Doplňování chybějících hodnot

1. Cíle programu

Účelem programu je umožnit uživateli doplnění chybějících hodnot v datech. Pro doplnění chybějících hodnot je možné použít dvě metody. První metoda nahrazuje chybějící hodnoty nejčastěji se vyskytující hodnotou. Druhá metoda používá k doplnění chybějících hodnot asociační pravidla a kombinuje je s metodou nahrazení chybějících hodnot pomocí nejčastěji se vyskytující hodnoty. Součástí programu je i možnost testování přesnosti odhadu a porovnání přesnosti obou metod. Program nabízí i možnost otestovat množství špatně určených chybějících hodnot druhou metodou při různém nastavení požadavku minimální podpory použitých pravidel. Program umožňuje diskretizaci spojitých veličin. Díky této funkci je umožněno doplnění chybějících hodnot i na základě závislostí mezi hodnotami hledaného atributu a hodnotami jiných atributů, které mohou být i spojité. Při doplňování chybějících hodnot pomocí asociačních pravidel je umožněno nalezené kategorie zpět převést na číselné hodnoty. Díky této nové funkci je tak možné doplňovat chybějící hodnoty i ve spojitých atributech.

2. Vstupní data

Jednotlivé atributy jsou označeny OT01 až OT99 – první řádek. Jednotlivé kategorie se označují písmeny A až Y. Písmeno Z označuje kategorii chybějících hodnot. Je možné vkládat i číselné hodnoty, které musí být před samotným doplňováním chybějících hodnot převedeny na kategorie pomocí diskretizace příslušnými tlačítky na zadávacím listu.

3. Popisy listů

3.1. Zadávací list

Do zadávacího listu se vkládají vstupní údaje o datovém souboru. Vždy je třeba zadat počet atributů (B4) - sloupců datového souboru a počet případů (B5) -řádků datového souboru. Počet hodnot (B6) se vypočítá sám.

V případě, že chcete použít funkci generování chybějících hodnot, je třeba vyplnit i požadovaný počet chybějících hodnot (B7) v procentech z celkového množství hodnot. Počet chybějících hodnot (B8) se vypočítá sám.

Pokud jsou jako vstupní data použity i spojité veličiny, je třeba vyplnit názvy spojitých atributů a požadovaný počet kategorií, na které mají být převedeny (sloupce H a J, začněte vyplňovat od řádku 22 a použijte tolik řádků, kolik je vstupních spojitých atributů). Maximální počet kategorií pro každý atribut je 25, minimální počet je 1 kategorie. Požadovaný počet kategorií musí být větší nebo roven počtu nechybějících hodnot příslušného atributu.

Zadávací list dále obsahuje i některé informace o výstupech z některých funkcí softwaru (konkrétně buňky B9 až B14). Prvním výstupem je skutečný počet vygenerovaných chybějících hodnot (B9) a skutečný počet vygenerovaných chybějících hodnot vztažený na celkový počet hodnot v procentech (B10). Počet skutečně vygenerovaných chybějících hodnot v procentech je obvykle nepatrně vyšší než počet požadovaných chybějících hodnot, vzhledem k nutnosti zaokrouhlení počtu vygenerovaných chybějících hodnot na nejbližší vyšší celé číslo.

Dalšími výstupy jsou výstupy z testování metod pro doplňování chybějících hodnot. Jsou to počet špatně určených chybějících hodnot metodou nejčastějšího výskytu (B11) a v procentech spočtený poměr špatně určených hodnot touto metodou k celkovému počtu chybějících hodnot (B12). Obdobně jsou obě hodnoty k dispozici i pro metodu využívající k doplnění chybějících hodnot asociační pravidla (B13 a B14). V této verzi programu i tato metoda doplňuje některé chybějící hodnoty pomocí nejčastějšího výskytu. Počet chybějících hodnot doplněných nejčastěji se vyskytující hodnotou, při doplňování chybějících hodnot metodou využívající asociační pravidla, je uveden v buňce B15.

