Fuzzy pravidlá ako metóda pre „dolovanie z údajov“

Fuzzy rules as a method for Data Mining

Ján Boháčik, Katedra informatiky FRI ŽU v Žiline

Abstract: Many current companies require processing large databases and capability to discover useful information in large databases. One of ways how to achieve these requirements is possible uses of the Knowledge Discovery in Databases Process. In the Data Mining Step of this Process IF-THEN rules are considered to be one of the most popular and effective representations of knowledge. Integration of fuzzy sets and Fuzzy Logics into the rules and into the whole Process can lead to increasing knowledge understandability and accuracy. In the paper existing types of IF-THEN rules are analyzed, a classification of them is offered and algorithms suitable for their discovery are summarized.

Key words: fuzzy rules, association rules, classification rules, data mining algorithms.

Abstrakt: Medzi základné požiadavky súčasných organizácii patrí spracovanie rozsiahlych databáz a schopnosť získavania užitočných informácii z veľkého množstva údajov. Jedným zo spôsobov ako ich zabezpečiť je využitie procesu Získavania znalostí z databáz. Vo fáze „dolovania z údajov“ sa v tomto procese považujú IF-THEN pravidlá za jedny z najpopulárnejších a najefektívnejších spôsobov reprezentácie znalostí. Integrácia fuzzy množín a fuzzy logiky do pravidiel a celého procesu môže viesť k zvýšeniu zrozumiteľnosti a presnosti znalostí. V nasledujúcom texte analyzujem existujúce typy IF-THEN pravidiel, ponúkam ich klasifikáciu a sumarizujem algoritmy vhodné na ich získavanie.

Kľúčové slová: fuzzy pravidlá, asociačné pravidlá, klasifikačné pravidlá, algoritmy na „dolovanie z údajov“.

1. Úvod

Učenie, ktorého cieľom je získavanie znalosti z údajov, sa v anglickej literatúre označuje „Data Mining“. V slovenčine sa presadzuje termín „dolovanie z údajov“. Výsledkom dolovania z údajov sú znalosti. Tieto znalosti môžu byť reprezentované rôznymi spôsobmi. Jedným z pre človeka najprístupnejších a zároveň najpopulárnejších spôsobov reprezentácie

znalostí sú pravidlá. Významnou skupinou pravidiel sú fuzzy pravidlá. Pre fuzzy pravidlá je nevyhnutná existencia fundamentov fuzzy logiky. Tie sa používajú ako jeden zo spôsobov opísania reálneho sveta ponímaním, ktoré je blízke ľudskému uvažovaniu a vnímaniu. Tento spôsob jeho popisu sa dokonca spomedzi všetkých známych javí byť najvhodnejší (23). A teda reprezentácia znalosti v podobe fuzzy pravidiel je nielen ľahko zrozumiteľná, ale umožňuje odstrániť i niektoré obmedzenia vyplývajúce z typu vstupných údajov a neurčitosti procesu ľudského poznávania v reálnom svete.

V kapitole 2 ponúkam klasifikáciu pravidiel používaných pri dolovaní z údajov. V nasledujúcich kapitolách 3 a 4 podrobnejšie diskutujem definície úloh o hľadaní asociačných pravidiel a algoritmy, ktoré ich umožňujú riešiť. Kapitola 5 obsahuje zhrnutie prínosov.

