Rozdział 6

Indukcja reguł 2U

6.1 Wstęp

Siła ekspresji reguł 2U ma dwa niezależne źródła. Pierwszym z nich są współczynniki wiarygodności irf i grf, gwarantujące możliwość wyrażania niepewności drugiego rzędu. Drugie źródło to uogólniona postać formuł atomowych, dopuszczających stosowanie wartości zbiorowych (w miejsce wartości pojedynczych) oraz operatorów relacyjnych działających na tych wartościach. Potrzeba posługiwania się niepewnością drugiego rzędu została uzasadniona w podrozdziale 3.9. Czy równie celowe jest rozszerzenie postaci formuł atomowych? Także i na to pytanie należy odpowiedzieć twierdząco. Wystarczy wskazać na przydatność takich formuł w medycznych systemach wspomagania decyzji, czy systemach monitorowania i nadzoru. Przykładowo, przy ocenie postępowania leczniczego podejmowanego w ataku astmy oskrzelowej niezbędna jest znajomość nie jednej, lecz wszystkich istotnych chorób współtowarzyszących, w szczególności – cukrzycy i schorzeń kardiologicznych. Sterowanie pracą linii produkcyjnych wymaga natomiast ciągłego testowania parametrów technologicznych pod kątem ich przynależności do wymaganych zakresów wartości. I w jednym, i w drugim wypadku, do reprezentacji danych potrzebne są zbiory wartości.

Aby móc w pełni skorzystać z możliwości oferowanych przez reguły 2U, należy zapewnić efektywność i wydajność projektowanego systemu PRS(2U). Przedstawione w podrozdziale 4.5 rozważania na temat jakości systemu regułowego z niepewnością mają charakter ogólny i abstrahują od szczegółowej metody reprezentacji wiedzy w systemie. Zaproponowana tam metryka jakości 4.11 ma więc odniesienie także do systemu PRS(2U). Z tego wynika, że przy konstrukcji bazy wiedzy systemu PRS(2U) należy zmierzać do pozyskania możliwie największego, wewnętrznie niesprzecznego i nieredundantnego, zbioru reguł 2U cechujących się wysoką wiarygodnością grf. W tym celu trzeba proces indukcji reguł oprzeć na dużym zbiorze danych trenujących. Zważywszy fakt, że ten proces ma być zautomatyzowany, należy dodatkowo zapewnić jednorodność powyższego zbioru.

Realizację wymienionych wymagań gwarantują dane atrybutowe i uniwersalny format reprezentacji tych danych. Mowa tu o formacie danych zbiorczych, który nadaje się także do reprezentacji danych indywidualnych. Jeśli dodatkowo wziąć pod uwagę możliwość mapowania schematów danych atrybutowych, to – przy użyciu tego formatu – ujednorodnienie zbioru dziedzinowych danych atrybutowych staje się realne.

6.2 Dane atrybutowe zbiorcze

6.2.1 Składnia i semantyka danych

Niech $\mathcal{D}_{\mathcal{U}}$, $\mathcal{T}_{\mathcal{DU}}$, $C_{\mathcal{TDU}} = \{\mathrm{C_{u1}}, \mathrm{C_{u2}}, \ldots, \mathrm{C_{un}}\}$, $\mathbf{A_S} = \{\mathrm{A_{s1}}, \mathrm{A_{s2}}, \ldots, \mathrm{A_{sn}}\}$ oznaczają, odpowiednio: dziedzinę problemową, wybraną konceptualizację dziedziny, skończony zbiór kategorii użytych w tej konceptualizacji i odpowiadający mu skończony zbiór atrybutów zbiorowych, tak jak w podrozdziale 5.2.

Definicja 6.1. Schematem danych atrybutowych (indywidualnych) ze zbiorami kwalifikowanymi w dziedzinie $\mathcal{D}$, zgodnym z konceptualizacją $\mathcal{T}_{\mathcal{DU}}$, nazywamy dowolny niepusty podzbiór $\mathbf{S_S} = \{\mathrm{A}_{si1}, \mathrm{A}_{si2}, \ldots, \mathrm{A}_{sim}\}$ zbioru atrybutów $\mathbf{A_S}$.

Definicja 6.2. Daną atrybutową (indywidualną) ze zbiorami kwalifikowanymi, zbudowaną według schematu $\mathbf{S_S} = \{\mathrm{A}_{si1}, \mathrm{A}_{si2}, \ldots, \mathrm{A}_{sim}\}$, nazywa się zbiór par uporządkowanych w postaci:

$$\{(\mathrm{A}_{si1}, n_{i1} vq_{i1}), (\mathrm{A}_{si2}, n_{i2} vq_{i2}) \ldots, (\mathrm{A}_{sim}, n_{im} vq_{im})\} \tag{6.1}$$

w którym vq_{ik} $(1 \leq k \leq m)$ oznacza dowolny kwalifikowany zbiór wartości – przykładów kategorii $C_{\mathrm{u}ik}$, a n_{ik} $(1 \leq k \leq m)$ – opcjonalny operator negacji ($n_{ik} \in \{\varepsilon, \neg\}$). Do zapisu powyższego zbioru stosuje się uproszczoną notację krotkową $\langle \mathrm{A}_{si1}: n_{i1} vq_{i1};\ \mathrm{A}_{si2}: n_{i1} vq_{i1};\ \ldots; \mathrm{A}_{sim}: n_{im} vq_{im} \rangle$ z dowolnym porządkiem wyszczególniania par $(\mathrm{A}_{sik}, n_{ik} vq_{ik})$.

Formalny opis danych atrybutowych ze zbiorami kwalifikowanymi jest możliwy na gruncie logiki atrybutowej ALSV(FD) (ang. *Attributive Logic with Set Values*) [103]. Tak jak w przypadku logik atrybutowych AAL i VAAL, alfabet tej logiki jest sumą rozłącznych zbiorów **O**, **A** i **W**, zawierających, odpowiednio: nazwy obiektów, nazwy atrybutów i wartości atrybutów. Jeśli przyjąć, że zbiór nazw atrybutów ma postać $\mathbf{A} = \{\mathrm{A}_1, \mathrm{A}_2, \ldots, \mathrm{A}_n\}$, a $\mathbf{W} = \mathbf{W_1} \cup \mathbf{W_2} \cup \ldots \cup \mathbf{W_n}$ jest sumą dziedzin wartościowania poszczególnych atrybutów ze zbioru **A**, to w logice ALSV(FD) każdy atrybut A_i należy postrzegać jako funkcję: $\mathrm{A}_i: \mathbf{O} \rightarrow \mathbf{2^{W_i}} \times \{=, \in, \subseteq, \supseteq, \sim\}$, stanowiącą odpowiednik bazodanowego operatora projekcji.

Formułą atomową logiki ALSV(FD) jest każde i tylko takie wyrażenie, które ma postać: $\mathrm{A}_i(o)\ op\ v_i$ lub $\neg(\mathrm{A}_i(o)\ op\ v_i)$, gdzie $\mathrm{A}_i \in \mathbf{A}$, $o \in \mathbf{O}$, $v_i \in \mathbf{2^{W_i}}$,

a $op \in \{=,\ \in, \subseteq, \supseteq, \sim\}$. Formuły złożone logiki ALSV(FD) powstają przez zastosowanie do formuł atomowych spójników: koniunkcji ∧, alternatywy ∨, implikacji ⇒ i równoważności ⇔ oraz kwantyfikatorów: ogólnego ∀ i szczegółowego ∃.

Danej atrybutowej 6.1 odpowiada na gruncie logiki ALSV(FD) formuła:

$$n_{i1}(\mathrm{A}_{si1}(o)\ op_{i1}\ v_{i1}) \wedge n_{i2}(\mathrm{A}_{si2}(o)\ op_{i2}\ v_{i2})\ \ldots n_{im}(\mathrm{A}_{sim}(o)\ op_{im}\ v_{im}) \quad (6.2)$$

w której postać operatora op_{ik} $(1 \le k \le m)$ zależy od postaci odpowiadającego mu kwalifikatora q_{ik}:

– jeśli kwalifikator zbioru jest równy $\oplus$, to operator w formule ma postać $\in$,

– jeśli kwalifikator zbioru jest równy $\odot$, to ma on postać $\supseteq$.

Z powyższego wynika, że do interpretacji danych atrybutowych ze zbiorami kwalifikowanymi wystarczą te formuły atomowe ALSV(FD), które są formułami klasycznymi w myśl definicji 5.2.

Definicja 6.3. Schematem danych atrybutowych zbiorczych w dziedzinie $\mathcal{D}$, zgodnym z konceptualizacją $\mathcal{T}_{\mathcal{DU}}$, nazywamy dowolną parę uporządkowaną $(\mathbf{S_{S1}},\ \mathbf{S_{S2}})$ podzbiorów $\mathbf{S_{S1}}$ i $\mathbf{S_{S2}}$ zbioru atrybutów $\mathbf{A_S}$, takich że:

– $\mathbf{S_{S2}} \neq \emptyset$,

– $\mathbf{S_{S1}} \cap \mathbf{S_{S2}} = \emptyset$.

Definicja 6.4. Daną atrybutową zbiorczą zbudowaną według schematu $(\mathbf{S_{S1}},\ \mathbf{S_{S2}})$, gdzie $\mathbf{S_{S1}} = \{\mathrm{A}_{si1}, \ldots, \mathrm{A}_{sim1}\}$ i $\mathbf{S_{S2}} = \{\mathrm{A}_{sj1}, \ldots, \mathrm{A}_{sjm2}\}$, nazywa się zbiór par uporządkowanych w postaci:

$$\left\{ \begin{matrix} (\mathrm{A}_{si1},\ n_{i1}vq_{i1}|d_i),\ \ldots,\ (\mathrm{A}_{sim1}, n_{im1}vq_{im1}|d_i), \\ (\mathrm{A}_{sj1}, n_{j1}vq_{j1}|d_{j1}), \ldots, (\mathrm{A}_{sjm2}, n_{jm2}vq_{jm2}|d_{jm2}) \end{matrix} \right\}, \quad (6.2)$$

w którym vq_{ik} $(1 \le k \le m_1)$ oraz vq_{jl} $(1 \le l \le m_2)$ oznaczają dowolne kwalifikowane zbiory wartości – przykładów kategorii, odpowiednio C_{ik} oraz C_{jl}, n_{ik} $(1 \le k \le m_1)$ oraz n_{jl} $(1 \le l \le m_2)$ – opcjonalne operatory negacji $(n_{ik}, n_{jl} \in \{\varepsilon, \neg\})$, a $d_i \in \mathbf{N}$ i d_{jl} $(1 \le l \le m_2) \in (\mathbf{N} \cup \{\mathbf{0}\})$ – liczności odnoszące się do tych zbiorów. Powyższe liczności muszą spełniać warunek: $\forall d_{jl}(1 \le l \le m_2)(d_{jl} \le d_i)$.