Tabulka začínající řádkem 17 slouží pouze pro výstup z testování nastavení podpory. Sloupec „Počet hodnot doplněných výskytem při doplňování AP“ uvádí počet hodnot doplněných nejčastěji se vyskytující hodnotou při použití metody využívající asociační pravidla. Sloupec „psuh“ označuje počet špatně určených chybějících hodnot. Sloupec „podp.“ uvádí nastavení minimální požadované podpory asociačních pravidel v počtech případů. Sloupec „podp. [%]“ pak uvádí totéž, ale v procentech z celkového počtu případů. Sloupec „psuh [%]“ uvádí počet špatně určených hodnot v procentech, vztažený k celkovému počtu chybějících hodnot. Sloupec „ppAP“ uvádí počet použitých asociačních pravidel pro doplňování chybějících hodnot.

3.2. Diskretizace

Tento list slouží při diskretizaci spojitých atributů pro účely následného převodu nalezených kategorií chybějících hodnot zpět na číselné hodnoty. První řádek tohoto listu obsahuje označení atributů (OT01 – OT99). Pro číselné atributy se zde ukládá průměrná hodnota z intervalu každé příslušné kategorie.

3.3. Vstupní data – spojitá

Do tohoto listu se vkládají vstupní data pokud obsahují jak číselné tak kategoriální atributy. První řádek tohoto listu obsahuje označení atributů (OT01 – OT99). Další řádky poté obsahují samotná data. Kategorie hodnot se zde označují písmeny A až Y. Kategorie Z je vyhrazena pro chybějící hodnoty. Číselné (spojité) atributy mohou obsahovat číselné hodnoty a chybějící hodnoty. Před samotným doplňováním chybějících hodnot pomocí asociačních pravidel musí být číselné hodnoty převedeny na kategorie pomocí diskretizace příslušným tlačítkem na zadávacím listu (viz. kapitola 3.1 a kapitola 4.).

3.4. Vstupní data

Do tohoto listu se vkládají vstupní data pro generování chybějících hodnot. A to pouze v případě, že se jedná pouze o kategoriální data, která neobsahují žádný spojitý atribut ani chybějící hodnoty. Jedná se tedy o kompletní datový soubor, ze kterého je generován datový soubor s chybějícími hodnotami, zejména pro účely testování metod pro doplňování chybějících hodnot.

První řádek tohoto listu obsahuje označení atributů (OT01 – OT99). Další řádky poté obsahují samotná data. Kategorie hodnot se zde označují písmeny A až Y. List je využíván i pro uložení výsledku diskretizace číselných atributů pokud vstupní data – spojitá neobsahovali chybějící hodnoty.

3.5. Nekompletní datový soubor

Tento list obsahuje nekompletní datový soubor. V případě, že je generován z kompletního datového souboru, je vyplněn automaticky stisknutím tlačítka „Vytvoř nekompletní datový soubor“. List může být také automaticky vyplněn při diskretizaci spojitých vstupních dat, pokud obsahovali chybějící hodnoty.

Tento list lze samozřejmě vyplnit i přímo. Přímo tento list vyplňujeme v případě, že se nejedná o testování metod a pouze doplňujeme chybějící hodnoty do nekompletního datového souboru a zároveň se jedná pouze o kategoriální data. Jinak používáme list Vstupní data – spojitá.

První řádek tohoto listu obsahuje označení atributů (OT01 – OT99). Další řádky poté obsahují samotná data. Kategorie známých hodnot se zde označují písmeny A až Y. Písmeno Z slouží pro označení kategorie chybějících hodnot.

3.6. Doplněný datový soubor – výskyt

Tento list slouží jako výstup pro metodu doplňující chybějící hodnoty pomocí nejčastějšího výskytu hodnoty atributu.

První řádek tohoto listu opět obsahuje označení atributů (OT01 – OT99). Další řádky obsahují samotná data, ve kterých jsou již chybějící hodnoty doplněny.

3.7. Asociační pravidla – zdroj

Tento list slouží jako vstup, do kterého lze zkopírovat vygenerovaná asociační pravidla v softwaru Weka. Program je otestován pro práci s vygenerovanými asociačními pravidly ve verzi softwaru Weka 3.6.1. Schopnost pracovat s vygenerovanými pravidly v jiných verzích softwaru Weka nebyla testována. Tento list vyplníme jednoduše tak, že vygenerovaná asociační pravidla v softwaru Weka zkopírujeme do schránky a poté vložíme do tohoto listu. Vygenerovaná asociační pravidla musí být seřazena sestupně podle jejich spolehlivosti.