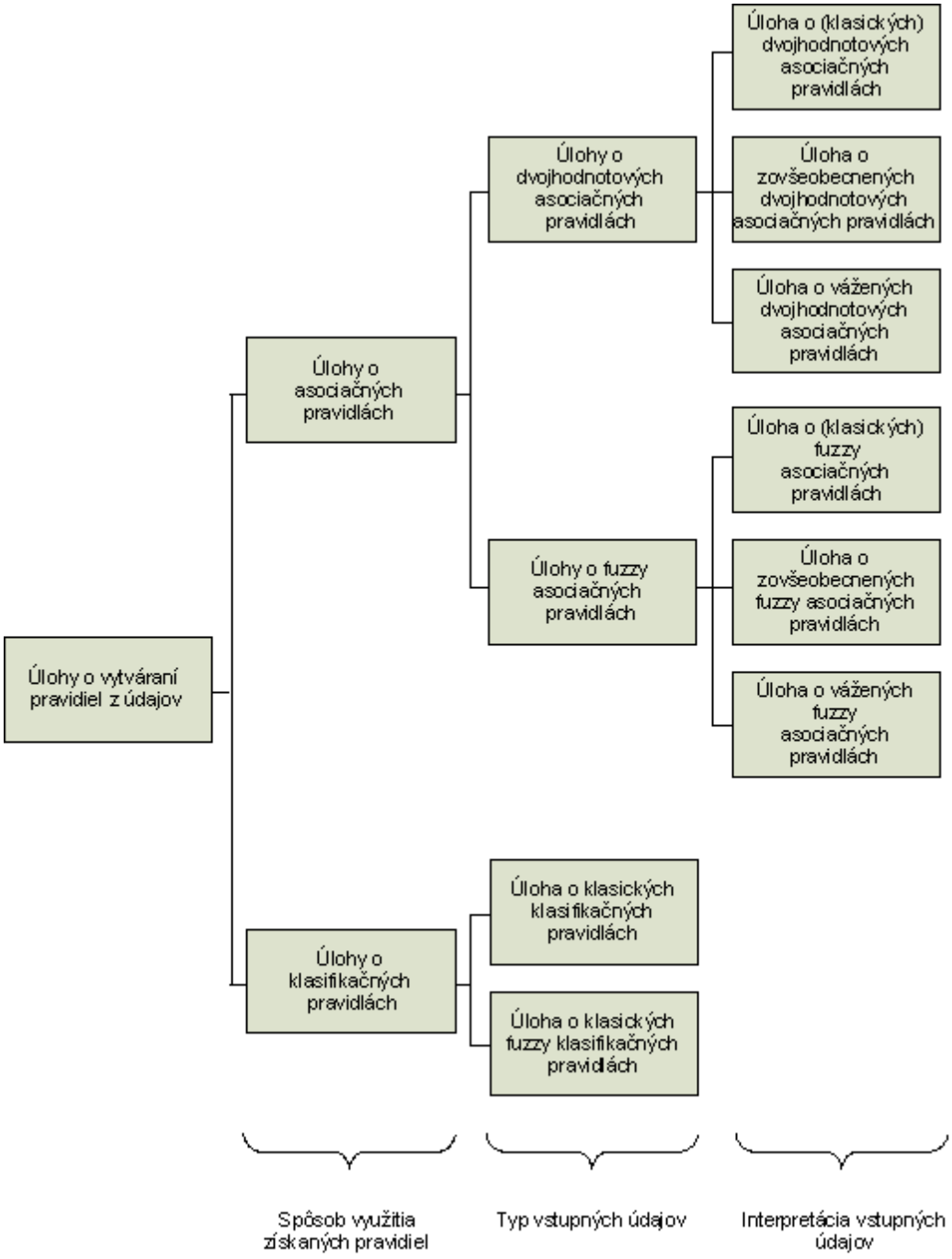
2. Klasifikácia fuzzy pravidiel

Medzi najpopulárnejšie spôsoby reprezentácie znalosti patria pravidlá IF-THEN nasledujúceho tvaru. Ich popularita je spôsobená tak variabilitou ich použitia, ako aj ich blízkosťou ľudskému uvažovaniu. Z toho dôvodu sa vytváranie pravidiel stáva stále dôležitejšou súčasťou dolovania z údajov. Na Obrázok 1 predkladám klasifikáciu úloh o vytváraní pravidiel z údajov. Túto klasifikáciu som navrhol na základe širokej škály mne dostupných odborných článkov.

Podľa spôsobu využitia získaných pravidiel som úlohy o vytváraní pravidiel rozlíšil na úlohy o klasifikačných pravidlách a úlohy o asociačných pravidlách. Pri úlohách o klasifikačných pravidlách sa vytvárajú pravidla, ktoré sa používajú na zaraďovanie inštancií do určitého konceptu (triedy). Toto zaraďovanie inštancií prebieha podľa hodnôt ich atribútov, resp. lingvistických premenných. Úlohy o asociačných pravidlách sú zamerané na vytváranie pravidiel, ktoré ukazujú zaujímavé asociácie medzi atribútmi, resp. lingvistickými premennými zozbieraných inštancií.

Úlohy na Obrázok 1 som ďalej podrobnejšie klasifikoval podľa typu vstupných údajov. Myslím tým typy údajov, ktoré sa používajú počas doby vytvárania pravidiel. Algoritmy určené pre vytváranie pravidiel pracujú primárne s kategoriálnymi atribútmi (2). Ich zovšeobecnením je nahradenie kategoriálnych atribútov lingvistickými premennými. Vtedy sa hovorí o ***fuzzy pravidlách***. Používanie fuzzy pravidiel poskytuje niekoľko výhod. Hlavnou výhodou je, že pri ich použití možno uvažovať nejasnosť (vágnosť) a nejednoznačnosť, ktorá sa objavuje v procese ľudského poznávania. Nejasnosť a nejednoznačnosť je dobre popísaná v (17). Zakomponovanie nejasnosti i nejednoznačnosti umožňuje, okrem iného, optimálnejšie

pracovať s numerickými atribútmi. Dosahuje sa to tzv. fuzzifikáciou numerických údajov. Algoritmy, ktoré nepracujú s fuzzy pravidlami, vyžadujú diskretizáciu. Pri disketizácií vstupných údajov dochádza k ich značnej deformácií. Tá má často za následok neočakávané zmeny výsledkov už pri malom rozdieli pôvodných numerických hodnôt atribútov (12). K fuzzifikácií, resp. k diskretizácií sa pristupuje buď pred samotným vytváraním pravidiel, alebo až počas vytvárania pravidiel ako napríklad v (30).



Obrázok 1: Mnou navrhnutá klasifikácia úloh o vytváraní pravidiel z údajov.

Pre asociačné pravidlá je dôležitá interpretácia závislosti medzi vstupnými atribútmi, resp. lingvistickými premennými. Na Obrázok 1 to dokumentuje časť Interpretácia vstupných údajov. Medzi vstupnými atribútmi sú niekedy definované is-a hierarchie. Tieto is-a hierarchie sú charakteristické pre zovšeobecnené asociačné pravidlá. Prvá zmienka o zovšeobecnených asociačných pravidlách je v (28). Tento článok ešte nepopisoval použiteľný algoritmus. Jeden z prvých takýchto algoritmov je publikovaný v (29). V (7) je popísaný algoritmus pre hľadanie zovšeobecnených fuzzy asociačných pravidiel s tzv. fuzzy is-a hierarchiami. Niekedy sa niektorým atribútom prikladá väčší význam než iným. To znamená, že každý atribút ma priradenú váhu. V takom prípade sa hovorí o vážených asociačných pravidlách. Spôsoby tvorby tohoto typu pravidiel sú popísané napríklad v (6) a (32). Vážené fuzzy asociačné pravidlá sa dajú vytvárať algoritmami, ktoré sú predstavené v (27) a (13).

V nasledujúcich dvoch kapitolách 3, 4 sa podrobnejšie venujem úlohe o klasických dvojhodnotových asociačných pravidlách, úlohe o zovšeobecnených dvojhodnotových asociačných pravidlách, úlohe o klasických klasifikačných pravidlách a úlohe o klasických fuzzy klasifikačných pravidlách. Pre tieto úlohy bolo vyvinutých viacero algoritmov. Z toho dôvodu sa dajú identifikovať spoločné princípy, ktoré tieto algoritmy využívajú. Pre každú z týchto úloh uvádzam exaktnú definíciu, základné princípy používané algoritmami pre ich riešenie, a odkazy na konkrétne články s popisom príslušných algoritmov.