W dalszym ciągu, do zapisu powyższego zbioru stosuje się uproszczoną notację krotkową:

$$\begin{aligned} < \mathrm{A}_{si1}&: n_{i1}vq_{i1}|d_i; \ldots;\ \mathrm{A}_{sim1}: n_{im1}vq_{im1}|d_i; \\ &\mathrm{A}_{sj1}: n_{j1}vq_{j1}|d_{j1} \ldots;\ \mathrm{A}_{sjm2}: n_{jm2}vq_{jm2}|d_{jm2}\ >, \end{aligned} \quad (6.3)$$

z dowolnym porządkiem wyszczególniania par w pierwszej części krotki (pary z indeksem i, w postaci $\mathrm{A}_{sik}: n_{ik}vq_{ik}|d_i$) oraz dowolnym porządkiem wyszczególniania par w drugiej części krotki (pary z indeksem j, w postaci $\mathrm{A}_{sjl}: n_{jl}vq_{jl}|d_{jl}$). Atrybuty z indeksem i nazywają się kluczowymi, a atrybuty z

indeksem j – niekluczowymi. Dla zapewnienia przejrzystości zapisu, niezależnie od postaci kwalifikowanych zbiorów wartości, wszystkie pary uporządkowane są specyfikowane jawnie.

Dla określenia sekwencji w postaci $n\ vq|d$, gdzie n oznacza potencjalny operator negacji, $n \in \{\varepsilon, \neg\})$, vq – zbiór kwalifikowany, a d – liczność odnoszącą się do tego zbioru, $d \in (\mathbf{N} \cup \{\mathbf{0}\})$, używa się w dalszym ciągu terminu „zbiór kwalifikowany z licznością".

Dana atrybutowa zbiorcza w postaci 6.3 „mieści" w sobie d_i danych atrybutowych indywidualnych ze zbiorami kwalifikowanymi, w postaci:

$$< \mathrm{A}_{si1}: n_{i1k} vq_{i1k}; \ldots;\ \mathrm{A}_{sim1}: n_{im1k} vq_{im1k};\ \mathrm{A}_{sj1}: n_{j1k} vq_{j1k};\ \ldots; \mathrm{A}_{sjm2}: n_{jm2k} vq_{jm2k} > \tag{6.4}$$

gdzie $1 \leq k \leq d_i$, spełniających zależności:

– $\forall(1 \leq l \leq m_1)\ (\mathrm{A}_{sil} \blacksquare_{il} vq_{il}) \leq_{\mathrm{AS}} (\mathrm{A}_{sil} \blacksquare_{ilk} vq_{ilk})$, gdzie $\blacksquare_{il}$ ($\blacksquare_{ilk}$) oznacza symbol relacyjny $=$, gdy n_{il} (n_{ilk}) jest operatorem pustym ε, lub symbol relacyjny $\neq$, gdy n_{il} (n_{ilk}) jest operatorem negacji $\neg$;

– $\forall(1 \leq l \leq m_2)\ \left(vq_{jl} =_{id} vq_{jlk}\right)$, gdzie $=_{id}$ jest symbolem relacji identyczności,

– $\forall(1 \leq l \leq m_2)\left(\sum_{k=1}^{d_i}\left(n_{jl} =_{id} n_{jlk}\ ?\ 1 : 0\right) = d_{jl}\right)$, gdzie $\left(n_{jl} =_{id} n_{jlk}\ ?\ 1 : 0\right)$ oznacza operator warunkowy dający w wyniku:

  – 1, gdy wartością wyrażenia relacyjnego $n_{jl} =_{id} n_{jlk}$ jest true,

  – 0, w przeciwnym wypadku.

Podsumowując, daną atrybutową zbiorczą można uzyskać przez złączenie takich danych (indywidualnych) ze zbiorami kwalifikowanymi, które są: zbudowane według tego samego schematu, parami do siebie „podobne" na odpowiadających sobie atrybutach kluczowych (podobieństwo definiowane za pomocą relacji $\leq_{\mathrm{AS}}$), oraz identyczne lub różniące się tylko występowaniem operatora negacji na odpowiadających sobie atrybutach niekluczowych.

Dane atrybutowe zbiorcze uzyskuje się w wyniku rozmaitych badań statystycznych, prowadzonych na wyselekcjonowanych grupach podobnych obiektów (zwłaszcza – osób) w celu zweryfikowania/ sfalsyfikowania pewnych ich właściwości, stanów lub zachowań. Przykładem są szeroko rozpowszechnione badania kliniczne, których celem jest zbadanie skuteczności określonych terapii farmakologicznych w leczeniu wskazanych chorób lub zaburzeń. Dobór obiektów do tych badań odbywa się w sposób rygorystyczny: każda cecha kluczowa obiektu, reprezentowana atrybutem kluczowym danej zbiorczej, musi mieć ustaloną, z góry narzuconą wartość, np. określona płeć pacjenta, wiek pochodzący z zadanego przedziału wiekowego, te same choroby główne i towarzyszące. Atrybuty niekluczowe reprezentują cechy potencjalnie różnicujące. Dla każdej z nich określa się pewną charakterystyczną wartość, której występowanie jest przed-

miotem badania klinicznego. Liczność odpowiedniego kwalifikowanego zbioru wartości (d_{jl}, $1 \leq l \leq m_2$, dla zbiorów wartości wewnątrz danej w postaci 6.2) określa liczbę tych spośród wszystkich poddanych badaniu obiektów (d_i dla danej w postaci 6.2), które osiągnęły tę charakterystyczną wartość. Przykłady danych atrybutowych zbiorczych uzyskanych w wyniku badań klinicznych można znaleźć w rozdziale 8.

Wiarygodne dane atrybutowe zbiorcze mają wyjątkową wartość. Jeśli taka dana reprezentuje odpowiednio dużą liczbę danych indywidualnych, to na jej podstawie można zaprojektować regułę 2U/ zbiór reguł 2U cechujących się odpowiednią wiarygodnością. Z kolei, jeśli liczba danych indywidualnych jest zbyt mała, to można najpierw podjąć próbę integracji danej atrybutowej zbiorczej z innymi podobnymi do niej danymi, a sam proces indukcji reguł 2U przeprowadzić w odniesieniu do danej zintegrowanej. Powyższy proces integracji danych atrybutowych zbiorczych opiera się na wielokrotnym wykonywaniu operacji sumy i iloczynu, zdefiniowanych w podrozdziale 6.2.2.

6.2.2 Algebra danych

Danym atrybutowym zbiorczym można nadać interpretację algebraiczną. Formalną definicję algebry takich danych (def. 6.10) poprzedza szereg definicji pomocniczych i lematów.

Definicja 6.5. Niech $v_i q_i$ oraz $v_j q_j$ oznaczają dwa zbiory kwalifikowane w postaci, odpowiednio: $v_i q_i = \{v_{i1}, v_{i2}, \dots, v_{iki}\} q_i$ i $v_j q_j = \{v_{j1}, v_{j2}, \dots, v_{jkj}\} q_j$, gdzie v_{il} $(1 \leq l \leq k_i)$, v_{jl} $(1 \leq l \leq k_j) \in \mathbf{W}$, a q_i, $q_j \in \{\oplus, \odot\}$. Zbiór kwalifikowany $v_i q_i$ jest podzbiorem zbioru kwalifikowanego $v_j q_j$, $v_i q_i \subseteq_q v_j q_j$, wtedy i tylko wtedy, gdy:

– $q_i = q_j = \oplus$ i $\{v_{j1}, v_{j2}, \dots, v_{jkj}\} \subseteq \{v_{i1}, v_{i2}, \dots, v_{iki}\}$ lub
– $q_i = q_j = \odot$ i $\{v_{i1}, v_{i2}, \dots, v_{iki}\} \subseteq \{v_{j1}, v_{j2}, \dots, v_{jkj}\}$ lub
– $q_i = \oplus$ i $q_j = \odot$ i $k_i = k_j = 1$ i $\{v_{i1}\} = \{v_{j1}\}$,

gdzie $\subseteq$ oznacza klasyczny teoriomnogościowy operator zawierania się zbiorów.

Ponadto, zanegowany zbiór kwalifikowany $\neg v_i q_i$ jest podzbiorem zanegowanego zbioru kwalifikowanego $\neg v_j q_j$, $\neg v_i q_i \subseteq_q \neg v_j q_j$, wtedy i tylko wtedy, gdy:

– $q_i = q_j = \oplus$ i $\{v_{i1}, v_{i2}, \dots, v_{iki}\} \subseteq \{v_{j1}, v_{j2}, \dots, v_{jkj}\}$ lub
– $q_i = q_j = \odot$ i $\{v_{j1}, v_{j2}, \dots, v_{jkj}\} \subseteq \{v_{i1}, v_{i2}, \dots, v_{iki}\}$ lub
– $q_i = \odot$ i $q_j = \oplus$ i $k_i = k_j = 1$ i $\{v_{i1}\} = \{v_{j1}\}$.

Relacja bycia podzbiorem $\subseteq_q$ nie zachodzi pomiędzy żadnymi zbiorem kwalifikowanym i zanegowanym zbiorem kwalifikowanym.

