebák, P. – Hustopecký, J. – Malá, I.: Vícerozměrné statistické metody II. Praha, Informatorium 2005.
[31] Hebák, P. – Hustopecký, J. – Pecáková, I. – Průša, M. – Řezanková, H. – Svobodová, A. – Vlach, P.: Vícerozměrné statistické metody III. Praha, Informatorium 2007.
[32] María Jesús BARCENA, Fernando TUSELL: Data imputation and file merging using the forest climbing algorithm. Dostupné z webu: http://www.et.bs.ehu.es/~etptupaf/pub/papiros/jos.pdf
[33] Webová stránka software SOLAS. [2011-01-11] dostupná pod odkazem: http://www.statistical-solutions-software.com/products-page/solas-for-missing-data-analysis/.
[34] Mařík, V. – Štěpánková, O. – Lažanský, J. – a kol.: Umělá inteligence 5. Academia, Praha 2007.
[35] Dokumentace procedury PROC MI na stránkách SAS [2011-01-11] Dostupné pod odkazem: http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_mi_sect013.htm