3. Úlohy o asociačných pravidlách

Definícia 1.1 (Úloha o klasických dvojhodnotových asociačných pravidlách, modifikácia z (15)):

Nech $\boldsymbol{U}$ je úplná množina inštancií $\boldsymbol{e}$. Nech je $\forall\, \boldsymbol{e} \in \boldsymbol{U}$ popísaná atribútmi $\boldsymbol{A} = \{A_1; ...; A_k; ...; A_n\}$. Nech $\forall\, A_k$ nadobúda jednu z hodnôt 0 alebo 1, t. j. $A_k = \{0;1\}$. Nech hodnota 0 (1) znamená, že A_k nie je (je) zahrnuté v inštancií $\boldsymbol{e} \in \boldsymbol{U}$. Nech $\boldsymbol{V} \subseteq \boldsymbol{U}$ je množina inštancií so známymi hodnotami všetkých atribútov v $\boldsymbol{A}$. Cieľom je nájsť **klasické dvojhodnotové asociačné pravidlá** tvaru:

$$\mathrm{R_i} = \mathrm{R_i}(\boldsymbol{E}_i^{Predpoklad}, \boldsymbol{E}_i^{Záver}) = \textbf{IF}\ \ \boldsymbol{E}_i^{Predpoklad}\ \ \textbf{THEN}\ \boldsymbol{E}_i^{Záver}\ \ \forall\ \ i = 1, 2, ..., m, \text{ kde}$$

$\boldsymbol{E}_i^{Predpoklad} \subset \mathbf{A}$, $\boldsymbol{E}_i^{Záver} \subset \mathbf{A}$, $\mathrm{M}(\boldsymbol{E}_i^{Predpoklad}) \geq 1$, $\mathrm{M}(\boldsymbol{E}_i^{Záver}) \geq 1$, $\boldsymbol{E}_i^{Predpoklad} \cap \boldsymbol{E}_i^{Záver} = \emptyset$. Nájdené pravidlá musia spĺňať stanovené charakteristiky (kritéria). Obvykle vždy aspoň minimálnu podporu ***minpod*** a minimálnu spoľahlivosť ***minspoľ***.

Na výpočet hodnôt charakteristík dvojhodnotových asociačných pravidiel možno použiť kontingenčnú tabuľku uvedenú v Tabuľka 1. Túto tabuľku som zostavil podľa (2). Nech je pre $\boldsymbol{X} \subseteq \mathbf{A}$ definovaná funkcia $\boldsymbol{X}(\boldsymbol{e})$, $\boldsymbol{e} \in \boldsymbol{V}$:

$$\boldsymbol{X}(\boldsymbol{e}) = \begin{cases} 1\text{; práve vtedy ak } \forall\, A_k \in \boldsymbol{X}\text{: pre inštanciu } \boldsymbol{e} \text{ je } A_k = 1 \cdot \\ 0\text{; inak} \end{cases}$$

Potom pre Tabuľka 1 platí: $a = \sum_{\boldsymbol{e} \in V} \mathbf{min}(\boldsymbol{E}_i^{\boldsymbol{Predpoklad}}(\boldsymbol{e});\ \boldsymbol{E}_i^{\boldsymbol{Záver}}(\boldsymbol{e}))$ - počet objektov spĺňajúcich predpoklad i záver, $b = \sum_{\boldsymbol{e} \in V} \mathbf{min}(\boldsymbol{E}_i^{\boldsymbol{Predpoklad}}(\boldsymbol{e});\ 1 - \boldsymbol{E}_i^{\boldsymbol{Záver}}(\boldsymbol{e}))$ - počet objektov spĺňajúcich predpoklad a nespĺňajúcich záver, $c = \sum_{\boldsymbol{e} \in V} \mathbf{min}(1 - \boldsymbol{E}_i^{\boldsymbol{Predpoklad}}(\boldsymbol{e});\ \boldsymbol{E}_i^{\boldsymbol{Záver}}(\boldsymbol{e}))$, $d = \sum_{\boldsymbol{e} \in V} \mathbf{min}(1 - \boldsymbol{E}_i^{\boldsymbol{Predpoklad}}(\boldsymbol{e});\ 1 - \boldsymbol{E}_i^{\boldsymbol{Záver}}(\boldsymbol{e}))$.

Tabuľka 1: Kontingenčná tabuľka pre asociačné pravidlá (2).

	Spĺňa $\boldsymbol{E}_i^{\boldsymbol{Záver}}$	Nespĺňa $\boldsymbol{E}_i^{\boldsymbol{Záver}}$
Spĺňa $\boldsymbol{E}_i^{\boldsymbol{Predpoklad}}$	a	b
Nespĺňa $\boldsymbol{E}_i^{\boldsymbol{Predpoklad}}$	c	d

Definíciu podpory a spoľahlivosti pravidla môžno formulovať nasledovne:

$$\text{Podpora}(\mathrm{R_i}) = \text{Podpora}(\boldsymbol{E}_i^{\boldsymbol{Predpoklad}} \cup \boldsymbol{E}_i^{\boldsymbol{Záver}}) = \frac{a}{a+b+c+d},$$

$$\text{Spoľahlivosť}(\mathrm{R_i}) = \frac{a}{a+b}.$$

Z Definícia 1.1 vyplýva, že asociačné pravidlá musia spĺňať stanovenú ***minpod*** a ***minspoľ***. Vďaka tomu sa dá úloha o klasických dvojhodnotových pravidlách dekomponovať na dva základne kroky predstavené v (1). Tieto kroky uvádzam s ohľadom na v práci zavedené definície a pojmy v nasledujúcej tabuľke Tabuľka 2:

Tabuľka 2: Všeobecný algoritmus pre úlohu o klasických dvojhodnotových pravidlách.