Definicja 6.6. Niech $\mathbf{Q_W}$ oznacza zbiór wszystkich kwalifikowanych zbiorów z licznością dających się skonstruować w odniesieniu do dziedziny $\mathbf{W}$, $\mathbf{Q_W} \stackrel{\text{def}}{=} \stackrel{\text{def}}{=} \{\varepsilon, \neg\} \times \mathbf{2^W} \times \{\oplus, \odot\} \times (\mathbf{N} \cup \{0\})$, a $n_j v_j q_j | d_j$ i $n_k v_k q_k | d_k$ oznaczają dowolne elementy tego zbioru. Zbiór z licznością $n_j v_j q_j | d_j$ jest podzbiorem zbioru z licznością $n_k v_k q_k | d_k$, $n_j v_j q_j | d_j \subseteq_{\mathbf{W}} n_k v_k q_k | d_k$, wtedy i tylko wtedy, gdy istnieje ciąg niezanegowanych bądź zanegowanych zbiorów kwalifikowanych $n_{k1} v_{k1} q_{k1}$, …, $n_{kl} v_{kl} q_{kl}$, gdzie $v_{k1}, \ldots, v_{kl} \in \mathbf{W}$, spełniających zależność:

$n_j v_j q_j \subseteq_q n_{k1} v_{k1} q_{k1} \subseteq_q \ldots \subseteq_q n_{kl} v_{kl} q_{kl} \subseteq_q n_k v_k q_k$.

Jak widać, liczności zbiorów kwalifikowanych nie mają wpływu na zachodzenie relacji $\subseteq_{\mathbf{W}}$. Z definicji 6.5 i 6.6 wynika oczywisty lemat 6.1.

Lemat 6.1. Relacja $\subseteq_{\mathbf{W}}$ jest relacją porządku częściowego.

Definicja 6.7. Niech $\mathbf{Q_W}$ oznacza zbiór kwalifikowanych zbiorów z licznością, o znaczeniu jak w definicji 6.6, a $n_i v_i q_i | d_i$ i $n_j v_j q_j | d_j$ oznaczają dwa dowolne zbiory kwalifikowane z licznością pochodzące z $\mathbf{Q_W}$. Sumą $\cup_{\mathbf{W}}$ i iloczynem $\cap_{\mathbf{W}}$ zbiorów kwalifikowanych z licznością nazywa się rozszerzenia klasycznych operacji teoriomnogościowych $\cup$ i $\cap$, zdefiniowane jak następuje:

$n_i v_i q_i | d_i \cup_{\mathbf{W}} n_j v_j q_j | d_j = n_k v_k q_k | d_k$, gdzie

– $v_k = v_i \cap v_j,\ q_k = \oplus,\ n_k = \varepsilon,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \oplus$ i $n_i = n_j = \varepsilon$,

– $v_k = v_i \cup v_j,\ q_k = \odot,\ n_k = \varepsilon,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \odot$ i $n_i = n_j = \varepsilon$,

– $v_k = v_i \cup v_j,\ q_k = \oplus,\ n_k = \neg,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \oplus$ i $n_i = n_j = \neg$,

– $v_k = v_i \cap v_j,\ q_k = \odot,\ n_k = \neg,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \odot$ i $n_i = n_j = \neg$,

– $v_k = \{\},\ q_k = \oplus,\ n_k = \varepsilon,\ d_k = d_i + d_j,$
jeśli $q_i \neq q_j$ lub $n_i \neq n_j$;

$n_i v_i q_i | d_i \cap_{\mathbf{W}} n_j v_j q_j | d_j = n_l v_l q_l | d_l$, gdzie

– $v_k = v_i \cup v_j,\ q_k = \oplus,\ n_k = \varepsilon,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \oplus$ i $n_i = n_j = \varepsilon$,

– $v_k = v_i \cap v_j,\ q_k = \odot,\ n_k = \varepsilon,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \odot$ i $n_i = n_j = \varepsilon$,

– $v_k = v_i \cap v_j,\ q_k = \oplus,\ n_k = \neg,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \oplus$ i $n_i = n_j = \neg$,

– $v_k = v_i \cup v_j,\ q_k = \odot,\ n_k = \neg,\ d_k = d_i + d_j,$
jeśli $q_i = q_j = \odot$ i $n_i = n_j = \neg$;

– $v_k = \{\},\ q_k = \odot,\ n_k = \varepsilon,\ d_k = d_i + d_j,$
jeśli $q_i \neq q_j$ lub $n_i \neq n_j$.

Lemat 6.2. Operacje sumy $\cup_{\mathbf{W}}$ i iloczynu $\cap_{\mathbf{W}}$ spełniają następujące zależności:

$$\forall\left(n_i v_i q_i|d_i, n_j v_j q_j|d_j \in \mathbf{Q_W}\right) =$$

$$= \left(n_i v_i q_i|d_i \ \cup_{\mathbf{W}}\ n_j v_j q_j|d_j = \sup\left(n_i v_i q_i|d_i, n_j v_j q_j|d_j\right)\right)$$

$$\forall\left(n_i v_i q_i|d_i, n_j v_j q_j|d_j \in \mathbf{Q_W}\right) =$$

$$= \left(n_i v_i q_i|d_i \ \cap_{\mathbf{W}}\ n_j v_j q_j|d_j = \inf\left(n_i v_i q_i|d_i, n_j v_j q_j|d_j\right)\right)$$

Prostą konsekwencją lematów 6.1 i 6.2 jest twierdzenie 6.1.

Twierdzenie 6.1. Algebra $\mathcal{A}_{\mathbf{W}} = (\mathbf{W}, \subseteq_{\mathbf{W}}, \cup_{\mathbf{W}}, \cap_{\mathbf{W}})$ jest kratą.

Relację porządku częściowego $\subseteq_{\mathbf{W}}$, operację sumy $\cup_{\mathbf{W}}$ i operację iloczynu $\cap_{\mathbf{W}}$ określone na zbiorze kwalifikowanych zbiorów z licznością $\mathbf{Q_W}$ można rozszerzyć na zbiór danych atrybutowych zbiorczych zbudowanych według schematu $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$.

Definicja 6.8. Niech $\mathbf{E_{S1,S2}}$ oznacza zbiór danych atrybutowych zbiorczych zbudowanych według schematu $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$ (def. 6.3), a e_f oraz e_g oznaczają dwie dane z tego zbioru, w poniższej postaci:

$$e_f = < \mathrm{A}_{si1}: n_{fi1} v q_{fi1}|d_{fi}; \ldots;\ \mathrm{A}_{sim1}: n_{fim1} v q_{fim1}|d_{fi};$$
$$\mathrm{A}_{sj1}: n_{fj1} v q_{fj1}|d_{fj1} \ldots;\ \mathrm{A}_{sjm2}: n_{fjm2} v q_{fjm2}|d_{fjm2} > \qquad (6.5)$$

$$e_g = < \mathrm{A}_{si1}: n_{gi1} v q_{gi1}|d_{gi}; \ldots;\ \mathrm{A}_{sim1}: n_{gim1} v q_{gim1}|d_{gi};$$
$$\mathrm{A}_{sj1}: n_{gj1} v q_{gj1}|d_{gj1} \ldots;\ \mathrm{A}_{sjm2}: n_{gjm2} v q_{gjm2}|d_{gjm2} > \qquad (6.6)$$

Dana atrybutowa zbiorcza e_f subsumuje daną atrybutową zbiorczą e_g, $e_f \subseteq_{\mathbf{S1,S2}} e_g$, wtedy i tylko wtedy, gdy spełnione są warunki:

– $\forall(1 \leq k \leq m_1)\ \left(n_{fik} v q_{fik}|d_{fk} \subseteq_{\mathbf{W}} n_{gk} v q_{gik}|d_{gk}\right)$,

– $\forall(1 \leq l \leq m_2)\ \left(n_{gil} v q_{gil}|d_{gl} \subseteq_{\mathbf{W}} n_{fil} v q_{fil}|d_{fl}\right)$.

Warunki te oznaczają, że relacja subsumcji zachodzi między parą tylko takich danych atrybutowych zbiorczych, które są:
– podobne w sensie relacji $\subseteq_{\mathbf{W}}$ na wszystkich atrybutach kluczowych,
– podobne w sensie relacji $\supseteq_{\mathbf{W}}$ na wszystkich atrybutach niekluczowych.

Z definicji 6.4 i 6.8 oraz z lematu 6.1 wynika natychmiast lemat 6.3.

Lemat 6.3. Relacja $\subseteq_{\mathbf{S1,S2}}$ jest relacją porządku częściowego.

Definicja 6.9. Niech $\mathbf{E}_{\mathbf{S1,S2}}$ oznacza zbiór danych atrybutowych zbiorczych zbudowanych według schematu $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$ (def. 6.3), a e_f oraz e_g oznaczają dwie dane z tego zbioru, w postaci 6.5 i 6.6. Sumą $\cup_{\mathbf{S1,S2}}$ i iloczynem $\cap_{\mathbf{S1,S2}}$ danych atrybutowych zbiorczych nazywa się operacje, które są zdefiniowane jak następuje:

$$e_f \cup_{\mathbf{S1,S2}} e_g = < \mathrm{A}_{si1}: n_{pi1} vq_{pi1} | d_{pi}; \ldots; \mathrm{A}_{sim1}: n_{pim1} vq_{pim1} | d_{pi}; \\ \mathrm{A}_{sj1}: n_{pj1} vq_{pj1} | d_{pj1} \ldots; \mathrm{A}_{sjm2}: n_{pjm2} vq_{pjm2} | d_{pjm2} >$$

$$e_f \cap_{\mathbf{S1,S2}} e_g = < \mathrm{A}_{si1}: n_{ri1} vq_{ri1} | d_{ri}; \ldots; \mathrm{A}_{sim1}: n_{rim1} vq_{rim1} | d_{ri}; \\ \mathrm{A}_{sj1}: n_{rj1} vq_{rj1} | d_{rj1} \ldots; \mathrm{A}_{sjm2}: n_{rjm2} vq_{rjm2} | d_{rjm2} > ,$$

gdzie

$$\forall (1 \le k \le m_1)\ n_{pik} vq_{pik} | d_{pi} = n_{fi1} vq_{fik} | d_{fi} \cap_{\mathbf{W}} n_{gi1} vq_{gik} | d_{gi} ,$$

$$\forall (1 \le l \le m_2)\ n_{pjl} vq_{pjl} | d_{pjl} = n_{fjl} vq_{fjl} | d_{fjl} \cup_{\mathbf{W}} n_{gjl} vq_{gjl} | d_{gjl} ,$$

$$\forall (1 \le k \le m_1)\ n_{rik} vq_{pik} | d_{pi} = n_{fi1} vq_{fik} | d_{fi} \cup_{\mathbf{W}} n_{gi1} vq_{gik} | d_{gi} ,$$

$$\forall (1 \le l \le m_2)\ n_{rjl} vq_{pjl} | d_{pjl} = n_{fjl} vq_{fjl} | d_{fjl} \cap_{\mathbf{W}} n_{gjl} vq_{gjl} | d_{gjl} .$$

Z definicji 6.9 i lematu 6.4 wynika następujący lemat 6.4.

