Hodnocení (ne)zajímavosti asociačních pravidel za využití báze znalostí

Přemysl Václav Duben, Stanislav Vojíř

Katedra informačního a znalostního inženýrství, FIS, Vysoká škola ekonomická v Praze
nám. W. Churchilla 1938/4, 13067 Praha 3 - Žižkov

{xdubp00|stanislav.vojir}@vse.cz

Abstrakt. Dolování asociačních pravidel je jednou z populárních data miningových metod, prostřednictvím které lze objevovat zajímavé vztahy v datech. Asociační pravidla jsou využitelná nejen pro analýzu chování zákazníků, ale také pro tvorbu klasifikačních modelů či pro detekci výjimek. Algoritmy pro hledání asociačních pravidel mají však také jednu nevýhodu – velké množství nalézaných výsledků. Při ručním řešení analytické otázky se musí data miningový expert probrat velkým množstvím pravidel a vybrat z nich jen ta opravdu zajímavá. V rámci tohoto příspěvku je představena metoda výběru (ne)zajímavých asociačních pravidel dle jejich podobnosti s pravidly dříve vybranými do báze znalostí. Tato metoda postprocessingu je implementována ve webovém data miningovém systému EasyMiner.

Klíčová slova: asociační pravidla, postprocessing, podobnost asociačních pravidel, báze znalostí, EasyMiner.

1 Motivace

Dolování asociačních pravidel je jedním ze základních typů data miningových úloh, vhodných nejen pro analýzu nákupních košíků, ale také pro objevování složitějších vztahů v datech. Nevýhodou algoritmů pro dolování asociačních pravidel jsou však velké nároky na následné zpracování výsledků. I v rámci jednoduchých úloh lze nalézt opravdu velké množství pravidel, která je následně nutné vyhodnotit z hlediska jejich zajímavosti pro řešenou analytickou otázku. Ruční vyhodnocení všech nalezených pravidel data miningovým či doménovým expertem je časově velmi náročné, neřkuli nemožné. Dlouhodobě jsou tedy hledány možnosti filtrování nalézaných pravidel ať již v průběhu zadání data miningové úlohy, v průběhu procesu dolování či v rámci zpracování jejích výsledků.

V rámci výzkumu realizovaného na Katedře informačního a znalostního inženýrství Vysoké školy ekonomické v Praze v projektech SEWEBAR a posléze EasyMiner (*http://easyminer.eu*) byly analyzovány možnosti zjednodušení procesu řešení úloh dolování asociačních pravidel a výběru adekvátních výsledků. V minulosti šlo o hledání obdobných asociačních pravidel za využití jejich zápisu ve formátu XML a jazyka *XQuery* [1]. Posléze byla v rámci projektu EasyMiner navržena a implementována

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 61-66.

báze znalostí umožňující shromažďování zajímavých a nezajímavých asociačních pravidel do ručně vytvářených *rule setů*, využitelných jak pro ruční vytváření klasifikačních modelů, tak pro sdílení doménových znalostí o potenciální zajímavosti pravidel. [2][3]

V rámci tohoto příspěvku je představena nově implementovaná metoda pro vizuální zjednodušení procházení asociačních pravidel nalezených v rámci data miningové úlohy na základě hodnocení jejich podobnosti s pravidly dříve přidanými do zvoleného rule setu. Daná metoda je implementována v grafickém uživatelském rozhraní webového data miningového systému EasyMiner (*http://easyminer.eu*).

Z hlediska zasazení do kontextu aktuálního výzkumu jsou v rámci dolování asociačních pravidel v systému EasyMiner využívány metody omezení nalézaných pravidel při zadávání data miningové úlohy a práce s doménovou znalostí (uloženou v bázi znalostí). Uživatel má zároveň možnost volitelně využít prořezávání pravidel za využití algoritmu CBA [4]. Z relevantních alternativních přístupů lze jmenovat například metody dolování asociačních pravidel s omezeními [5], již zmíněné metody prořezávání nalézaných pravidel, nalézání zajímavých pravidel za využití ontologií [6] či metody vizualizace a shlukování obdobných pravidel [7].

2 Báze znalostí

V rámci projektu EasyMiner je implementována komplexní znalostní báze shromažďující informace o atributech využívaných pro řešení data miningových úloh a jejich metodách předzpracování a také o pravidlech shromážděných v rámci rule setů uložených taktéž v bázi znalostí. Báze znalostí byla poprvé představena v [2]. Posléze byl návrh dokončen a implementován v [3].

Báze znalostí se vnitřně skládá z trojice základních typů uložených informací: 1. metainformace o zpracovávaných datech – meta-atributy a jejich formáty, 2. informace o vhodných metodách předzpracování konkrétních typů dat (formátů meta-atributů) na atributy použitelné pro data mining, 3. konkrétní uložená pravidla (a jejich seskupení do rule setů).