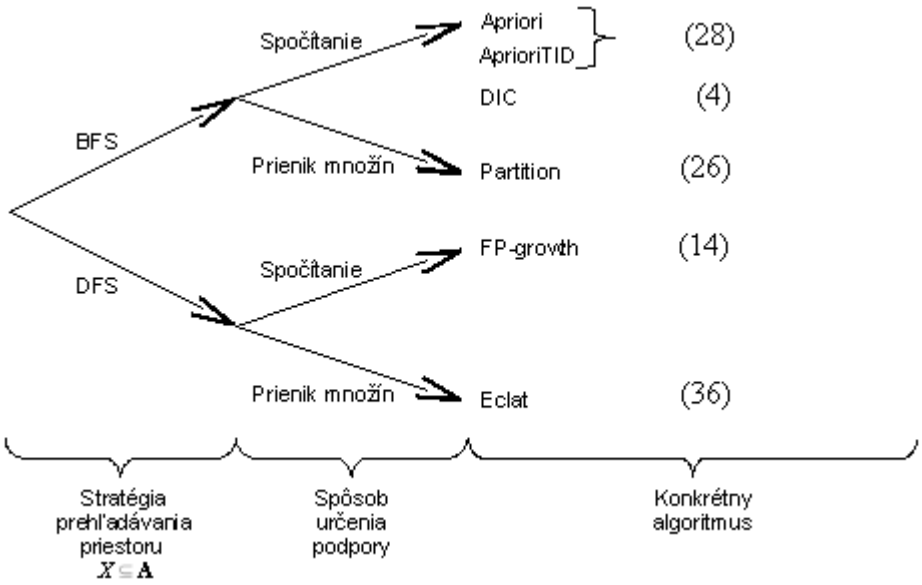
Krok 1	Nájdi množinu $\boldsymbol{F}$ všetkých ***frekventovaných položkových množín***. Pre množinu $\boldsymbol{F}$ platí $\boldsymbol{F}=\{\boldsymbol{X}\subseteq\mathbf{A}\,|\,\mathrm{M}(\boldsymbol{X})\geq 1,\ \mathrm{Podpora}(\boldsymbol{X})\geq \boldsymbol{minpod}\}$.
Krok 2	Pre $\forall\,\boldsymbol{X}\in\boldsymbol{F},\ \mathrm{M}(\boldsymbol{X})\geq 2$ vypočítaj všetky možné $\mathrm{Spoľahlivosť}(\mathrm{R})$, kde $\mathrm{R}=\mathbf{IF}\ \ \boldsymbol{X}-\boldsymbol{Y}\ \ \mathbf{THEN}\,\boldsymbol{Y},\ \boldsymbol{Y}\subset\boldsymbol{X},\ \boldsymbol{X}-\boldsymbol{Y}\neq\emptyset,\ \boldsymbol{Y}\neq\emptyset,\ \boldsymbol{X}-\boldsymbol{Y}\cap\boldsymbol{Y}=\emptyset$. Do výslednej množiny vytvorených pravidiel zaraď také pravidlá, pre ktoré platí $\mathrm{Spoľahlivosť}(\mathrm{R})\geq \boldsymbol{minspoľ}$.

Vo všeobecnom algoritme v Tabuľka 2 sa ukázal byť zložitým **Krok 1**. Dôvodom je skutočnosť, že $\mathrm{M}(\boldsymbol{F})$ rastie exponenciálne s lineárnym nárastom $\mathrm{M}(\mathbf{A})$. Z toho dôvodu je cieľom väčšiny časovo efektívnych algoritmov znížiť $\mathrm{M}(\boldsymbol{F})$ pri zachovaní optimálneho výsledku. Takmer všetky algoritmy to riešia prehľadávaním určitého špeciálneho stromu frekventovaných položkových množín (15). Tento strom je založený na matematických vetách predstavených v (1). Jeho prehľadávanie sa uskutočňuje dvoma základnými spôsobmi:

- **Prehľadávanie do šírky (BFS)** - $\mathrm{Podpora}(\boldsymbol{X})\ \ \forall\,\boldsymbol{X}\in\boldsymbol{F}$, kde $\mathrm{M}(\boldsymbol{X})=k$ je určená pred určením $\mathrm{Podpora}(\boldsymbol{X})\ \ \forall\,\boldsymbol{X}\in\boldsymbol{F}$, kde $\mathrm{M}(\boldsymbol{X})=k$-1.
- **Prehľadávanie do hĺbky (DFS)** – Rekurzívne prechádzanie stromovou štruktúrou a z toho vyplývajúce rekurzívne určovanie podpory.

Na určovanie podpory pri prehľadávaní stromu frekventovaných položkových množín sa obvykle uplatňuje jeden z nasledujúcich spôsobov:

- **Spočítanie** – Pre $\forall\,\boldsymbol{e}\in\boldsymbol{V}$ a pre aktuálnu $\boldsymbol{X}\in\boldsymbol{F}$ sa položí $a=a+\boldsymbol{X}(\boldsymbol{e})$, kde a je z kontingenčnej tabuľky v Tabuľka 1.
- **Prienik množín** – Ku $\forall\,\boldsymbol{e}\in\boldsymbol{V}$ je priradený jednoznačný identifikátor ID. Pre $\forall\,\boldsymbol{X}\subseteq\boldsymbol{A}$ s $\mathrm{M}(\boldsymbol{X})=1$ je definovaný ID zoznam označený $\boldsymbol{X}$.IDZoznam. Tento zoznam obsahuje ID všetkých $\boldsymbol{e}\in\boldsymbol{V}$ takých, že $\boldsymbol{X}(\boldsymbol{e})=1$. Zodpovedajúcim spôsobom je $\forall\,\boldsymbol{C}\subseteq\boldsymbol{A},\ \mathrm{M}(\boldsymbol{C})>1$ definovaný $\boldsymbol{C}$.IDZoznam tak, že pre $\boldsymbol{C}=\boldsymbol{X}\cup\boldsymbol{Y},\ \boldsymbol{X}\subseteq\boldsymbol{A},\ \boldsymbol{Y}\subseteq\boldsymbol{A}$ sa $\boldsymbol{C}$.IDZoznam$=\boldsymbol{X}$.IDZoznam$\cap\boldsymbol{Y}$.IDZoznam. Z dôvodu zvýšenia efektívnosti prienikov sú ID zoznamy usporiadané vo vzostupnom poradí. Pre a uvedené v kontingenčnej tabuľke Tabuľka 2 platí, že $a=\mathrm{M}(\boldsymbol{C}.\mathrm{IDZoznam})$.



Obrázok 2: Klasifikácia algoritmov pre riešenie úlohy o klasických dvojhodnotových *asociačných pravidlách (15).*

Na Obrázok 2 uvádzam klasifikáciu algoritmov podľa spôsobu prehľadávania stromu frekventovaných položkových množín, t. j. podľa spôsobu prehľadávania priestoru riešení, a podľa spôsobu určovania podpory pri prehľadávaní.

Úloha o klasických fuzzy asociačných pravidlách je priamym zovšeobecnením doteraz analyzovanej úlohy o klasických dvojhodnotových asociačných pravidlách. Rozdiel je v tom, že sa už nepoužívajú dvojhodnotové atribúty $\boldsymbol{A}=\{A_1; ...; A_k; ...; A_n\}$. Ich funkciu preberajú a rozširujú lingvistické premenné. Mnoho algoritmov pre riešenie tejto úlohy využíva princípy, ktoré diskutujem v predchádzajúcich odstavcoch. Obsahujú však určité modifikácie pre zabezpečenie funkčnosti s ohľadom na fuzzy interpretáciu a využitie fuzzy logiky. Typickým príkladom je algoritmus v (12).

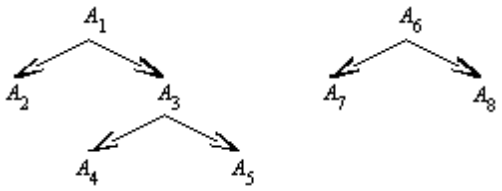
Definícia 1.2:

Nech je daná množina $\boldsymbol{A}=\{A_1; ...; A_k; ...; A_n\}$. **Množina is-a hierarchií medzi prvkami v** $\boldsymbol{A}$ je množina acyklických orientovaných grafov, ktorých vrcholy sú nejaké $A_k \in \mathbf{A}$. Pre všetky tieto grafy navyše platí, že každý z nich obsahuje $\forall A_k \in \mathbf{A}$ najviac raz. Tiež platí, že dva rôzne grafy nemôžu obsahovať rovnaký A_k.

Predpokladajme, že v množine is-a hierarchií existuje orientovaná hrana z A_k ku A_l. Hovoríme, že A_l je následník A_k, resp. A_k je predok A_l. Píšeme následník$(A_k)=A_l$, resp. predchodca$(A_l)=A_k$.

Príklad 1.1:

Nech je daná množina $\boldsymbol{A}=\{A_1; A_2; ...; A_7\}$. Nech je medzi jej prvkami definovaná množina is-a hierarchií zobrazených na obrázku Obrázok 3. Potom sa dá písať následník$(A_3)=A_4 \vee A_5$, predchodca$(A_8)=A_6$.



Obrázok 3: Príklad množiny is-a hierarchií.

Definícia 1.3 (Úloha o zovšeobecnených dvojhodnotových asociačných pravidlách, modifikácia zo (29):

Úloha o zovšeobecnených dvojhodnotových asociačných pravidlách je úloha o klasických dvojhodnotových asociačných pravidlách s tým rozdielom, že medzi atribútmi v $\boldsymbol{A}=\{A_1; ...; A_k; ...; A_n\}$ je definovaná množina is-a hierarchií. Hľadajú sa ***zovšeobecnené dvojhodnotové asociačné pravidlá***, ktoré spĺňajú podmienky uvedené pre pravidlá v definícií úlohy o (klasických) dvojhodnotových asociačných pravidlách. Naviac však žiadne $A_k \in \boldsymbol{E_i^{Záver}}$ nie je predok žiadnej $A_l \in \boldsymbol{E_i^{Predpoklad}}$.

Zovšeobecnené dvojhodnotové asociačné pravidlá zachytávajú asociácie medzi atribútmi $A_k \in \mathbf{A}$ na ľubovoľnej úrovni is-a hierarchií. Teda nielen na najnižšej úrovni ako je to v klasickom prípade. Zabraňuje to situácií, keď by napríklad pravidlo **IF** A_2 **THEN** A_7 spĺňalo pri atribútoch z Príklad 1.1 minimálnu podporu ***minpod*** a minimálnu spoľahlivosť ***minspoľ***, zatiaľ čo pravidlo **IF** A_1 **THEN** A_7 už nie.

Hlavným problémom pri hľadaní zovšeobecnených dvojhodnotových asociačných pravidiel je voľba minimálnej požadovanej podpory (2). Ak je jej stanovená hodnota príliš vysoká,

nenájdu sa pravidlá na najnižšej úrovni. Nie je ani zaručené nájdenie všeobecnejšieho pravidla pre nízku spoľahlivosť takéhoto pravidla. Ak je minimálna požadovaná podpora pravidiel príliš vysoká, dochádza ku kombinatorickej explózii. Navyše sa atribúty A_k z vyšších úrovni vyskytujú v nájdených pravidlách vo všetkých možných vzájomných kombináciách.

Jednými z prvých efektívnych algoritmov pre riešenie úlohy o zovšeobecnených dvojhodnotových asociačných pravidlách boli Cumulate a EstMerge (29). V (8) je algoritmus, ktorý používa rôznu minimálnu podporu na rôznych úrovniach hierarchie.

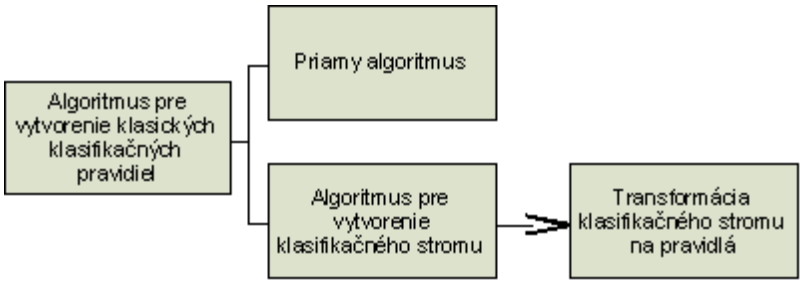
4. Úlohy o klasifikačných pravidlách

Definícia 1.4 (Úloha o klasických klasifikačných pravidlách, modifikácia z (35)):

Nech $\boldsymbol{U}$ je úplná množina inštancií $\boldsymbol{e}$. Nech je $\forall\, \boldsymbol{e} \in \boldsymbol{U}$ popísaná atribútmi $\boldsymbol{A} = \{A_1; ...; A_k; ...; A_n\}$ a nech $\forall\, A_k$ nadobúda jednu z obvykle malého počtu kategoriálnych hodnôt $A_k = \{a_{k,1}; ...; a_{k,l}; ...; a_{k,n_k}\}$. Nech $\boldsymbol{V} \subseteq \boldsymbol{U}$ je množina inštancií, pre ktoré sú známe hodnoty A_k. Nech je $\forall\, \boldsymbol{e} \in \boldsymbol{V}$ klasifikovaná (priradená) do tzv. ***triedy***, t.j. do jednej z hodnôt atribútu $C = \{c_1; ...; c_2; ...; c_L\}$. Pri tejto úlohe sa vytvárajú ***klasické klasifikačné pravidlá***, ktoré majú nasledujúci tvar:

$$\textbf{IF}\ \boldsymbol{E_i^{Predpoklad}}\ \textbf{THEN}\ C\ \text{is}\ c_j\ \ \forall\ i = 1, 2, ..., m, \text{ kde}$$