Lemat 6.4. Niech $\mathbf{E}_{\mathbf{S1,S2}}$ oznacza zbiór danych atrybutowych zbiorczych zbudowanych według schematu $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$ (def. 6.3). Operacje sumy $\cup_{\mathbf{S1,S2}}$ i iloczynu $\cap_{\mathbf{S1,S2}}$ spełniają następujące zależności:

$$\forall (e_f, e_g \in \mathbf{E}_{\mathbf{S1,S2}}) \left(e_f \cup_{\mathbf{S1,S2}} e_g = \sup(e_f, e_g) \right)$$

$$\forall (e_f, e_g \in \mathbf{E}_{\mathbf{S1,S2}}) \left(e_f \cap_{\mathbf{S1,S2}} e_g = \inf(e_f, e_g) \right)$$

Definicja 6.10. Algebrą danych atrybutowych zbiorczych zbudowanych według schematu $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$ nazywa się algebrę:

$$\mathcal{A}_{\mathbf{S1,\,S2}} = \left(\mathbf{E}_{\mathbf{S1,S2}},\ \subseteq_{\mathbf{S1,S2}},\ \cup_{\mathbf{S1,S2}},\ \cap_{\mathbf{S1,S2}} \right) \tag{6.7}$$

Konsekwencją lematów 6.3 i 6.4 jest twierdzenie 6.2.

Twierdzenie 6.2. Algebra $\mathcal{A}_{\mathbf{S1,\,S2}} = \left(\mathbf{E}_{\mathbf{S1,S2}},\ \subseteq_{\mathbf{S1,S2}},\ \cup_{\mathbf{S1,S2}},\ \cap_{\mathbf{S1,S2}} \right)$ jest kratą.

6.3 Algorytmy indukcji i selekcji reguł

Indukcyjne uczenie reguł 2U odbywa się cyklicznie, w dwóch etapach. Pierwszy z nich polega na łączeniu danych atrybutowych zbiorczych według pewnego klucza. W wyniku tego łączenia, zwanego dalej integracją danych atrybutowych

zbiorczych, powstaje dana o charakterze wirtualnym. W drugim etapie, na podstawie tej danej, odbywa się właściwy proces generacji reguły 2U. Proces ten obejmuje konstrukcję przesłanki i konkluzji reguły oraz obliczanie jej współczynników wiarygodności.

Integracja danych atrybutowych zbiorczych. Idea integrowania danych ma na celu konstruowanie takich (wirtualnych) danych zbiorczych, które będą „mieścić w sobie" największą możliwą liczbę danych atrybutowych indywidualnych. W oczywisty sposób, powyższe dane indywidualne muszą być do siebie podobne w sensie formalnym. Przy spełnieniu odpowiednich ograniczeń ilościowych, dana zbiorcza może być użyta do bezpośredniej generacji reguły z niepewnością.

Niech $\mathbf{E_{S1,S2}}$ oznacza zbiór danych atrybutowych zbiorczych zbudowanych według schematu $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$, a e – dowolną daną z tego zbioru. W dalszym ciągu zakłada się, że e jest w postaci:

$$< \mathrm{A}_{si1}: n_{i1} vq_{i1} | d_i; \ldots; \mathrm{A}_{sim1}: n_{im1} vq_{im1} | d_i; \\ \mathrm{A}_{sj1}: n_{j1} vq_{j1} | d_{j1} \ldots; \mathrm{A}_{sjm2}: n_{jm2} vq_{jm2} | d_{jm2} > \tag{6.8}$$

gdzie poszczególne elementy zachowują swoje oznaczenia z definicji 6.3.

Ze względu na zakładaną interpretację danej atrybutowej zbiorczej (str. 105), daną e można by zastąpić, przy zachowaniu semantyki zbioru $\mathbf{E_{S1,S2}}$, następującym zbiorem danych atrybutowych zbiorczych z jednym atrybutem niekluczowym:

$$\left\{ \begin{array}{l} < \mathrm{A}_{si1}: n_{i1} vq_{i1} | d_i; \ldots; \mathrm{A}_{sim1}: n_{im1} vq_{im1} | d_i; \mathrm{A}_{sj1}: n_{j1} vq_{j1} | d_{j1} >, \\ < \mathrm{A}_{si1}: n_{i1} vq_{i1} | d_i; \ldots; \mathrm{A}_{sim1}: n_{im1} vq_{im1} | d_i; \mathrm{A}_{sj2}: n_{j2} vq_{j2} | d_{j2} >, \\ \qquad \ldots \\ < \mathrm{A}_{si1}: n_{i1} vq_{i1} | d_i; \ldots; \mathrm{A}_{sim1}: n_{im1} vq_{im1} | d_i; \\ \qquad\qquad \mathrm{A}_{sjm2}: n_{jm2} vq_{jm2} | d_{jm2} > \end{array} \right\} \tag{6.9}$$

Zgodnie z definicją danej atrybutowej zbiorczej (def. 6.3), wszystkie liczności odnoszące się do zbiorów powiązanych w krotce z atrybutami kluczowymi są identyczne i maksymalne w całej krotce, czyli niemniejsze od liczności odnoszących się do zbiorów powiązanych w niej z atrybutami niekluczowymi. W szczególności, dla danej atrybutowej zbiorczej w postaci 6.8 zachodzi warunek: $\forall d_{jl} (1 \leq l \leq m_2) \big(d_i \geq d_{jl}\big)$. Użyta w warunku relacja arytmetyczna ma charakter nieostry; może się więc zdarzyć, że dla jednego lub więcej atrybutów niekluczowych, krotność powiązanego z nim zbioru będzie równa krotności maksymalnej. Niech $\mathbf{A_{Si}} = \{\mathrm{A}_{sjc1}, \mathrm{A}_{sjc2}, \ldots, \mathrm{A}_{sjcz}\} \subseteq \{\mathrm{A}_{sj1}, \mathrm{A}_{sj2}, \ldots, \mathrm{A}_{sjm2}\}$ oznacza niepusty zbiór takich atrybutów, $\mathbf{A_{Si}} \neq \emptyset$. Bez konsekwencji dla poprawności semantycznej danej atrybutowej zbiorczej, każdy z atrybutów ze zbioru $\mathbf{A_{Si}}$ można przesunąć, wraz z odpowiadającym mu zbiorem kwalifikowanym, na pozycję atrybutu kluczowego danej. Liczba l_c możliwych kombinacji takich przesunięć wynosi:

– $l_c = \left(2^{|\mathbf{A_{Sj}}|} - 1\right)$, gdy $\{\mathrm{A}_{sjc1}, \mathrm{A}_{sjc2}, \ldots, \mathrm{A}_{sjcz}\} \subset \{\mathrm{A}_{sj1}, \mathrm{A}_{sj2}, \ldots, \mathrm{A}_{sjm2}\}$,

– $l_c = \left(2^{|\mathbf{A_{Sj}}|} - 2\right)$, gdy $\{\mathrm{A}_{sjc1}, \mathrm{A}_{sjc2}, \ldots, \mathrm{A}_{sjcz}\} = \{\mathrm{A}_{sj1}, \mathrm{A}_{sj2}, \ldots, \mathrm{A}_{sjm2}\}$ (zbiór atrybutów niekluczowych nie może być zbiorem pustym).

W taki sposób można powiększyć zbiór danych atrybutowych zbiorczych z jednym atrybutem niekluczowym uzyskiwanych z danej wyjściowej w postaci 6.8.

Opisany proces „rozkładu" będzie prowadzony dla wszystkich danych ze zbioru $\mathbf{E_{S1,S2}}$. W dalszej kolejności, każda uzyskana w ten sposób dana atrybutowa zbiorcza z jednym atrybutem niekluczowym zostanie zintegrowana ze wszystkimi innymi danymi podobnymi do niej w sensie relacji subsumcji szczegółowej.

Definicja 6.11. Schematem danych atrybutowych zbiorczych z jednym atrybutem niekluczowym w dziedzinie $\mathcal{D}$, zgodnym z konceptualizacją $\mathcal{T}_{\mathcal{DU}}$, nazywamy dowolną parę uporządkowaną $\left(\mathbf{S_{S1}}, \mathrm{A}_{sj}\right)$, zbudowaną z podzbioru $\mathbf{S_{S1}}$ zbioru atrybutów $\mathbf{A_S}$ i atrybutu $\mathrm{A}_{sj} \in \mathbf{A_S}$ spełniającego warunek:

– $\mathbf{S_{S1}} \cap \{\mathrm{A}_j\} = \emptyset$.

Definicja 6.12. Daną atrybutową zbiorczą z jednym atrybutem niekluczowym zbudowaną według schematu $\left(\mathbf{S_{S1}}, \mathrm{A}_{sj}\right)$, gdzie $\mathbf{S_{S1}} = \{\mathrm{A}_{si1}, \ldots, \mathrm{A}_{sim1}\}$, nazywa się zbiór par uporządkowanych w postaci:

$$\left\{(\mathrm{A}_{si1},\ n_{i1}vq_{i1}|d_i),\ \ldots,\ (\mathrm{A}_{sim1}, n_{im1}vq_{im1}|d_i), \left(\mathrm{A}_{sj}, n_j vq_j|d_j\right)\right\} \qquad (6.10)$$

w którym n_{ik}, vq_{ik} $(1 \leq k \leq m_1)$ oraz d_i zachowują oznaczenia z definicji 6.4, vq_j oznacza dowolny kwalifikowany zbiór wartości – przykładów kategorii C_j, n_j – opcjonalny operator negacji, $n_j \in \{\varepsilon, \neg\}$), a $d_j \in (\mathbf{N} \cup \{\mathbf{0}\})$ – liczność zbioru vq_j, spełniającą wymóg: $d_{jl} \leq d_i$.