Jednotlivé *meta-atributy* označují základní typy dat, ve své podstatě jde o abstraktní skupiny atributů – např. „věk“. Dané typy dat se posléze vyskytují v konkrétních *formátech* – např. „věk v letech“.

Pro jednotlivé *formáty* jsou definovány *metody předzpracování* uživatelem nahraných dat do podoby datových *atributů* použitelných pro dolování asociačních pravidel. Báze znalostí obsahuje informace o již definovaných metodách seskupení hodnot pomocí množin či intervalů.

V rámci řešení data miningové úlohy nejprve uživatel nahraje svá vlastní data do aplikace EasyMiner. Následně je vyzván k jejich předzpracování – vytvoření atributů z datových sloupců obsažených v nahraných datech. K předzpracování jsou využívány metody ručně definovaných množin či ručně či automaticky definovaných intervalů.

V rámci předzpracování jsou data *namapována na meta-atributy* uložené v bázi znalostí (a jejich *konkrétní formáty*). Ve výchozím stavu jsou formáty automaticky

vytvářeny na základě nahraných dat, volitelně má však uživatel možnost znovu-použít již existující metodu předzpracování – čímž dojde k namapování datového sloupce na již existující formát/meta-atribut. Na základě mapování na formáty meta-atributů poté mohou být vyhodnoceny jako podobné (či totožné) také atributy, které se sice liší ve svých názvech, ale vznikly ze stejných datových sloupců.

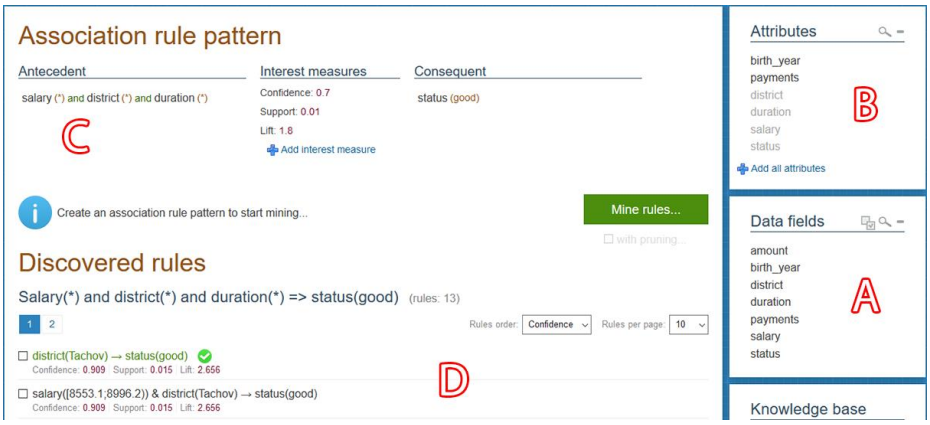
V průběhu dolování má následně uživatel možnost přidávat vybraná asociační pravidla do báze znalostí – za účelem vytváření klasifikačních modelů či za účelem hodnocení (ne)zajímavosti u dalších, podobných pravidel.

3 Hodnocení zajímavosti asociačních pravidel

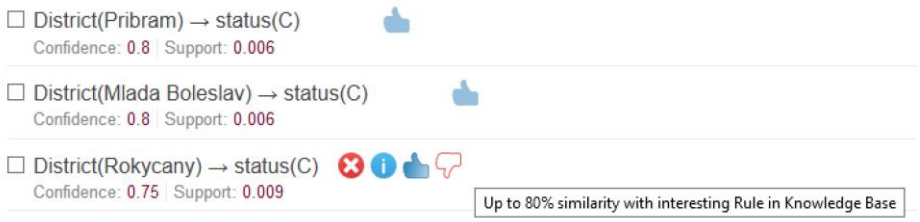
V rámci procesu dolování asociačních pravidel v grafickém rozhraní systému Easy-Miner (**Obr. 1**) uživatel nejprve předzpracuje datové sloupce na atributy (**Obr. 1** A, B). Následně definuje data miningovou úlohu vytvořením vzoru hledaných asociačních pravidel (**Obr. 1** C) – tj. umístí zvolené atributy do antecedentu či konsekventu vzoru pravidel a definuje minimální hodnoty požadovaných měr zajímavosti.[1] Po spuštění dolování jsou nalezené výsledky zobrazeny přímo v rámci uživatelského rozhraní v sekci *Discovered rules* (**Obr. 1** D). V rámci této sekce má uživatel možnost pravidla řadit dle použitých měr zajímavosti a také označovat pravidla, která považuje za zajímavá či naopak za nezajímavá.