$\boldsymbol{E_i^{Predpoklad}} = A_{i_1}$ is a_{j_1} **AND** A_{i_2} is a_{j_2} **AND** ... **AND** $A_{i_{n_i}}$ is $a_{j_{n_j}}$. Navyše, $\boldsymbol{E_i^{Predpoklad}}$ obsahuje minimálne jeden atribút a žiaden atribút sa neopakuje. Cieľom je s použitím vytvorených pravidiel určovať hodnoty $c_j \in C$ pre $\boldsymbol{e} \in \boldsymbol{U}$ so známymi hodnotami A_k.



Obrázok 4: Mnou navrhnuté základné členenie algoritmov pre riešenie úlohy *o klasických klasifikačných pravidlách.*

Na základe mne dostupných odborných článkov som navrhol klasifikáciu, ktorá rozlišuje dva základné spôsoby pre vytváranie klasických klasifikačných pravidiel (Obrázok 4). Jedným z nich sú priame algoritmy. Tie obvykle využívajú buď pokrývanie množín, alebo elimináciu

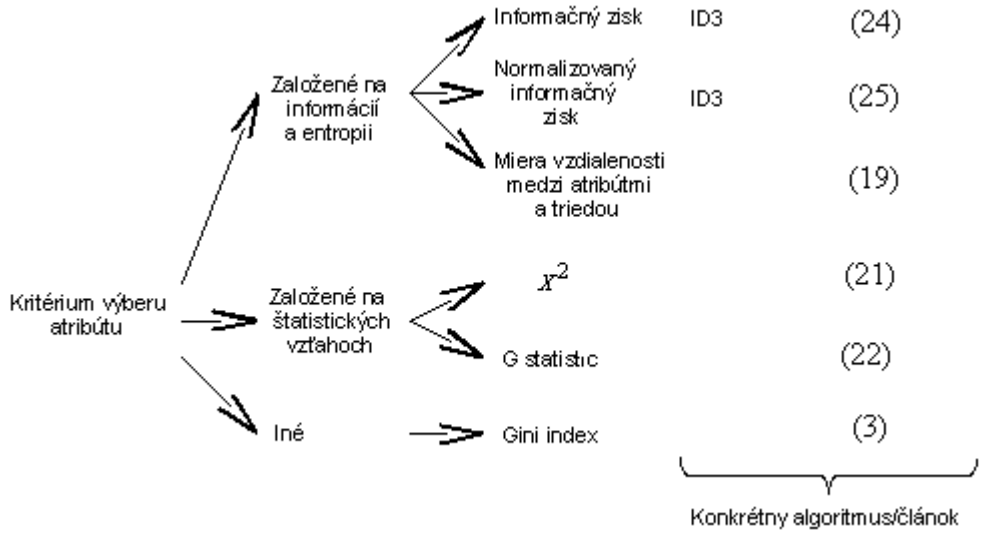
atribútov. Pri pokrývaní množín sa v priestore atribútov postupne vyberajú oblasti, ktoré obsahujú inštancie prevažne jednej triedy. Tieto oblasti sa následne transformujú na klasické klasifikačné pravidla. Algoritmy tohto typu sú AQ (20), CN2 (10), CN4 (5). Elimináciu atribútov používa algoritmus RITIO (34). Tento algoritmus eliminuje najmenej významné atribúty čo najbližšie k začiatku vytvárania pravidiel. Zároveň sa pokúša zlepšiť efektívnosť vytvárania pravidiel a zvýšiť ich klasifikačnú presnosť.

Druhým uvedeným spôsobom na Obrázok 4 je vytváranie klasifikačného stromu, ktorý sa následne transformuje na pravidlá. Všeobecný algoritmus pre zostavenie klasifikačného stromu predkladám v Tab 1.5:

Tabuľka 3: Všeobecný algoritmus pre vytvorenie klasifikačného stromu.

Krok 1	Vytvor koreň stromu a asociuj s ním atribút, ktorý najlepšie vyhovuje kritériu výberu. Vytvor vetvu pre každú možnú hodnotu vybraného atribútu. Asociuj s práve vytvorenými vetvami príslušné hodnoty vybraného atribútu. Priraď vetvy ku koreňu. Označ vetvy za nespracované.
Krok 2	Ak neexistuje nespracovaná vetva, **KONIEC**. Inak vyber nespracovanú vetvu. Označ ju za spracovanú. Ďalej postupuj jedným z dvoch nasledujúcich možností: • Ak nie je splnená zastavovacia podmienka, priraď k práve označenej vetve uzol. S uzlom asociuj atribút, ktorý vyberieš podľa kritéria výberu. K uzlu priraď jednu vetvu pre každú možnú hodnotu vybraného atribútu. S vetvami asociuj príslušné hodnoty. Pripoj vetvy k vybranému uzlu. • Ak je splnená zastavovacia podmienka, priraď k práve označenej vetve list. Z listom asociuj hodnotou z C. Opakuj **Krok 2**.

Kľúčovým problémom pri vytváraní klasifikačného stromu je výber atribútu pre asociáciu s uzlom. Na základe štúdia odborných článkov som dospel ku klasifikácií, ktorú uvádzam na Obrázok 5. Na tomto obrázku som vyznačil jednotlivé možnosti pre výber atribútu. K týmto možnostiam som zároveň priradil odkazy na konkrétne algoritmy, ktoré ich využívajú.