W dalszym ciągu, do zapisu danych atrybutowych zbiorczych z jednym atrybutem niekluczowym używa się uproszczonej notacji krotkowej (6.3).

Definicja 6.13. Niech $\mathbf{E_{S1,}}_{\mathrm{Asj}}$ oznacza zbiór danych atrybutowych zbiorczych z jednym atrybutem niekluczowym, zbudowanych według schematu $\left(\mathbf{S_{S1}}, \mathrm{A}_{sj}\right)$ (def. 6.3), a $e1_f$ oraz $e1_g$ oznaczają dwie dane z tego zbioru, w postaci:

$$e1_f = < \mathrm{A}_{si1}: n_{fi1}vq_{fi1}|d_{fi}; \ldots; \mathrm{A}_{sim1}: n_{fim1}vq_{fim1}|d_{fi};\ \mathrm{A}_{sj}: n_{fj}vq_{fj}|d_{fj} > \qquad (6.11)$$

$$e1_g = < \mathrm{A}_{si1}: n_{gi1}vq_{gi1}|d_{gi}; \ldots; \mathrm{A}_{sim1}: n_{gim1}vq_{gim1}|d_{gi};\ \mathrm{A}_{sj}: n_{gj}vq_{gj}|d_{gj} > \qquad (6.12)$$

Dana atrybutowa zbiorcza z jednym atrybutem niekluczowym $e1_f$ subsumuje szczegółowo daną atrybutową zbiorczą z jednym atrybutem niekluczowym $e1_g$, $e1_f \subseteq_{\mathbf{S1},\mathrm{Asj}} e_g$, wtedy i tylko wtedy, gdy spełnione są warunki:

– $\forall(1 \le k \le m_1)\ (n_{fik}vq_{fik}|d_{fk} \subseteq_{\mathbf{W}} n_{gk}vq_{gik}|d_{gk})$,

– $n_{fj}vq_{fj}|d_{fj} \subseteq_{\mathbf{W}} n_{gj}vq_{gj}|d_{gj}$ i $n_{gj}vq_{gj}|d_{gj} \subseteq_{\mathbf{W}} n_{fj}vq_{fj}|d_{fj}$.

Z definicji 6.5 i 6.6 wynika, że drugi z wymienionych warunków jest równoważny żądaniu: $n_{fj} = n_{gj}$ i $vq_{fj} = vq_{fj}$. Relacja $\subseteq_{\mathbf{S1},\mathrm{Asj}}$ jest w oczywisty sposób podzbiorem relacji $\subseteq_{\mathbf{S1},\{\mathrm{Asj}\}}$.

Z definicji 6.13 i 6.8 oraz z lematu 6.1 wynika natychmiast lemat 6.5.

Lemat 6.5. Relacja $\subseteq_{\mathbf{S1},\mathrm{Asj}}$ jest relacją porządku częściowego.

Definicja 6.14. Niech $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$ oznacza schemat danych atrybutowych zbiorczych, a $(\mathbf{S_{S1\_k}}, \mathrm{A}_{sj})$, gdzie $\mathbf{S1} \subseteq \mathbf{S1\_k} \subset (\mathbf{S1} \cup \mathbf{S2})$; $\mathrm{A}_{sj} \in (\mathbf{S_{S2}} \backslash \mathbf{S_{1\_k}})$ – pochodny względem $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$ schemat danych atrybutowych zbiorczych z jednym atrybutem kluczowym.

Niech $\mathbf{E_{S1,S2}}$ oznacza zbiór danych atrybutowych zbiorczych zbudowanych według schematu $(\mathbf{S_{S1}}, \mathbf{S_{S2}})$, a $\mathbf{E_{S1\_k},Asj}$ – zbiór wszystkich danych z jednym atrybutem niekluczowym, uzyskanych przez „rozkład" danych ze zbioru $\mathbf{E_{S1,S2}}$ i zbudowanych według schematu $(\mathbf{S_{S1\_k}}, \mathrm{A}_{sj})$.

Niech $e1_f$ oznacza pewną daną ze zbioru $\mathbf{E_{S1\_k},Asj}$, a $\{e1_{g1}, e1_{g2}, \ldots, e1_{gh}\} \subseteq \mathbf{E_{S1\_k},Asj}$ – podzbiór tych wszystkich elementów ze zbioru $\mathbf{E_{S1\_k},Asj}$, które są integrowalne z $e1_f$, tzn. spełniają warunek: $e1_f \subseteq_{\mathbf{S1\_k},\mathrm{Asj}} e1_{gi}$, gdzie $1 \le i \le h$.

Daną wirtualną $e1_{f\_vir}(e1_f, \mathbf{E_{S1,S2}})$ oblicza się jako iloczyn:

$$e1_{f\_vir}(e1_f, \mathbf{E_{S1,S2}}) = e1_{g1} \cap_{\mathbf{S1},\{\mathrm{Asj}\}} e1_{g2} \cap_{\mathbf{S1},\{\mathrm{Asj}\}} \ldots \cap_{\mathbf{S1},\{\mathrm{Asj}\}} e1_{gh}$$

Generacja reguły 2U. Uzyskana w wyniku integracji dana wirtualna $e1_{f\_vir}(e1_f, \mathbf{E_{S1,S2}})$ zostanie przetransformowana do postaci reguły 2U opartej na schemacie $\mathbf{S_S} = \mathbf{S_{S1}} \cup \mathbf{S_{S2}}$ i zwanej R_{e1f}. Powyższa transformacja odbywa się według następującego schematu:

$e1_{f_{vir}}(e1_f, \mathbf{E_{S1,S2}}) =$

$= < \mathrm{A}_{si1}: n_{fi1}vq_{fi1}|d_{fi}; \ldots; \mathrm{A}_{simk}: n_{fimk}vq_{fimk}|d_{fi};\ \mathrm{A}_{sj}: n_{fj}vq_{fj}|d_{fj} >$

$\Rightarrow$

it is declared with $\mathrm{grf}(p_r)$:

$\mathrm{A}_{si1}\ \blacksquare_{fi1}\ vq_{fi1}, \ldots, \mathrm{A}_{simk}\ \blacksquare_{fimk}\ vq_{fimk} \rightarrow \mathrm{A}_{sj}\ \blacksquare_{fj}\ vq_{fj}$ with $\mathrm{irf}(p_c) =$

$= R_{e1f}$

gdzie $\blacksquare_{fil}$ $(1 \leq l \leq m_k)/\blacksquare_{fj}$ oznacza operator $=$, gdy n_{fil} /n_{fj} jest symbolem pustym (ε) oraz operator $\neq$, gdy n_{fil} /n_{fj} jest symbolem negacji (¬), a p_c i p_r oznaczają wartości współczynników wiarygodności, odpowiednio wewnętrznego irf i zewnętrznego grf.

Algorytm obliczania wartości współczynników wiarygodności. Obliczanie współczynnika irf będzie się odbywać na etapie transformacji danej wirtualnej, na podstawie następującego wzoru:

$$p_c = \frac{d_{fj}}{d_{fi}} \tag{6.13}$$

Znacznie trudniej jest zaproponować poprawną metodę wyznaczania współczynnika grf. Brak uznanych metod algorytmicznych zachęca do rozwiązania tego problemu przy użyciu technik heurystycznych. Jedną z takich propozycji przedłożono w pracy [112]. Rozwiązanie opiera się na ocenie dwóch właściwości reguły 2U: wagi i precyzji. Waga reguły zależy od mocy wirtualnej danej zbiorczej, mierzonej liczbą danych indywidualnych reprezentowanych tą daną zbiorczą, i od wyrazistości reguły, mierzonej odległością współczynnika p_c od punktów krańcowych zakresu $\langle 0; 1 \rangle$. Można ją więc wyznaczyć na etapie transformacji danej wirtualnej, podobnie jak sam współczynnik p_c. Z kolei, precyzja reguły zależy od stopnia wzajemnego podobieństwa rzeczywistych danych zbiorczych, z których uzyskano wirtualną daną zbiorczą: im mniejsze zróżnicowanie tych danych, tym większa precyzja reguły. Ten stopień należy oszacować w procesie integracji danych – po jego zakończeniu dane potrzebne do obliczeń stają się niedostępne.

W rozwiązaniu zaproponowanym w [112], wpływ obu wymienionych właściwości na współczynnik grf przyjęto wyrażać następującym wzorem:

$$p_r = \min\left(\mathrm{wg}\left(R_{e1f}\right), \mathrm{acc}\left(R_{e1f}\right)\right) \tag{6.14}$$

gdzie wg i acc są nazwami funkcji wyznaczających, odpowiednio wagę i precyzję reguły, i przyjmujących wartości z przedziału $\langle 0; 1 \rangle$. Na podstawie doświadczeń związanych z użyciem reguł 2U do wnioskowania w pewnej konkretnej dziedzinie, wzór ten można zmodyfikować lub zastąpić innym, gwarantującym lepszą efektywność i wydajność systemu.

Selekcja reguł. Celem konstrukcji reguły 2U jest jej wykorzystanie w systemie regułowym do prowadzenia wnioskowań. Nowo wygenerowaną regułę umieszcza się w bazie wiedzy tego systemu. Zanim to jednak nastąpi, należy:

- wykluczyć zachodzenie warunków poddających w wątpliwość dodatni monotoniczny charakter zależności reprezentowanej regułą (twierdz. 5.1),
- wykluczyć zachodzenie sprzeczności pomiędzy badaną regułą i dowolną z reguł znajdujących się w bazie wiedzy (def. 5.7),
- wykluczyć istnienie w bazie wiedzy reguły subsumującej badaną regułę (def. 5.8).