3.8. Asociační pravidla

Tento list obsahuje asociační pravidla zapsaná ve strukturované formě. List lze samozřejmě vyplnit ručně, ale zejména lze využít automatické vyplnění stisknutím tlačítka „Rozeber asociační pravidla“ na „zadávacím listu“. Tento list je pak vyplněn na základě vyplněného listu „Asociační pravidla – zdroj“.

První řádek tohoto listu obsahuje hlavičku. Sloupec „ID“ obsahuje pořadové číslo pravidla. Je použito číslo pravidla z listu „Asociační pravidla – zdroj“. Pokud je list vyplňován jiným způsobem, než automaticky, lze pravidla číslovat od jedné do celkového počtu pravidel.

Sloupec „spolehlivost“ obsahuje spolehlivost pravidla. Tato je udávána v rozsahu 0 až 1 (0=0%, 1=100%).

Sloupec „podpora [počet]“ obsahuje podporu pravidla, vyjádřenou počtem případů.

Sloupec „podpora [-]“ obsahuje podporu pravidla, vztaženou na celkový počet případů.

Sloupec „závěr (atribut)“ obsahuje atribut v závěru pravidla.

Sloupec „závěr (hodnota)“ obsahuje hodnotu atributu v závěru pravidla.

Další sloupce (OT01 až OT99) obsahují hodnoty v předpokladu pravidla, příslušné k jednotlivým atributům. Není-li hodnota některého z atributů vyplněna, znamená to, že tento atribut v předpokladu pravidla není obsažen.

Předpokladem pro doplnění chybějících hodnot pomocí asociačních pravidel je, že asociační pravidla v tomto listu budou seřazena sestupně dle jejich spolehlivosti.

3.9. Doplněný datový soubor – AP.

Tento list slouží jako výstup pro metodu doplňující chybějící hodnoty pomocí asociačních pravidel.

První řádek tohoto listu opět obsahuje označení atributů (OT01 – OT99). Další řádky obsahují samotná data, ve kterých jsou již chybějící hodnoty doplněny.

3.10. Doplněný datový soubor – AP – sp.

Tento list slouží jako výstup v případě, že původní datový soubor obsahoval i číselné atributy. Datový soubor lze převést zpět tak aby atributy byly stejného typu jako v původním datovém souboru. U číselných atributů jsou převedeny i nalezené kategorie.

První řádek tohoto listu opět obsahuje označení atributů (OT01 – OT99). Další řádky obsahují samotná data, ve kterých jsou již chybějící hodnoty doplněny.

3.11. Asociační pravidla - původní

Tento list slouží pro zálohování strukturované podoby asociačních pravidel při testování nastavení podpory. Pro urychlení práce programu jsou se vzrůstajícím požadavkem podpory asociační pravidla ve strukturované podobě s nedostatečnou podporou z listu „Asociační pravidla“ umazávána.

Záloha strukturovaných asociačních pravidel probíhá automaticky na začátku procesu testování nastavení podpory a na konci procesu opět probíhá obnova strukturovaných asociačních pravidel do listu „Asociační pravidla“. Je ovšem doporučeno strukturovaná asociační pravidla před spuštěním testování nastavení podpory ještě na víc zálohovat.

4. Funkce programu

Pro usnadnění práce s programem je možné zapnout program do několika módů, které zpřístupňují v různé míře jednotlivé funkce programu. V rámci těchto módů dochází také k zamykání některých listů a tlačítek, které se v příslušném módu nepoužívají.

4.1. Generování nekompletního souboru

Generování nekompletního datového souboru slouží zejména k účelu testování metod pro doplňování chybějících hodnot. V tomto režimu software pouze vygeneruje požadovaný počet chybějících hodnot ve vstupním kompletním datovém souboru. Chybějící hodnoty mají rovnoměrné rozdělení.

Postup při generování nekompletního datového souboru:

1. V zadávacím listu vyplňte počet případů a počet atributů vstupního kompletního datového souboru.

2. V zadávacím listu vyplňte požadovaný počet chybějících hodnot v procentech (B7).

3. Pro vkládání dat včetně číselných atributů použijte list 'Vstupní data – spojitá'. Pokud vkládáte pouze kategoriální data, použijte list 'Vstupní data' (podrobnosti v kapitole 3 – list Vstupní data – spojitá a list Vstupní data).

4. Pokud jsou jako vstupní data použity i spojité veličiny, je třeba vyplnit názvy spojitých atributů a požadovaný počet kategorií, na které mají být převedeny (sloupce H a J v zadávacím listu , začněte vyplňovat od řádku 22 a použijte tolik řádků, kolik je vstupních spojitých atributů). Maximální počet kategorií pro každý atribut je 25, minimální počet je 1 kategorie. Požadovaný počet kategorií musí být větší nebo roven počtu nechybějících hodnot příslušného atributu.

5. V případě potřeby diskretizujte spojité atributy tlačítkem „ Diskretizace spojitých atributů - vstupní data“.

6. Tlačítkem 'Vytvoř nekompletní datový soubor' vygenerujete požadovaný nekompletní datový soubor s chybějícími hodnotami. Výstupní nekompletní datový soubor naleznete v listu „Nekompletní datový soubor“ (podrobnosti v kapitole 3 – list Nekompletní datový soubor).

4.2. Doplnění chybějících hodnot pomocí asociačních pravidel

Tento režim programu slouží pro doplnění chybějících hodnot pomocí metody, která kombinuje asociační pravidla a nejčastější výskyt. Je tak vyřešen problém, kdy pro všechny chybějící hodnoty neexistují příslušná asociační pravidla pro jejich doplnění. Zároveň je použita metoda doplnění chybějících hodnot pomocí nejčastěji se vyskytující veličiny tam, kde relativní četnost výskytu nejčastěji se vyskytující hodnoty je vyšší, než spolehlivost použitelného asociačního pravidla s nejvyšší spolehlivostí.

Postup při doplnění chybějících hodnot pomocí asociačních pravidel:

1. V zadávacím listu vyplňte počet případů a počet atributů nekompletního datového souboru k doplnění.

2. Pro vkládání dat včetně číselných atributů použijte list 'Vstupní data – spojitá'. Pokud vkládáte pouze kategoriální data, použijte list 'Nekompletní datový soubor' (podrobnosti v kapitole 3 – list Vstupní data – spojitá a list Nekompletní datový soubor).

3. Pokud jsou jako vstupní data použity i spojité veličiny, je třeba vyplnit názvy spojitých atributů a požadovaný počet kategorií, na které mají být převedeny (sloupce H a J v zadávacím listu , začněte vyplňovat od řádku 22 a použijte tolik řádků, kolik je vstupních spojitých atributů). Maximální počet kategorií pro každý atribut je 25, minimální počet je 1 kategorie. Požadovaný počet kategorií musí být větší nebo roven počtu nechybějících hodnot příslušného atributu.

4. V případě potřeby diskretizujte spojité atributy tlačítkem „Diskretizace spojitých atributů - nekompletní datový soubor“.

5. Do listu 'Asociační pravidla - zdroj' vložte vygenerovaná asociační pravidla ze softwaru Weka (podrobnosti v kapitole 3 – list Asociační pravidla – zdroj). Tento krok, společně s krokem 6, můžete vynechat v případě, že

můžete asociační pravidla zapsat přímo ve strukturované podobě. V tomto případě asociační pravidla ve strukturované podobě zadáváte přímo do listu Asociační pravidla (podrobnosti v kapitole 3 – list Asociační pravidla).

6. Tlačítkem 'Rozeber asociační pravidla' převedete nestrukturovaný zápis asociačních pravidel do strukturované podoby. Výstup z tohoto kroku je v listu „Asociační pravidla“ (podrobnosti v kapitole 3 – list Asociační pravidla). Tento krok, společně s krokem 5, můžete vynechat v případě, že můžete asociační pravidla zapsat přímo ve strukturované podobě. V tomto případě asociační pravidla ve strukturované podobě zadáváte přímo do listu Asociační pravidla (podrobnosti v kapitole 3 – list Asociační pravidla).

7. Tlačítkem 'Najdi chybějící hodnoty pomocí asociačních pravidel' spusťte doplnění chybějících hodnot. Výstup z tohoto kroku je v listu Doplněný datový soubor – AP (podrobnosti v kapitole 3 - Doplněný datový soubor – AP).

8. V případě, že byl nekompletní datový soubor generován z kompletního, můžete zjistit úspěšnost algoritmu stisknutím tlačítka „Spočti úspěšnost algoritmu hledání chybějících hodnot - asociační pravidla“. Výsledky poté najdete v „Zadávacím listu“ v buňkách B13 a B14 (týká se pouze nalezených kategorií)

9. V případě, že původní datový soubor obsahoval i číselné atributy, můžete tlačítkem „převeď atributy na spojité“ získat zpět datový soubor tak aby atributy byly stejného typu jako v původním datovém souboru. U číselných atributů jsou převedeny na číselné hodnoty i nalezené kategorie.

4.3. Doplnění chybějících hodnot metodou nejčastějšího výskytu

Tento režim programu slouží pro doplnění chybějících hodnot metodou nejčastějšího výskytu.

Postup při doplnění chybějících hodnot metodou nejčastějšího výskytu:

1. V zadávacím listu vyplňte počet případů a počet atributů nekompletního datového souboru k doplnění.

2. Pro vkládání dat včetně číselných atributů použijte list 'Vstupní data – spojitá'. Pokud vkládáte pouze kategoriální data, použijte list 'Nekompletní datový soubor' (podrobnosti v kapitole 3 – list Vstupní data – spojitá a list Nekompletní datový soubor).

3. Pokud jsou jako vstupní data použity i spojité veličiny, je třeba vyplnit názvy spojitých atributů a požadovaný počet kategorií, na které mají být převedeny (sloupce H a J v zadávacím listu , začněte vyplňovat od řádku

22 a použijte tolik řádků, kolik je vstupních spojitých atributů). Maximální počet kategorií pro každý atribut je 25, minimální počet je 1 kategorie. Požadovaný počet kategorií musí být větší nebo roven počtu nechybějících hodnot příslušného atributu.

4. V případě potřeby diskretizujte spojité atributy tlačítkem „Diskretizace spojitých atributů - nekompletní datový soubor“.

5. Tlačítkem 'Najdi chybějící hodnoty pomocí metody nejčastějšího výskytu' spusťte doplnění chybějících hodnot. Výstup z tohoto kroku je v listu Doplněný datový soubor – výskyt (podrobnosti v kapitole 3 - Doplněný datový soubor – výskyt).

6. V případě, že byl nekompletní datový soubor generován z kompletního, můžete zjistit úspěšnost algoritmu stisknutím tlačítka „Spočti úspěšnost algoritmu hledání chybějících hodnot – výskyt“. Výsledky poté najdete v „Zadávacím listu“ v buňkách B11 a B12 (týká se pouze nalezených kategorií).

V případě, že u kategoriálního atributu chybí hodnoty u všech případů, vrací tato funkce do doplněného datového souboru místo chybějících hodnot hodnotu A.

4.4. Porovnání obou metod

Tento režim programu slouží zejména k porovnání obou metod pro doplňování chybějících hodnot. V tomto režimu jsou zpřístupněny veškeré funkce programu mimo funkce testování nastavení podpory.

Postup při porovnání obou metod pro doplňování chybějících hodnot:

1. V zadávacím listu vyplňte počet případů a počet atributů vstupního kompletního datového souboru.

2. V zadávacím listu vyplňte požadovaný počet chybějících hodnot v procentech (B7).

3. Pro vkládání dat včetně číselných atributů použijte list 'Vstupní data – spojitá'. Pokud vkládáte pouze kategoriální data, použijte list 'Vstupní data' (podrobnosti v kapitole 3 – list Vstupní data – spojitá a list Vstupní data).

4. Pokud jsou jako vstupní data použity i spojité veličiny, je třeba vyplnit názvy spojitých atributů a požadovaný počet kategorií, na které mají být převedeny (sloupce H a J v zadávacím listu , začněte vyplňovat od řádku 22 a použijte tolik řádků, kolik je vstupních spojitých atributů). Maximální počet kategorií pro každý atribut je 25, minimální počet je 1 kategorie. Požadovaný počet kategorií musí být větší nebo roven počtu nechybějících hodnot příslušného atributu.

5. V případě potřeby diskretizujte spojité atributy tlačítkem „ Diskretizace spojitých atributů - vstupní data“.

6. Tlačítkem 'Vytvoř nekompletní datový soubor' vygenerujete požadovaný nekompletní datový soubor s chybějícími hodnotami. Výstupní nekompletní datový soubor z tohoto kroku naleznete v listu „Nekompletní datový soubor“ (podrobnosti v kapitole 3 – list Nekompletní datový soubor).

7. Tlačítkem 'Najdi chybějící hodnoty pomocí metody nejčastějšího výskytu' spusťte doplnění chybějících hodnot. Výstup z tohoto kroku je v listu Doplněný datový soubor – výskyt (podrobnosti v kapitole 3 - Doplněný datový soubor – výskyt).

8. Do listu 'Asociační pravidla - zdroj' vložte vygenerovaná asociační pravidla ze softwaru Weka (podrobnosti v kapitole 3 – list Asociační pravidla – zdroj). Tento krok, společně s krokem 9, můžete vynechat v případě, že můžete asociační pravidla zapsat přímo ve strukturované podobě. V tomto případě asociační pravidla ve strukturované podobě zadáváte přímo do listu Asociační pravidla (podrobnosti v kapitole 3 – list Asociační pravidla).

9. Tlačítkem 'Rozeber asociační pravidla' převedete nestrukturovaný zápis asociačních pravidel do strukturované podoby. Výstup z tohoto kroku je v listu „Asociační pravidla“ (podrobnosti v kapitole 3 – list Asociační pravidla). Tento krok, společně s krokem 8, můžete vynechat v případě, že můžete asociační pravidla zapsat přímo ve strukturované podobě. V tomto případě asociační pravidla ve strukturované podobě zadáváte přímo do listu Asociační pravidla (podrobnosti v kapitole 3 – list Asociační pravidla).

10. Tlačítkem 'Najdi chybějící hodnoty pomocí asociačních pravidel' spusťte doplnění chybějících hodnot. Výstup z tohoto kroku je v listu Doplněný datový soubor – AP (podrobnosti v kapitole 3 - Doplněný datový soubor – AP).

11. Zjistěte úspěšnost algoritmu pro doplňování chybějících hodnot metodou nejčastějšího výskytu stisknutím tlačítka „Spočti úspěšnost algoritmu hledání chybějících hodnot – výskyt“. Výsledky poté najdete v „Zadávacím listu“ v buňkách B11 a B12 (týká se pouze nalezených kategorií).

12. Zjistěte úspěšnost algoritmu pro doplňování chybějících hodnot pomocí asociačních pravidel stisknutím tlačítka „Spočti úspěšnost algoritmu hledání chybějících hodnot - asociační pravidla“. Výsledky poté najdete v „Zadávacím listu“ v buňkách B13, B14 a B15 (týká se pouze nalezených kategorií).

13. V případě, že původní datový soubor obsahoval i číselné atributy, můžete tlačítkem „převeď atributy na spojité“ získat zpět datový soubor tak aby

atributy byly stejného typu jako v původním datovém souboru. U číselných atributů jsou převedeny na číselné hodnoty i nalezené kategorie (týká se pouze doplňování pomocí asociačních pravidel).

4.5. Testování nastavení podpory

Tento režim dává uživateli možnost zjistit, jak ovlivňuje nastavení požadavku minimální podpory doplňování chybějících hodnot metodou asociačních pravidel kombinovanou s metodou doplnění chybějících hodnot pomocí metody nejčastějšího výskytu. Tento režim je dostupný pouze pro kategoriální data.

Postup při testování nastavení podpory:

1. V zadávacím listu vyplňte počet případů a počet atributů vstupního kompletního datového souboru.

2. V zadávacím listu vyplňte požadovaný počet chybějících hodnot v procentech (B7).

3. Pro vkládání dat včetně číselných atributů použijte list 'Vstupní data – spojitá'. Pokud vkládáte pouze kategoriální data, použijte list 'Vstupní data' (podrobnosti v kapitole 3 – list Vstupní data – spojitá a list Vstupní data).

4. Pokud jsou jako vstupní data použity i spojité veličiny, je třeba vyplnit názvy spojitých atributů a požadovaný počet kategorií, na které mají být převedeny (sloupce H a J v zadávacím listu , začněte vyplňovat od řádku 22 a použijte tolik řádků, kolik je vstupních spojitých atributů). Maximální počet kategorií pro každý atribut je 25, minimální počet je 1 kategorie. Požadovaný počet kategorií musí být větší nebo roven počtu nechybějících hodnot příslušného atributu.

5. V případě potřeby diskretizujte spojité atributy tlačítkem „ Diskretizace spojitých atributů - vstupní data“.

6. Tlačítkem 'Vytvoř nekompletní datový soubor' vygenerujete požadovaný nekompletní datový soubor s chybějícími hodnotami. Výstupní nekompletní datový soubor z tohoto kroku naleznete v listu „Nekompletní datový soubor“ (podrobnosti v kapitole 3 – list Nekompletní datový soubor).

7. Tlačítkem 'Najdi chybějící hodnoty pomocí metody nejčastějšího výskytu' spusťte doplnění chybějících hodnot. Výstup z tohoto kroku je v listu Doplněný datový soubor – výskyt (podrobnosti v kapitole 3 - Doplněný datový soubor – výskyt). Tento krok můžete společně s krokem 10 vynechat, pokud nemáte zájem o zjištění úspěšnosti doplnění chybějících hodnot metodou nejčastějšího výskytu.

8. Do listu 'Asociační pravidla - zdroj' vložte vygenerovaná asociační pravidla ze softwaru Weka (podrobnosti v kapitole 3 – list Asociační pravidla – zdroj). Tento krok, společně s krokem 9, můžete vynechat v případě, že můžete asociační pravidla zapsat přímo ve strukturované podobě. V tomto případě asociační pravidla ve strukturované podobě zadáváte přímo do listu Asociační pravidla (podrobnosti v kapitole 3 – list Asociační pravidla).

9. Tlačítkem 'Rozeber asociační pravidla' převedete nestrukturovaný zápis asociačních pravidel do strukturované podoby. Výstup z tohoto kroku je v listu „Asociační pravidla“ (podrobnosti v kapitole 3 – list Asociační pravidla). Tento krok, společně s krokem 8, můžete vynechat v případě, že můžete asociační pravidla zapsat přímo ve strukturované podobě. V tomto případě asociační pravidla ve strukturované podobě zadáváte přímo do listu Asociační pravidla (podrobnosti v kapitole 3 – list Asociační pravidla).

10. Zjistěte úspěšnost algoritmu pro doplňování chybějících hodnot metodou nejčastějšího výskytu stisknutím tlačítka „Spočti úspěšnost algoritmu hledání chybějících hodnot – výskyt“. Výsledky poté najdete v „Zadávacím listu“ v buňkách B11 a B12. Tento krok můžete společně s krokem 7 vynechat, pokud nemáte zájem o zjištění úspěšnosti doplnění chybějících hodnot metodou nejčastějšího výskytu.

11. Tlačítkem „Testování nastavení podpory“ spustíte testování nastavení podpory. Výsledky poté najdete v „Zadávacím listu“ v buňkách B13, B14 a B15 a v tabulce začínající na řádku 17. Podrobnosti k jednotlivým údajům jsou uvedeny v kapitole 3.1.

5. Omezení

Počet atributů je omezen na 99

Počet případů je omezen maximálním počtem řádků v jednom listu, přičemž od tohoto maximálního počtu řádků je třeba odečíst jeden řádek obsahující označení atributů (OT01 - OT99). (Ve verzi MS Excel 2007 je maximální počet řádků v jednom listu 65536 řádků.)

Počet různých kategorií je omezen na písmena A až Y (25 kategorií). Kategorie Z je určena pro chybějící hodnoty. Nově je v této verzi programu možné vkládat i číselné hodnoty, které musí být před samotným doplňováním chybějících hodnot převedeny na kategorie pomocí diskretizace příslušnými tlačítky na zadávacím listu.

Počet vkládaných asociačních pravidel je omezený maximálním počtem řádků v jednom listu, přičemž od tohoto maximálního počtu řádků je třeba odečíst počet řádků, které nejsou využity pro uložení asociačních pravidel. Jedná

se o hlavičku obsahující informace o generování asociačních pravidel. (Ve verzi MS Excel 2007 je maximální počet řádků v jednom listu 65536 řádků.)

Software je testován ve verzi MS Excel 2007.