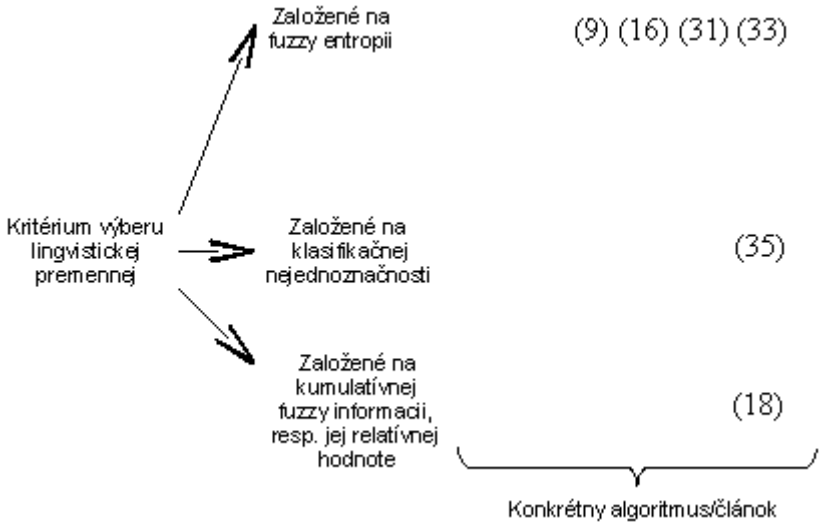
Obrázok 5: Mnou navrhnutá klasifikácia algoritmov pre vytváranie klasifikačného stromu (klasifikácia podľa kritéria výberu atribútu).

Definícia 1.5 (Úloha o klasických fuzzy klasifikačných pravidlách):

Nech $\boldsymbol{U}$ je úplná množina inštancií $\boldsymbol{e}$. Nech je $\forall\, \boldsymbol{e} \in \boldsymbol{U}$ popísaná lingvistickými premennými $\boldsymbol{A} = \{A_1; ...; A_k; ...; A_n\}$. Nech sú tieto lingvistické premenné asociované s $\boldsymbol{U}$ a nech $\forall\, A_k$ nadobúda jednu z obvykle malého počtu lingvistických výrazov $A_k = \{a_{k,1}; ...; a_{k,l}; ...; a_{k,n_k}\}$. Nech $\boldsymbol{V} \subseteq \boldsymbol{U}$ je množina inštancií $\boldsymbol{e}$, pre ktoré sú známe hodnoty $\mu_{a_{k,l}}(\boldsymbol{e}) \forall\, a_{k,l} \in A_k,\ \forall\, A_k \in \mathbf{A}$. Nech je $\forall\, \boldsymbol{e} \in \boldsymbol{V}$ klasifikovaná hodnotami $\mu_{c_j}(\boldsymbol{e}),\ \forall\, c_j \in C$, kde $C = \{c_1; ...; c_2; ...; c_L\}$ je lingvistická premenná asociovaná s $\boldsymbol{U}$. $\boldsymbol{C}$ často nazývam tiež ***triedna lingvistická premenná***. V rámci tejto úlohy sa vytvárajú ***klasické fuzzy klasifikačné pravidlá*** nasledujúceho tvaru:

$$\textbf{IF}\ \boldsymbol{E}_i^{\boldsymbol{Predpoklad}}\ \textbf{THEN}\ C\ \text{is}\ c_j\ \forall\ i = 1, 2, ..., m, \text{ kde}$$

$\boldsymbol{E}_i^{\boldsymbol{Predpoklad}} = A_{i_1}$ is a_{j_1} **AND** A_{i_2} is a_{j_2} **AND** ... **AND** $A_{i_{n_i}}$ is $a_{j_{n_j}}$. $\boldsymbol{E}_i^{\boldsymbol{Predpoklad}}$ navyše obsahuje minimálne jednu lingvistickú premennú a žiadna lingvistická premenná sa v ňom neopakuje. Cieľom je na základe vytvorených pravidiel určovať pre nejaké $\boldsymbol{e} \in \boldsymbol{U}$ hodnoty $\mu_{c_j}(\boldsymbol{e}),\ \forall\, c_j \in C$.



Obrázok 6: Mnou navrhnutá klasifikácia algoritmov pre vytváranie fuzzy klasifikačného stromu (klasifikácia podľa kritéria výberu lingvistickej premennej).

Analýzou odborných článkov som zistil, že klasické fuzzy klasifikačné pravidlá sú obvykle vytvárané transformáciou z fuzzy klasifikačného stromu. Všeobecný algoritmus pre vytvorenie fuzzy klasifikačného stromu zodpovedá algoritmu v Tabuľka 3 ak sa nahradí pojem atribút za pojem lingvistická premenná a pojem hodnota atribútu za pojem lingvistický výraz definovaný pre lingvistickú premennú. Kľúčovým je kritérium výberu lingvistickej premennej pre asociáciu s uzlom (11). Na obrázku Obrázok 6 uvádzam svoju klasifikáciu algoritmov podľa tohto kritéria.

5. Záver

Hlavným prínosom tejto práce je :

a) Súhrnná analýza fázy „dolovania z údajov“ v procese Získavania znalostí z databáz ak sa predpokladá použitie (fuzzy) IF-THEN pravidiel;

b) Navrhnutá klasifikácia IF-THEN pravidiel používaných vo fáze „dolovania z údajov“;

c) Definovanie úloh o vytváraní pravidiel jednotnou terminológiou;

d) Sumarizácia princípov algoritmov pre vytváranie IF-THEN pravidiel.

Tento príspevok bol vypracovaný ako súčasť denného doktorandského štúdia na Katedre informatiky Fakulty riadenia a informatiky Žilinskej univerzity.

6. Zoznam bibliografických odkazov

(1) Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proc. of ACM SIGMOD International Conference on Management Data, Washington (1993), 914-925.

(2) Berka, P.: Knowledge Discovery in Databases. Academia – the Publisher of the Czech Academy of Sciences, Prague, 2003. (in Czech)

(3) Breiman, L., Friedman, J. H., Olsen, R. A., Stone, C. J.: Classification and Regression Trees. Wadsworth & Brooks, Monterey, 1984.

(4) Brin, S., Motwani, R., Silverstein, C.: Beyond Market baskets: Generalizing association rules to correlations. In: Proc. Of 1997 ACM SIGMOD International Conference on Management of Data, ACM Press, Tucson (1997), 265-276.

(5) Bruha, I., Kočková, S.: A support for decision making: Cost-sensitive learning system. Artificial Intelligence in Medicine 6 (1994) 67-82.

(6) Cai, CH., Fu, A., Cheng, CH., Kwong, WW.: Mining association rules with weighted items. In: Proc. of International Database Engineering and Application Symposium IDEAS-98, IEEE Computer Society, Washington DC (1998), 68-77.

(7) Chen, G., Wei., Q.: Fuzzy association rules and the extended mining algorithms. Information Sciences—Informatics and Computer Science: An International Journal 147 (2002) 201-228.

(8) Chung, F., Lui, Ch.: A post-analysis framework for mining generalized association rules with multiple minimum supports. In: Proc. of KDD-2000 Workshop on Post-Processing in Machine Learning and Data Mining (2000).

(9) Cios, K. J., Sztandera, L. M.: Continuous ID3 algorithm with fuzzy entropy measures. In: Proc. of 1st IEEE International Conference on Fuzzy Systems, San Diego (1992), 469-476.

(10) Clark, P., Niblett, T.: The CN2 induction algorithm. Journal of Machine Learning 3 (1989) 261-283.

(11) Davidovskaja, I., Kovalík, Š.: Usage of fuzzy decision trees for classification tasks. In: Proc. of 5th International Conference on New Information Technologies, Belarus State Economic University, Minsk (2002), 179-183. (in Russian)

(12) Gynesei, A.: A Fuzzy approach for mining quantitative association rules. TUCS Technical Report 336 (2000).

(13) Gyenesei, A.: Mining weighted association rules for fuzzy quantitative items. TUCS Technical Report 346 (2000).

(14) Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proc. of ACM-SIGMOD International Conference on Management Data, ACM, Dallas (2000), 1-12.

(15) Hipp, J., Güntzer, U., Nakhaeizadeh, G.: Algorithms for association rule mining – A general survey and comparison. SIGKDD Explorations 2 (2000) 58-64.

(16) Ichihashi, H., Shirail, T., Nagasaka, K., Miyoshi, T.: Neuro-Fuzzy ID3. Fuzzy Sets and Systems 81 (1996) 151-167.

(17) Klir, G. J.: Where do we stand on measures of uncertainty, ambiguity, fuzziness and the like? Fuzzy Sets and Systems 24 (1987) 141-160.

(18) Levashenko, V., Zaitseva, E.: Usage of new information estimations for induction of fuzzy decision trees. In: Proc. of 3rd IEEE Int. Conference on Intelligent Data Eng. and Automated Learning, Kluwer Publisher, Manchester (2002), 493-499.

(19) Mantaras, R. L.: A distance based attribute selection measure for decision tree induction. Journal of Machine Learning (1991) 103-115.

(20) Michalski, R. S.: On the quasi-minimal solution of the general covering problem. In: Proc. of 5th International Symposium on Information Processing FCIP'69, Bled (1969), 125-128.

(21) J. Mingers. Expert Systems – Rule inducion with statistical data. Journal of The Operations Research Society 38, p. 39-47, 1987.

(22) Mingers, J.: An empirical comparison of selection measures for decision-tree induction. Journal of Machine Learning 3 (1989) 319-342.

(23) Pazúrik, P.: Multi-Interval Discretization of Continuous Attributes. Master's Thesis, University of Žilina, Žilina, 2005. (in Slovak)

(24) Quinlan, J. R.: Discovering rules by induction from large collections of examples. In: D. Michie, editor, Expert Systems in the Micro Electronic Age, Edinburgh University Press, Edinburgh (1979), 168-201.

(25) Quinlan, J. R.: Induction of Decision Trees. Journal of Machine Learning 1 (1986) 81-106.

(26) Savasere, A., Omiecinski, E., Navathe, S.: An efficient algorithm for mining association rules in large databases. Technical Report GIT-CC-95-04, Atlanta (1995)

(27) Shu-Yue, J., Tsang, E., Yengg, D., Daming, S.: Mining fuzzy association rules with weighted items. In: Proc. of IEEE International Conference on Systems, Man and Cybernetics (2000) 1906-1911.

(28) Strikant, R., Agrawal, R.: Fast algorithms for mining associaton rules. In: Proc. of 20th International Conference on Very Large Data Bases, Morgan Kaufmann, Santiago de Chile (1994), 487-499.

(29) Strikant, R., Agrawal, R.: Mining generalized association rules. In: Proc. of 21st International Conference on Very Large Data Bases, Morgan Kaufmann, Zurich (1995), 407-419.

(30) Strikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proc. of ACM SIGMOD International Conference on Management on Data, ACM Press, Montreal (1996), 1-12.

(31) Tani, T., Sakoda, M.: Fuzzy modeling by ID3 algorithm and its application to prediction. In: Proc. of the IEEE International Conference on Fuzzy Systems, Ed. IEEE, San Diego (1992), 923-930.

(32) Tao, F., Murtagh, F., Farid, M.: Weighted association rule mining using weighted support and significance framework. In: Proc. of 9th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, ACM Press, Washington DC (2003), 661-666.

(33) Weber, R.: Fuzzy ID3: A class of methods for automatic knowledge acquisition. In: Proc. of 2nd International Conference on Fuzzy Logic and Neural Networks, Iizuka (1992), 265-268.

(34) Wu, X., Urpani, D.: Induction by attribute elimination. IEEE Transactions on Knowledge and Data Engineering 11 (1999) 805-812.

(35) Yuan, Y., Shaw, M. J.: Induction of fuzzy decision trees. Fuzzy Sets and Systems 69 (1995) 125-139.

(36) Zaki, M. J., Parthasarathy, S., Ogihara, M., Li, W.: New algorithms for fast discovery of association rules. Report: TR651, University of Rochester, (1997).

7. Adresa autora:

Ján Boháčik, Ing.
Katedra informatiky
Fakulta riadenia a informatiky
Žilinská univerzita
Univerzitná 8215/1
010 26 Žilina
Jan.Bohacik@gmail.com

Zaradené na publikáciu v januári 2009 vydavateľstvom časopisu Journal of Information Technologies (ISSN 1337-7469), t.j. Katedrou informatiky FPV UCM v Trnave.