Wszystkie wymienione zadania mają charakter globalny. Z tego powodu, zostaną wykonane dopiero po wygenerowaniu wszystkich reguł. Najtrudniejsze z zadań, czyli badanie reguły pod kątem dodatniej monotoniczności, będzie prowadzone dwuetapowo. Pierwszy etap będzie polegał na wyselekcjonowaniu takiego, możliwie największego, podzbioru zbioru wszystkich reguł, który gwarantuje niezachodzenie zależności 5.11 i 5.12.W ramach tego etapu, zostaną wyodrębnione wszystkie takie pary zbiorów kwalifikowanych (vq_i, vq_j), których pierwszy element występuje w przesłance pewnej reguły, a drugi – w konkluzji tej reguły. Następnie, dla każdej pary, zostaną zliczone reguły, w których:

(a) oba elementy występują w regule w postaci niezanegowanej vq_i i vq_j (liczba o symbolu $n\_vq_{i+}\_vq_{j+}$),
(b) oba elementy występują w regule w postaci zanegowanej $\neg\mathrm{vq_i}$ i $\neg\mathrm{vq_j}$ (liczba o symbolu $\mathrm{n\_vq_{i-}\_vq_{j-}}$),
(c) pierwszy z elementów występuje w postaci niezanegowanej $\mathrm{vq_i}$, a drugi – w postaci zanegowanej $\neg\mathrm{vq_j}$ (liczba o symbolu $\mathrm{n\_vq_{i+}\_vq_{j-}}$)
(d) pierwszy z elementów występuje w postaci zanegowanej $\neg\mathrm{vq_i}$, a drugi – w postaci niezanegowanej $\mathrm{vq_j}$ (liczba o symbolu $\mathrm{n\_vq_{i-}\_vq_{j+}}$).

W kolejności, nastąpi sprawdzenie relacji pomiędzy liczbami $(n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-})$ i $(n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-})$. Możliwe są następujące przypadki:

– $(n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-}) \gg (n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-})$ – za poprawny należy uznać w regule układ elementów (a) i (b), za niepoprawny – układ elementów (c) i (d),
– $(n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-}) \ll (n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-})$ – za poprawny należy uznać w regule układ elementów (c) i (d), za niepoprawny – układ elementów (a) i (b),
– $(n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-}) \approx (n\_vq_{i+}\_vq_{j+} + n\_vq_{i-}\_vq_{j-})$ – poprawność poszczególnych układów elementów jest trudna do zweryfikowania.

Niech R_{e1f} oznacza wygenerowaną regułę 2U w postaci:

it is declared with grf(p_r):
$\mathrm{A_{si1}}\ \blacksquare_{fi1}\ vq_{fi1}, \ldots, \mathrm{A}_{simk}\ \blacksquare_{fimk}\ vq_{fimk} \rightarrow \mathrm{A_{sj}}\ \blacksquare_{fj}\ vq_{fj}$ with irf(p_c)

Jeśli każdy z układów $(\blacksquare_{fil} vq_{fil},\ \blacksquare_{fj}\ vq_{fj})$ $(1 \le l \le m_k)$ jest poprawny, to regułę R_{e1f} należy dołączyć do podzbioru w niezmienionej postaci. Jeśli każdy z wymienionych układów jest niepoprawny, to regułę należy dołączyć do podzbioru w zmodyfikowanej postaci $R_{e1f}{}'$:

it is declared with grf(p_r):
$\mathrm{A_{si1}}\ \blacksquare_{fi1}\ vq_{fi1}, \ldots, \mathrm{A}_{simk}\ \blacksquare_{fimk}\ vq_{fimk} \rightarrow \mathrm{A_{sj}}\ \blacksquare_{fj}{}'\ vq_{fj}$ with irf($p_c{}'$)

gdzie $\blacksquare_{fj}{}'$ oznacza operator odwrotny do $\blacksquare_{fj}$, a $p_c{}' = 1 - p_c$.

Jeśli są pomiędzy rozważanymi układami takie, których poprawności nie da się określić, to regułę R_{e1f} należy pominąć przy konstrukcji podzbioru.

Drugi etap badania reguł pod kątem dodatniej monotoniczności będzie prowadzony równolegle z ich testowaniem pod względem sprzeczności. W wyselekcjonowanym podzbiorze reguł, każda para reguł sprzecznych (def. 5.7) i każda para reguł spełniających zależność 5.13 lub 5.14 zostanie oznaczona jako „podejrzana". Oznaczenia będę miały przy tym charakter wielokrotny. Następnie, reguła/reguły z największą liczbą oznaczeń zostaną usunięte z podzbioru. Proces oznaczania i usuwania kolejnych reguł będzie powtarzany cyklicznie do momentu, dopóki liczba oznaczeń nie spadnie do zera.

W ostatnim etapie, uzyskany podzbiór reguł będzie testowany pod kątem występowania reguł nadmiarowych, czyli subsumowanych przez inne reguły z tego podzbioru (def. 5.8). Wszystkie wykryte reguły nadmiarowe zostaną bezwarunkowo usunięte z podzbioru.

Zaprezentowany algorytm selekcji reguł do bazy wiedzy systemu PRS(2U) daje gwarancję niesprzeczności zaprojektowanej bazy wiedzy i niewystępowania reguł nadmiarowych w tej bazie. Ponadto, algorytm zapewnia, że zależności reprezentowane przez wyselekcjonowane reguły będą – z wysokim prawdopodobieństwem – dodatnie monotoniczne.

Opisana rygorystyczna selekcja reguł będzie prowadzić do istotnego zmniejszenia liczności wyjściowego zbioru reguł. W wypadku ewidencji mocno skonfliktowanej wewnętrznie, może to doprowadzić do uzyskania małej, a przez to słabej bazy wiedzy. Alternatywnie, można podjąć próbę zastosowania innych, heurystycznych metod selekcji reguł. W procesie selekcji można się kierować, między innymi, kryterium zewnętrznego współczynnika wiarygodności grf.

6.4 Implementacja procesu projektowania bazy wiedzy

Jak wynika z wcześniejszych rozważań, proces projektowania bazy wiedzy systemu PRS(2U) będzie przebiegał w dwóch krokach. Pierwszy z nich polega na cyklicznym wykonywaniu integracji danych atrybutowych zbiorczych i generacji reguł 2U. Każdy cykl tego procesu jest sterowany postacią jednej z dostępnych danych atrybutowych zbiorczych z jednym atrybutem kluczowym. Drugi krok procesu to „przycinanie" zbioru wygenerowanych reguł 2U. Ma ono na celu eliminację ze zbioru tych reguł, które naruszają pewne istotne wymagania jakościowe stawiane bazie wiedzy systemu PRS(2U).

Powyższy proces projektowania bazy wiedzy zaimplementowano w języku C#, przy wykorzystaniu kolekcji standardowych. Na rysunku 6.1 zaprezentowano kluczowy fragment kodu implementującego poszczególne działania. Jest on wzorowany na programie zamieszczonym w [113].

```
Clear(KB);
foreach (d_i in D)  //D - wektor danych atrybutowych zbiorczych
  { Remove(D, d_i);
    K_i = key(d_i); U_i = nonkey(d_i); C_i = common(d_i);
    foreach (E_ij in PowerSet(U_i))
      if (IsSubset(E_ij, C_i))
         { Subtract(U_i, E_ij); Sum(K_i, E_ij);
           foreach (f_ijk in U_i)
             { d_vir = d_i;
               foreach (d_h in D)
                 { K_h = key(d_h); U_h = nonkey(d_h); C_h = common(d_h);
                   flag1 = 1;
                   foreach (A_ig in K_i)
                     if (!(Contains(C_h, A_ig) &&
                           preserves(d_i, d_h, A_ig))
                        { flag1 = 0; break; }
                   if (!flag1)
                      continue;
                   if (!(Contains(U_h f_ijk) &&
                         Contains(C_h, f_ijk) &&
                         preserves(d_i, d_h, f_ijk)) &&
                        !(Contains(U_h, f_ijk) &&
                          !Contains(C_h, f_ijk) &&
                          preserves(d_i, d_h, f_ijk) &&
                          preserves(d_h, d_i, f_ijk)))
                      continue;

                   d_vir = integrate(d_vir, d_h);
                 }
               R_ijk = generate(d_vir, K_i, f_ijk);
               Add(KB, R_ijk);
             }
           Sum(U_i, E_ij);
         }
    Add(D, d_i);
  }
Prune1(KB);
Prune2(KB);
foreach (R_s in KB)
  { Remove(KB, R_s); flag2 = 1;
    foreach (R_t in KB)
      if (subsumes(R_t, R_s))
         { flag2 = 0; break; }
      if (flag2)
         Add(KB, R_s);
  }
```

Kod 6.1. Implementacja procesu projektowania bazy wiedzy systemu PRS(2U)

Użyte w kodzie funkcje: `Clear`, `Add`, `Remove`, `Contains`, `IsSubset`, `Sum`, `Subtract` i `PowerSet` mają charakter generyczny i reprezentują operacje na zbiorach, odpowiednio: czyszczenia zbioru, dodawania elementu do zbioru, usuwania elementu ze zbioru, weryfikacji przynależności elementu do zbioru, weryfikacji zawierania się zbiorów, sumowania zbiorów, odejmowania zbiorów i tworzenia zbioru potęgowego dla wskazanego zbioru. Z kolei funkcje: `key`, `non_key`, `preserves`, `common`, `integrate` i `generate` odnoszą się do danych atrybutowych zbiorczych, implementowanych w postaci słowników. Funkcje te służą do, odpowiednio: wyodrębnienia zbioru atrybutów kluczowych; wyodrębnienia zbioru atrybutów niekluczowych; testowania zachodzenia relacji $\subseteq_q$ pomiędzy zbiorami kwalifikowanymi przypisanymi temu samemu atrybutowi w obrębie dwóch różnych danych atrybutowych; wyodrębnienia atrybutów, do których przypisano zbiory kwalifikowane z maksymalną licznością (w obrębie danej), złączenia wskazanej danej atrybutowej z daną atrybutową wyjściową, wygenerowania reguły 2U na podstawie danej atrybutowej wirtualnej. Funkcja `subsumes` służy do weryfikacji faktu subsumowania drugiej reguły 2U przez pierwszą. Złożone operacje `Prune1` i `Prune2` odpowiadają za: wstępne badanie zależności dodatniej monotonicznej reguł ze wskazanej bazy wiedzy oraz zaawansowane badanie zależności dodatniej monotonicznej i niesprzeczności reguł ze wskazanej bazy wiedzy.