Při procházení výsledků má uživatel možnost uložit vybraná pravidla do báze znalostí – pomocí jejich označení za zajímavá či nezajímavá. Následně jsou všechna ostatní pravidla hodnocena z hlediska podobnosti s takto uloženými pravidly a dle jejich podobnosti je k nim doplněna informace o tom, zda je nejpodobnější pravidlo uloženo jako (ne)zajímavé. Dané porovnání je realizováno nejen ve výsledcích jedné dílčí data miningové úlohy, ale také u všech úloh následujících. Tj. pokud uživatel např. označí zvolené pravidlo za nezajímavé, bude jako nezajímavé označeno i v případě, že se vyskytne ve výsledcích některé z následujících úloh. Označování zajímavosti pravidel je znázorněno na **Obr. 2**. K aktivaci vyhodnocení podobnosti daného jednotlivých pravidel je využíván AJAX.

[1] Uživatel má k dispozici míry zajímavosti *konfidence*, *podpora* a *lift*. Volitelně lze zapnout také automatické prořezání zobrazených výsledků algoritmem *CBA* [4].



Obr. 1. Uživatelské rozhraní systému EasyMiner



Obr. 2. Ukázka vyhodnocení podobnosti asociačních pravidel s pravidly ve znalostní bázi

Pro porovnání s obsahem báze znalostí je využívána množina atributů nacházejících se v antecedentu či konsekventu konkrétního asociačního pravidla a posléze jejich konkrétních hodnot. Celé porovnání je za účelem rychlosti realizováno jako dvoufázové. V první fázi dochází u pravidel nalezených nad jedním konkrétním datovým souborem. V tomto případě dochází nejprve k porovnání textové podobnosti antecedentu a konsekventu daných pravidel.

V rámci podrobného porovnání je pravidlo: 1. rozděleno na jednotlivé atributy obsažené v antecedentu/konsekventu, 2. k jednotlivým atributům jsou v rámci báze znalostí dohledány odpovídající formáty meta-atributů (na základě informace o předzpracování dat), 3. dochází k dekomponování hodnot. Tímto způsobem jsou porovnávána i pravidla získaná na základě analýz rozdílných datových souborů, pokud dané atributy využívají stejné formáty meta-atributů.

Z hlediska dekomponování je například pravidlo

salary_intervals([10000;20000]) & district(Tachov) → status(good)

nejprve rozděleno na jednotlivé atributy. Následně jsou k těmto atributům dohledány jejich konkrétní formáty meta-atributů – antecedent: *salary/czk([10000,20000]), district/towns(Tachov)*, konsekvent: *status/sets(good)*. V posledním kroku jsou hodnoty

atributů nahrazeny identifikátory intervalů či množin hodnot definovaných v bázi znalostí. Z hlediska daného příkladu je hodnota atributu *good* definována jako množina původních hodnot datového sloupce {„A“, „B“}.

Pro stanovení podobnosti je nejprve vyhodnocena shoda atributů v antecedentu/konsekventu pravidla a posléze shoda jejich hodnot. Každá část pravidla má váhu 50 % celkového hodnocení. Pro uživatelskou srozumitelnost je následně podobnost převedena na hodnoty ze škály 0 – 100 %. V současné variantě nejsou porovnávány hodnoty použitých měr zajímavosti.[2]

4 Závěr

V rámci tohoto příspěvku byla představena metoda hodnocení zajímavosti nalézaných asociačních pravidel za využití pravidel dříve ručně přidaných do báze znalostí. Daná metoda byla implementována v systému EasyMiner.

V rámci dalšího vývoje by mělo dojít k podrobnějšímu uživatelskému a výkonnostnímu testování implementovaných algoritmů, testování vhodnosti (ne)zahrnutí hodnot měr zajímavosti pro stanovení podobnosti pravidel a k usnadnění ručního mapování nahraných dat na existující formáty meta-atributů báze znalostí.

Literatura

1. Kliegr, T., Hazucha, A., Marek, T.: Instant feedback on discovered association rules with PMML-based query-by-example. In: International Conference on Web Reasoning and Rule Systems, Springer Berlin Heidelberg (2011), pp. 257-262.
2. Vojíř, S.: Concept of Semantic Knowledge Base for Data Mining of Business Rules. In: Znalosti 2014 Exhibice, Edukace a nacházení Expertů - Exhibition, Education and Expert finding. Praha: KIZI FIS (2014) pp. 132-136.
3. Vojíř, S.: Učení business rules z výsledků dolování GUHA asociačních pravidel. Vysoká škola ekonomická v Praze (2016).
4. Kliegr, T., Kuchař, J., Sottara, D., Vojíř, S.: Learning Business Rules with Association Rule Classifiers. In: International Workshop on Rules and Rule Markup Languages for the Se-mantic Web, Springer (2014), pp. 236–250.
5. Dan Nguyen, Bay Vo: Mining Class-Association Rules with Constraints. In: {Knowledge and Systems Engineering, Springer (2014) pp. 307-318.
6. Saravanam, D., