Nové metody a postupy v oblasti přístrojové techniky, automatického řízení a informatiky
Ústav přístrojové a řídicí techniky ČVUT v Praze, odbor automatického řízení a inženýrské informatiky, Technická 4, 166 07 Praha 6

VYUŽITÍ PROSTŘEDKŮ VYTĚŽOVÁNÍ ZNALOSTÍ Z DATABÁZÍ V SYSTÉMU MONITOROVÁNÍ VÝVOJE KRAJINY

Jiří Bíla a Jakub Jura

Abstrakt: Článek informuje o jedné linii výzkumu sledované v rámci výzkumného úkolu „Vývoj metod stanovení toků energie a látek ve vybraných ekosystémech. Návrh a ověření principů hodnocení hospodářských zásahů pro zajištění podmínek autoregulace a rozvoje biodiverzity“. Je stručně uvedena základní struktura systému měřicích stanic a struktura databáze. Hlavní pozornost je věnována technice vytěžování znalostí z databází. V daném případě jsou blíže popsány dvě metody: metoda využívající tzv. *konceptuálních svazů,* která vede k objevování znalostí ve formě systému pravidel a metoda *rough množin*, která umožňuje vytvářet aproximace datových a znalostních struktur. Demonstrace metod je provedena na segmentu databáze získaného z jedné z měřicích stanic monitorovacího systém.

Klíčová slova: Databáze, vytěžování znalostí z databází, konceptuální svaz, rough množiny, aproximace dat, systém pravidel.

1. Úvod

Cílem projektu [2] je popis energetických toků a vývoje biodiverzity v krajině. Krajina se zdá stabilní, ale zásahy do krajiny probíhají dlouho a stále. Výzkum se věnuje zejména mikro-meteorologii, termodynamice atmosféry v těsné blízkosti povrchů různých lokalit (zelený porost, rašeliniště, betonová plocha, …), velkému a malému vodnímu cyklu, [1]. Data pro monitorovací systém jsou získávána z 11 měřicích stanic rozmístěných ve vybraných lokalitách ekosystému. Každá stanice měří 13 veličin (zpravidla v intervalech 6 minut). Data jsou archivována, testována a editována v databázi, jsou využívána k nejrůznějším výpočtům (např. k výpočtům dělení sluneční radiace – dopadající, odražené, zachycené v povrchu, …) a k objevování dosud neznámých vztahů a znalostí (Data Mining and Knowledge Discoveries).

2. Struktura databáze systému

Struktura databáze na obr. 1 a má následující komponenty:

- *Centrální modul*: Umožňuje vstup do systému a koordinační funkce jednotlivých bloků.
- Třída *Uživatelské rozhraní* centrálního modulu (**TUR_CM**): Představuje obrazovku uživatele s možností spouštění základních funkcí jednotlivých bloků.
- *Třída měřicích stanic*: Je nadřazenou třídou pro koordinaci funkcí měřicích stanic.
- *Třída měřicí stanice* (**TMS**): Obsahuje základní manipulační operace s daty měřicích stanic.
- Třída *Uživatelské rozhraní* měřicí stanice (**TUR_MS**): Představuje obrazovku uživatele pro ovládání manipulačních operací s daty měřicích stanic.
- Třída *Řízení* (**TŘízení**): Je třídou s operacemi předávání řízení mezi jednotlivými bloky.

3. Konceptuální svazy a rough množiny

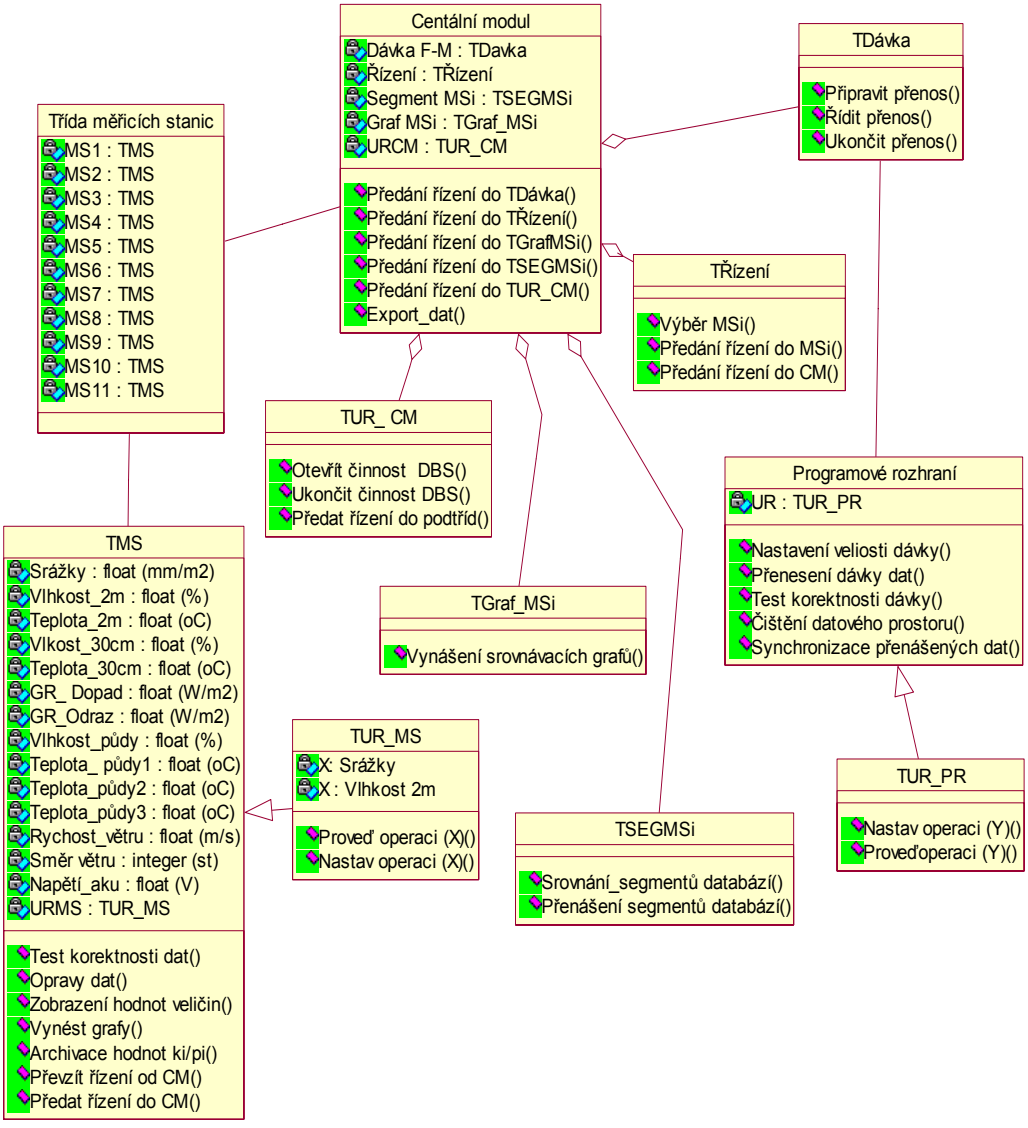
3.1 Konceptuální svaz

Definice 1: Trojici $C = (O, I, R)$, (O je množina objektů, I je množina položek (atributů) a R je binární relace $R \subseteq O \times I$) nazýváme *kontextem vytěžování dat* (Data Mining Context), [3].

Přirozeným sdružováním objektů a vztahů mezi objekty a jejich množinami položek získáme tzv. *konceptuální svaz* na kontextu vytěžování dat (dále označovaný L). Každý prvek svazu L odvozený z kontextu C je dvojice $\langle X, Y \rangle$, kde X je množina objektů $X \subseteq O$ a Y je množina položek $Y \subseteq I$. Každá dvojice $\langle X, Y \rangle$ splňuje následující podmínky vzhledem k relaci R.

$$X = \{x \in O \mid \forall y \in Y, x \, R \, y \}, \tag{1}$$

$$Y = \{y \in I \mid \forall x \in X, x \, R \, y \}. \tag{2}$$



Obr.1. Schéma databázového systému

X je největší množina objektů popsaná vlastnostmi z Y, a Y je největší množina položek společná všem objektům z X. Prostřednictvím parciálního uspořádání "< " (které je vyjádřeno daným svazem) lze zkonstruovat tzv. *Hassův diagram*:

- Hrana z H_1 do H2 existuje pokud H1 < H2 a není žádný prvek H3 svazu takový, že H1 < H3 < H2.
- H1 je předchůdce prvku H2 (H2 je následník prvku H1).
- Dvojice ⟨X, Y⟩ představuje uzel v Hassův diagramu.

Hassův diagram vyjadřuje vztah "*zobecnění/specifikace*" mezi konceptuálními uzly. Hassův diagram uvažován jako acyklický graf s jedinou dodatečnou podmínkou: každá dvojice uzlů má jediného společného nejbližšího předchůdce a jediného následníka.

***Příklad 1**:* Použití řádků databáze pro metodu konceptuálního svazu není přímé. Položky v řádcích databáze je nutno sdružit v *monitorovací třídy* (např. standardní podmínky malého vodního cyklu v lokalitě L1, rozpad malého vodního cyklu V L1, standardní evaporace a transpirace rostlin v L1, vysušování oblasti L1, ….) a *situace* (např. fragmenty databáze (soubory 13 složkových vektorů) popisujících pohyb měřených veličin v daném monitorovacím období), které přicházejí. Pak lze uplatnit techniky vytěžování dat. Řekněme, že máme 5 monitorovacích tříd a fragment databáze transformovaný do 5 opakovaných situací. Příslušnost jednotlivých situací k monitorovacím třídám popisuje matice na obr. 2. Nyní můžeme použít pojmů kontextů na konceptuálním svazu: Uvažujme *kontext vytěžování dat C = (*{A_0, A_1, A_2, A_3, A_4},{3, 4, 7, 8, 9}, *R)* s relací R reprezentovanou v tabulce $\mathbf{M_G}$ na obr. 2. Hassův diagram, který reprezentuje daný kontext *C,* je na obr.3.

$\mathbf{M_G}$	3	4	7	8	9
A_0	1	1	1	1	1
A_1	1	1			
A_2		1	1	1	1
A_3	1		1	1	
A_4	1	1	1		1

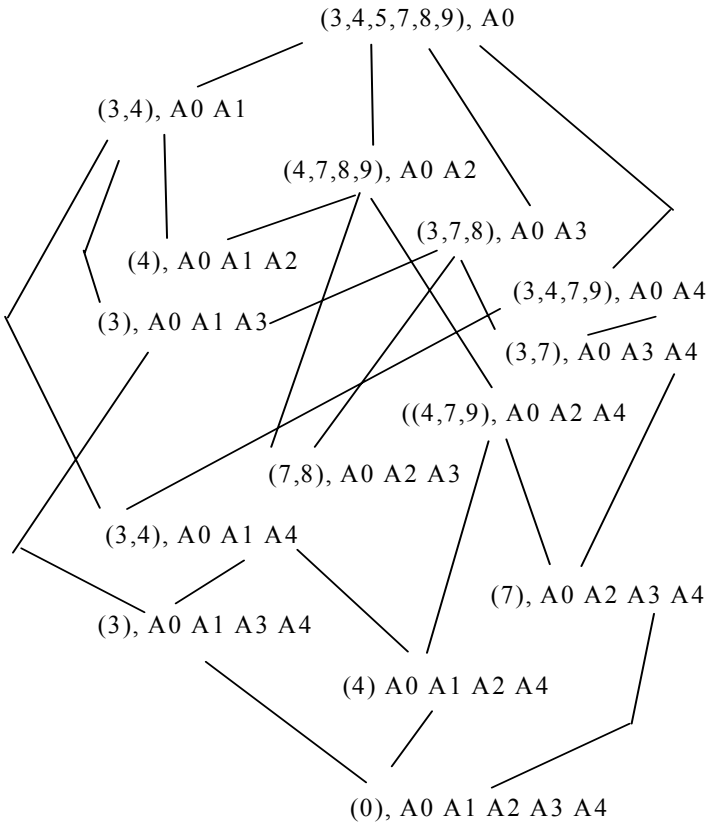
Obr. 2. Příklad relace R na kontextu vytěžování dat

Získání *pravidel* z daného Hassova diagramu je procedurálně velmi jednoduché. Pravidla ovšem nemusejí pokrývat celou databázi (celý vybraný fragment databáze). Proto ke každé množině pravidel získané z Hassova diagramu jsou přiloženy jisté záruky jejich věrohodnosti a platnosti, vzhledem k dané databázi. Mezi nejjednodušší způsoby, jak tyto záruky věrohodnosti a platnosti vyjádřit, patří výpočet dvou limitujících veličin, označovaných v [3] jako *podpora* (*support* - **Supp**) a *spolehlivost* (*confidence* - **Conf**).
Uvažujeme množinu situací $\boldsymbol{S}$ a množinu monitorovacích tříd $\boldsymbol{A}$ (obě množiny jsou předpokládány konečné). Asociační pravidlo je výraz $A_i \Rightarrow A_j$, kde $A_i, A_j \in \boldsymbol{A}$, $A_i, A_j \neq \emptyset$, $A_i \cap A_j = \emptyset$.
Pravidlo $A_i \Rightarrow A_j$ znamená: "každá situace *s,* která je indikována (**Ict**) monitorovací třídou A_i (**Ict**(*s*, A_i)), je indikována také třídou A_j, (**Ict**(*s*, A_j))".
Podporu a *spolehlivost* pravidla $A_i \Rightarrow A_j$ vyjadřujeme následovně:

$$\text{supp}(A_i, \boldsymbol{S}) = (\#(s \in \boldsymbol{S} \mid \mathbf{Ict}(s, A_i))/ (\#(\boldsymbol{S})), \tag{3}$$

$$\mathbf{Supp}\ (A_i \Rightarrow A_j, \boldsymbol{S}) = \text{supp}(A_i \cup A_j, \boldsymbol{S}), \tag{4}$$

$$\mathbf{Conf}\ (A_i \Rightarrow A_j, \boldsymbol{S}) = \mathbf{Supp}\ (A_i \Rightarrow A_j, \boldsymbol{S}) / \text{supp}(A_i). \tag{5}$$



Obr.3. Hassův diagram sestrojený z matice $\mathbf{M_G}$.

Systém pravidel sestrojený z Hassova diagramu na obr. 3 s vypočítanými hodnotami **Supp** a **Conf** je uveden v tabulce Tab.1. (Nejsou uvedena pravidla se **Supp** nebo **Conf** = 0.)

3.2 *Rough množiny*

Původní definice *rough množiny* náleží Pawlakovi [4] a dostaneme se k ní přes následující posloupnost pojmů:

Definice 2: *Relace nerozlišitelnosti.* Mějme universum prvků U, množinu atributů A a množiny hodnot V_{ai}, kterých mohou atributy a_i z množiny A, nabývat ($V = \cup V_{ai}$) .

Uvažujeme funkci g: $U \times A \rightarrow V$, která zajišťuje hodnoty atributů pro prvky universa U. Pomocí funkce g je zavedena *relace nerozlišitelnosti* RE(A) (vzhledem k atributu a_i) následovně :

$$x_1, x_2 \in U,\ (x_1\ \text{RE}(A)\ x_2) \Leftrightarrow (g(x_1, a_i) = g(x_2, a_i)). \tag{6}$$

Pawlak dále zavedl pojem "**informačního systému**" jako dvojice $\langle U, A\rangle$ (s implicitně existující funkcí g: $U \times A \rightarrow V$, (viz. výše)).
Pomocí relace nerozlišitelnosti RE(A) a pomocí základních operací na množině U byly dále definovány následující konstrukce a míry. Většina z nich je zavedena pro řešení následujícího problému " *Které prvky universa U a s jakou jistotou (vyjádřeno kvalitativně), aproximují podmnožinu* $X \subset U$, *která nás zajímá* ".

Tab.1.

Pravidlo No. i	**Pravidlo r_i**	**Supp(r_i)**	**Conf (r_i)**
1	$A_1 \Rightarrow A_2$	0.2	0.5
2	$A_1 \Rightarrow A_3$	0.2	0.5
3	$A_1 \Rightarrow A_4$	0.4	1
4	$A_2 \Rightarrow A_3$	0.4	0.5
5	$A_3 \Rightarrow A_4$	0.4	0.66
6	$A_1\ A_2 \Rightarrow A_4$	0.2	1
7	$A_2\ A_4 \Rightarrow A_4$	0.2	0.33
8	$A_2\ A_3 \Rightarrow A_4$	0.2	0.5
9	$A_2\ A_4 \Rightarrow A_3$	0.2	0.33
10	$A_1\ A_3 \Rightarrow A_4$	0.2	1
11	$A_3\ A_4 \Rightarrow A_1$	0.2	0.5

Definice 3: **Dolní aproximace** je popis objektů, o nichž lze s jistotou tvrdit, že náležejí do podmnožiny. Dolní aproximace bývá někdy nazývána pozitivní oblastí $\mathbf{Posi_{RE}}(X)$.

$$\mathbf{Posi_{RE}}(X) = \cup \{Y \mid (Y \in (U/RE)) \text{ AND } (Y \subseteq X)\}, \tag{7}$$

kde U/RE je *faktorová množina* zkonstruovaná na U podle relace RE(A).

Definice 4: **Horní aproximace** je množina prvků z U, které mohou (možná) patřit do X. Označuje se $\mathbf{Poss_{RE}}(X)$ (possibility) a je definována následovně :

$$\mathbf{Poss_{RE}}(X) = \cup \{Y \mid (Y \in U/RE) \text{ AND } (Y \cap X \neq \varnothing)\}. \tag{8}$$

Definice 5: Množinový rozdíl mezi horní a dolní aproximací X se nazývá **hraniční množina** $\mathbf{Bound_{RE}}(X)$

$$\mathbf{Bound_{RE}}(X) = \mathbf{Poss_{RE}}(X) - \mathbf{Posi_{RE}}(X). \tag{9}$$

Definice 6: **Rough množina** (hrubá, aproximovaná množina) je podmnožina X universa U, která je definována pomocí horní a dolní aproximace ($\mathbf{Poss_{RE}}(X)$, $\mathbf{Posi_{RE}}(X)$) a pro kterou platí

$$\mathbf{Bound_{RE}}(X) \neq \varnothing. \tag{10}$$

Pomocí pojmu *rough množiny* lze definovat přibližnou přesnost α_{RE}, s jakou nalezená aproximace reprezentuje vybranou množinu X:

$$\alpha_{RE}(X) = \mathbf{card}\ (\mathbf{Posi}_{RE}(X)) / \mathbf{card}\ (\mathbf{Poss}_{RE}(X)). \quad (11)$$

***Pokračování příkladu 1**:* Uvažujme nyní množinu X, která je podmíněna pravou stranou pravidel: P = A_4. Množina **X** je reprezentována příslušnými řádky tabulky X = {3, 5, 6, 7, 8, 10}. Zajímá nás, jak tuto množinu bude aproximovat množina podmíněná levou stranou pravidel L =A_3.
Tzn., že naše relace nerozlišitelnosti je v daném případě definována výrazem

$RE1(A) = \{y|\ y, x \in U \mid (y\ \ RE1(A)\ x)) \Leftrightarrow (g(y, L)) = g(x, L)) = A_3.)\}$.

Na základě vztahů (7) - (9) můžeme nyní formovat příslušné množiny a míry :

$\mathbf{Posi}_{RE1}(X) = \{5, 8, 10\}$, $\mathbf{Poss}_{RE1}(X) = \{5, 8, 10, 11\}$.

$\mathbf{Bound}_{RE1}(X) = \mathbf{Poss}_{RE1}(X) - \mathbf{Posi}_{RE1}(X) = \{5, 8, 10, 11\} - \{5, 8, 10\} = 11$.

$\alpha_{RE1}(X) = \mathbf{card}\ (\mathbf{Posi}_{RE1}(X))\ /\ \mathbf{card}\ (\mathbf{Poss}_{RE1}(X)) = 3/4 = 0.75$.

Pravá strana pravidel $\mathbf{A_4}$ je tedy aproximována levou stranou pravidel $\mathbf{A_3}$ ze 75 %.

4. Závěr

V předloženém článku byla naznačena aplikace dvou metod vytěžování znalostí z databází. Jako datová podpora byly použity podmínky z databáze monitorovacího systému. Přínos článku je i v tom, že ukazuje, do jaké formy je nutno atributy a položky databáze transformovat, aby bylo možno použít uvedené metody vytěžování. Úspěšnost vytěžování ovšem závisí na definování monitorovacích tříd a obecného tvaru „provozních“ situací.

Poděkování
Předložený článek náleží do výzkumu podporovaného grantovým projektem č. 2B06023.

Literatura

[1] M. Kravčik, J. Pokorný, J. Kohutiar, M. Kováč a E. Tóth: *Water for the Recovery of the Climate. A New Water Paradigm*. Typopress-publishing house s.r.o., Košice. 2008.

[2] J. Pokorný, V. Jirka, L. Pechar, J. Bíla, M. Hofreiter, R. Petrová, J. Zicha, F. Kobrzek a J. Mareček.: Dílčí zpráva k projektu 2B 06023. MŠMT, Česká republika, 2008.

[3] Z. Yi, S. Jianliang, Da Ruan and S. Pengfei: Interesting Rough Lattice-based Implication Rules Discovery. (Da Ruan, J.Kacprzyk, M.Fedrizzi (Eds..)), *Soft Computing for Risk Evaluation and Management,* Heidelberg, Springer-Verlag, 2001, pp. 155-169.

[4] Z. Pawlak: *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers. 1991.

[5] J. Bila and J. Jura.: Fuzzy Concepts in the Detection of Unexpected Situations. *Acta Polytechnica.* Vol. 47, No.1., 2007, pp. 5-8.