6.5 Przykłady indukcji i selekcji reguł

Zamieszczone w niniejszym podrozdziale przykłady ilustrują kolejno: rozkład danej atrybutowej zbiorczej na zbiór danych atrybutowych z jednym atrybutem kluczowym, integrację danych z jednym atrybutem kluczowym, transformację danej atrybutowej wirtualnej do postaci reguły 2U, projektowanie bazy wiedzy systemu PRS(2U), poprzez: wstępną selekcję reguł 2U, zaawansowaną selekcję reguł 2U, usuwanie reguł nadmiarowych.

Przykład 6.1. Niech $\mathbf{S_{S1,S2}}$ oznacza schemat danych atrybutowych zbiorczych w postaci $\mathbf{S_{S1,S2}} = (\{\text{Wydz\_Rodz}, \text{Kier}\}, \{\text{plec}, \text{rok}, \text{sprzet}\})$, a e_1 – daną atrybutową zbiorczą zbudowaną według tego schematu, w postaci:

$$
\begin{aligned}
e_1 = < \ & \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32; \\
& \text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32; \\
& \text{plec} : \{\text{M}\} \odot |32; \\
& \text{rok} : \{3\} \odot |32; \\
& \text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot |19 >
\end{aligned}
$$

gdzie poszczególne atrybuty i wartości zachowują swoje oznaczenia z rozdziału 5 (jedyna modyfikacja polega na zmianie pierwszych liter w nazwach atrybutów niekluczowych z wielkich na małe).

Powyższą daną można zastąpić następującym, semantycznie równoważnym zbiorem danych atrybutowych zbiorczych z jednym atrybutem kluczowym:

$\{ < \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_a)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{plec} : \{\text{M}\} \odot |32 >,$

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_b)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{rok} : \{3\} \odot |32 >,$

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_c)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot |19 >$

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_d)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{Plec} : \{\text{M}\} \odot |32;$
$\text{rok} : \{3\} \odot |32 >,$

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_e)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{Plec} : \{\text{M}\} \odot |32;$
$\text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot |19 >,$

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_f)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{Rok} : \{3\} \odot |32;$
$\text{plec} : \{\text{M}\} \odot |32 >,$

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_g)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{Rok} : \{3\} \odot |32;$
$\text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot |19 >,$

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_h)
$\text{Kier} : \{\text{Inf}, \text{AiR}\} \oplus |32;$
$\text{Plec} : \{\text{M}\} \odot |32;$
$\text{Rok} : \{3\} \odot |32;$
$\text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot |19 > \}$. □

Przykład 6.2. Niech e_1 oznacza daną atrybutową zbiorczą w postaci identycznej jak w przykładzie 6.1, a e_2 i e_3 – dwie nowe dane atrybutowe zbiorcze, zbudowane według tego samego schematu $\mathbf{S_{S1,S2}}$, w postaci:

$e_2 = <$ Wydz_Rodz : {WInf_st} $\odot$ |45;
Kier : {Inf} $\odot$ |45;
plec : {M} $\odot$ |41;
rok : {2,3} $\oplus$ |39;
sprzet : {Lap} $\odot$ |43 >
$e_3 = <$ Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |71;
Kier : {Inf, AiR} $\oplus$ |71;
plec : {M, K} $\oplus$ |71;
rok : $\neg${3} $\odot$ |36;
sprzet : {K_st, Lap} $\odot$ |68 >

Integracja danych atrybutowych zbiorczych z jednym atrybutem niekluczowym uzyskanych przez rozkład e_1 (dane $e_a - e_h$) z danymi atrybutowymi zbiorczymi z jednym atrybutem niekluczowym uzyskanymi przez analogiczny rozkład e_2 i e_3 da w wyniku następujący zbiór danych atrybutowych wirtualnych:

{ < Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |77; (e_{vir_a})
Kier : {Inf, AiR} $\oplus$ |77;
plec : {M} $\odot$ |73 >, pc=0.95; pr=0,81
< Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |32; (e_{vir_b})
Kier : {Inf, AiR} $\oplus$ |32
rok : {3} $\odot$ |32 >, pc=1; pr=0.95 /0.85
< Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |103; (e_{vir_c})
Kier : {Inf, AiR} $\oplus$ |103;
sprzet : {K_st, Lap} $\odot$ |87 > pc=0.84 pr=0.86
< Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |32; (e_{vir_d})
Kier : {Inf, AiR} $\oplus$ |32;
Plec : {M} $\odot$ |32;
rok : {3} $\odot$ |32 >, pc=1; pr=0.95 /0.85 (nadmiar)
< Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |32; (e_{vir_e})
Kier : {Inf, AiR} $\oplus$ |32;
Plec : {M} $\odot$ |32;
sprzet : {K_st, Lap} $\odot$ |19 >, pc=0.59; pr=0.66
< Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |32; (e_{vir_f})
Kier : {Inf, AiR} $\oplus$ |32;
Rok : {3} $\odot$ |32;
plec : {M} $\odot$ |32 >, pc=1; pr=0.95 /0.85
< Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |32; (e_{vir_g})
Kier : {Inf, AiR} $\oplus$ |32;
Rok : {3} $\odot$ |32;
sprzet : {K_st, Lap} $\odot$ |19 >, pc=0.59; pr=0.66

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32;$ (e_{vir_h})
Kier : $\{\text{Inf}, \text{AiR}\} \oplus |32;$
Plec : $\{\text{M}\} \odot |32;$
Rok : $\{3\} \odot |32;$
sprzet : $\{\text{K\_st}, \text{Lap}\} \odot |19 > \}$. pc=1; pr=0.95 /0.85

Przykładowo, dana wirtualna e_{vir_a} jest wynikiem integracji danej e_a z następującą daną e_i, uzyskaną przez rozkład danej atrybutowej zbiorczej e_2:

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}\} \odot |45;$ (e_i)
Kier : $\{\text{Inf}\} \odot |45;$
plec : $\{\text{M}\} \odot |41 >$

Dana e_i jest zbudowana według tego schematu co e_a ($\mathbf{S_{S1,\{Plec\}}} = = (\{\text{Wydz\_Rodz}, \text{Kier}\}, \{\text{plec}\})$) i jest podobna do niej w sensie relacji $\subseteq_{\mathbf{S1},\{\mathbf{Plec}\}}$ ze względu na spełnienie warunku:

$(\{\text{WInf\_st}, \text{WInf\_nst}\} \oplus |32 \subseteq_{\mathbf{W}} \{\text{WInf\_st}\} \odot |45) \wedge$
$(\{\text{Inf}, \text{AiR}\} \oplus |77 \subseteq_{\mathbf{W}} \{\text{Inf}\} \odot |45) \wedge$
$(\{\text{M}\} \odot |41 \subseteq_{\mathbf{W}} \{\text{M}\} \odot |41)$

Wymagane podobieństwo nie występuje natomiast pomiędzy daną e_a a następującą daną e_j, uzyskaną przez rozkład danej atrybutowej zbiorczej e_3 i zbudowaną według schematu $\mathbf{S_{S1,\{Plec\}}}$:

$< \text{Wydz\_Rodz} : \{\text{WInf\_st}, \text{WInf\_nst}\} \odot |45;$ (e_j)
Kier : $\{\text{Inf}, \text{AiR}\} \odot |45;$
plec : $\{\text{M}, \text{K}\} \oplus |41 >$

Ze względu na niespełnienie warunku:

$\{\text{M}\} \odot |32 \subseteq_{\mathbf{W}} \{\text{M}, \text{K}\} \oplus |41$,

zachodzi ostatecznie:

$e_{vir\_a} = e_a \cap_{\mathbf{S1},\{\text{Plec}\}} e_i =$
$= < \text{Wydz\_Rodz} : (\{\text{WInf}_{\text{st}}, \text{WInf}_{\text{nst}}\} \oplus) \cap_{\mathbf{W}} (\{\text{WInf}_{\text{st}}\} \odot)|(32 + 45);$
Kier : $(\{\text{Inf}, \text{AiR}\} \oplus) \cap_{\mathbf{W}} (\{\text{Inf}\} \odot)(32 + 45);$
plec : $(\{\text{M}\} \odot) \cap_{\mathbf{W}} (\{\text{M}\} \odot)|(32 + 41) > .$

Analogiczne procesy integracji dla danych atrybutowych zbiorczych z jednym atrybutem kluczowym uzyskanych przez rozkład e_2 i e_3 dadzą w wyniku 7 kolejnych danych atrybutowych wirtualnych, zwanych umownie $e_{vir\_i} - e_{vir\_o}$. □

Przykład 6.3. Niech e_{vir_a} oznacza daną atrybutową wirtualną w postaci identycznej jak w przykładzie 6.2:

$<$ Wydz_Rodz : {WInf_st, WInf_nst} $\oplus$ |77;
Kier : {Inf, AiR} $\oplus$ |77;
plec : {M} $\odot$ |73 $>$.

Przy użyciu algorytmu zdefiniowanego w podrozdziale 6.3, daną tę można przekształcić do postaci następującej reguły 2U, zwanej dalej R_{evir_a}:

it is declared with grf(p_r):
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
$\rightarrow$ plec = {M} $\odot$ with irf(p_c)

gdzie $p_c = \frac{73}{77} \approx 0.95$, a $p_r = \min\left(\mathrm{wg}(R_{evir\_a}), \mathrm{acc}(R_{evir\_a})\right) =$

$= \min\left(\min\left(1 - 2 \cdot 1.96 \cdot \sqrt{\frac{0.95 \cdot 0.05}{77}}, 0.95\right), \frac{2 \cdot \frac{1 \cdot 32 + \frac{45}{2}}{77} + 1}{3}\right) \approx \min(0.90, 0.81) =$

$= 0.81$. W obliczeniach współczynnika p_r wzięto pod uwagę:
– przy oznaczaniu wartości wagi wg – współczynnik $p_c' = \min(p_c, 0.95)$,
– przy oznaczaniu wartości precyzji acc – tylko względną precyzję formuł zawartych w przesłance reguły [112].

W wyniku analogicznego procesu transformacji danych wirtualnych $e_{vir\_b} - e_{vir\_o}$, zostaną wygenerowane reguły $R_{evir\_b} - R_{evir\_o}$. □

Przykład 6.4. W zbiorze wygenerowanych 15 reguł 2U znajdują się, między innymi, reguła R_{evir_b}, w postaci:

it is declared with grf(0.85):
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
$\rightarrow$ rok = {3} $\odot$ with irf(1.0)

oraz reguła R_{evir_m}, w postaci:

it is declared with grf(0.77):
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
$\rightarrow$ rok = $\neg${3} $\odot$ with irf(0.51)

Wstępne badanie zbioru wszystkich reguł pod kątem dodatniej monotoniczności prowadzi do konkluzji, że:

$n\_\{\text{WInf\_st, WInf\_nst}\} \oplus_+ \_\{3\} \odot_+ + n\_\{\text{WInf\_st, WInf\_nst}\} \oplus_- \_\{3\} \odot_- =$
$n\_\{\text{WInf\_st, WInf\_nst}\} \oplus_+ \_\{3\} \odot_- + n\_\{\text{WInf\_st, WInf\_nst}\} \oplus_- \_\{3\} \odot_+ =$
$= 2$

W takiej sytuacji, nie jest możliwe określenie charakteru zależności pomiędzy elementami $\{\text{WInf\_st}, \text{WInf\_nst}\} \oplus$ i $\{3\} \odot$, a tym samym – określenie poprawności reguł R_{evir_b} i R_{evir_m}. Z tego powodu, obie wymienione reguły zostaną usunięte ze zbioru (oprócz nich, zostaną usunięte także reguły R_{evir_d} i R_{evir_n}). □

Przykład 6.5. W kolejności, nastąpi proces zaawansowanego badania reguł pod kątem dodatniej monotoniczności oraz pod kątem niesprzeczności. Badaniu zostanie poddana każda para reguł 2U pochodzących z 11-elementowego zbioru, uzyskanego w poprzednim etapie projektowania bazy wiedzy systemu PRS(2U). W zbiorze tym znajdują się, między innymi:
– reguła R_{evir_c}, w postaci:

$$\begin{aligned}&\text{it is declared with grf}(0.86):\\ &\quad \text{Wydz\_Rodz} = \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus,\\ &\quad \text{Kier} = \{\text{Inf}, \text{AiR}\} \oplus,\\ &\rightarrow \quad \text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot \text{ with irf}(0.84)\ ,\end{aligned}$$

– reguła R_{evir_e}, w postaci:

$$\begin{aligned}&\text{it is declared with grf}(0.66):\\ &\quad \text{Wydz\_Rodz} = \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus,\\ &\quad \text{Kier} = \{\text{Inf}, \text{AiR}\} \oplus,\\ &\quad \text{Plec} = \{\text{M}\} \odot,\\ &\rightarrow \quad \text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot \text{ with irf}(0.59)\ ,\end{aligned}$$

– reguła R_{evir_g}, w postaci:

$$\begin{aligned}&\text{it is declared with grf}(0.66):\\ &\quad \text{Wydz\_Rodz} = \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus,\\ &\quad \text{Kier} = \{\text{Inf}, \text{AiR}\} \oplus,\\ &\quad \text{Rok} = \{3\} \odot,\\ &\rightarrow \quad \text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot \text{ with irf}(0.59)\ ,\end{aligned}$$

– reguła R_{evir_h}, w postaci:

$$\begin{aligned}&\text{it is declared with grf}(0.66):\\ &\quad \text{Wydz\_Rodz} = \{\text{WInf\_st}, \text{WInf\_nst}\} \oplus,\\ &\quad \text{Kier} = \{\text{Inf}, \text{AiR}\} \oplus,\\ &\quad \text{Plec} = \{\text{M}\} \odot,\\ &\quad \text{Rok} = \{3\} \odot,\\ &\rightarrow \quad \text{sprzet} : \{\text{K\_st}, \text{Lap}\} \odot \text{ with irf}(0.59)\ .\end{aligned}$$

Łatwo zauważyć, że dla każdej pary reguł: (R_{evir_c}, R_{evir_e}), (R_{evir_c}, R_{evir_g}) i (R_{evir_c}, R_{evir_h}) zachodzi warunek 5.13, świadczący o tym, że przynajmniej jedna z reguł w parze jest zbudowana nieprawidłowo. Na podsta-

wie tego spostrzeżenia, przytoczone reguły zostaną oznaczone etykietą „podejrzana": reguła R_{evir_c} – trzykrotnie, a reguły R_{evir_e}, R_{evir_g} i R_{evir_h} – jednokrotnie.

Dla całego, 11-elementowego zbioru reguł zostanie wyodrębnionych 6 par reguł, dla których zachodzi warunek 5.13 lub 5.14. Równocześnie, żadna para reguł nie będzie pozostawać w sprzeczności (def. 5.7). W rezultacie, po przeprowadzeniu pełnego badania, jedna reguła (R_{evir_c}) zostanie oznaczona etykietą „podejrzana" trzykrotnie, trzy reguły (R_{evir_e}, R_{evir_h}, R_{evir_o}) – dwukrotnie, a dwie reguły (R_{evir_a}, R_{evir_i}) – jednokrotnie.

Po usunięciu ze zbioru reguły R_{evir_c}, a następnie reguły R_{evir_o} (w nowej sytuacji – jedynej oznaczonej etykietą „podejrzana" dwukrotnie), w zbiorze pozostanie 9 reguł, z których dwie (R_{evir_a}, R_{evir_i}) będą nadal „podejrzane". Ponieważ obie te reguły są oznaczone etykietą „podejrzana" jednokrotnie, ostatnia operacja usunięcia reguły odbędzie się na podstawie dodatkowego kryterium – zewnętrznego współczynnika wiarygodności grf. Ze względu na zależność: $grf(R_{evir_a}) = 0.81 < 0.83 = grf(R_{evir_i})$, ze zbioru zostanie usunięta reguła R_{evir_a}.

Przykład 6.6. Ostatni etap projektowania bazy wiedzy systemu PRS(2U) polega na poszukiwaniu i eliminacji reguł nadmiarowych. W bieżącym, 8-elementowym zbiorze reguł, jedyną nadmiarową będzie reguła R_{evir_h}, w postaci:

it is declared with grf(0.66):
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
Plec = {M} $\odot$,
Rok = {3} $\odot$,
→ sprzet : {K_st, Lap} $\odot$ with irf(0.59) .

Jest ona subsumowana zarówno przez regułę R_{evir_e}:

it is declared with grf(0.66):
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
Plec = {M} $\odot$,
→ sprzet : {K_st, Lap} $\odot$ with irf(0.59) ,

jak i regułę R_{evir_g}:

it is declared with grf(0.66):
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
Rok = {3} $\odot$,
→ sprzet : {K_st, Lap} $\odot$ with irf(0.59) .

Po usunięciu ze zbioru reguły R_{evir_g}, do bazy wiedzy $\mathrm{KB}_r(\boldsymbol{R\_2u})$ projektowanego systemu PRS(2U) kandydują ostatecznie reguły:

it is declared with grf(0.66): (R_{evir_e})
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
Plec = {M} $\odot$,
$\rightarrow$ sprzet : {K_st, Lap} $\odot$ with irf(0.59) ,

it is declared with grf(0.85): (R_{evir_f})
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
Rok = {3} $\odot$
$\rightarrow$ plec = {M} $\odot$ with irf(1.00) ,

it is declared with grf(0.66): (R_{evir_g})
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
Rok = {3} $\odot$
$\rightarrow$ sprzet : {K_st, Lap} $\odot$ with irf(0.59) ,

it is declared with grf(0.83): (R_{evir_i})
Wydz_Rodz = {WInf_st} $\odot$,
Kier = {Inf} $\odot$
$\rightarrow$ plec = {M} $\odot$ with irf(0.91) ,

it is declared with grf(0.83): (R_{evir_j})
Wydz_Rodz = {WInf_st} $\odot$,
Kier = {Inf} $\odot$
$\rightarrow$ rok = {2,3} $\oplus$ with irf(0.87) ,

it is declared with grf(0.89): (R_{evir_k})
Wydz_Rodz = {WInf_st} $\odot$,
Kier = {Inf} $\odot$,
$\rightarrow$ sprzet : {Lap} $\odot$ with irf(0.96) ,

it is declared with grf(0.90): (R_{evir_l})
Wydz_Rodz = {WInf_st, WInf_nst} $\oplus$,
Kier = {Inf, AiR} $\oplus$,
$\rightarrow$ plec = {M, K} $\oplus$ with irf(1.00) }

Pobieżna analiza przytoczonych reguł prowadzi do wniosku, że ostatnia z nich (R_{evir_l}) ma charakter nadmiarowy. Wynika on nie z istnienia w bazie wiedzy $\mathrm{KB}_r(\boldsymbol{R\_2u})$ reguły subsumującej regułę R_{evir_l}, lecz ze znajomości

dziedziny wartościowania atrybutu Plec. Następujący fakt ma charakter aksjomatyczny:

it is declared with grf(1.00):
    true
→ plec = {M, K} ⊕ with irf(1.00) }

Faktu tego nie da się jednak wygenerować w procesie klasycznego uczenia indukcyjnego. Potrzebna jest do tego znajomość ontologii dziedzinowej (patrz rozdz. 7).

Użyte w przykładach dane atrybutowe zbiorcze są danymi sztucznymi, skonstruowanymi dla celów ilustracyjnych. Choć uprawdopodobniono je tak, by mogły pochodzić z rzeczywistej ewidencji danych, to jednak ich liczba nie uprawnia do uznania wyindukowanych reguł za realne, a zaproponowanej bazy wiedzy za pełną. W systemie z taką bazą wiedzy nie można by przeprowadzić żadnego ciekawego wnioskowania.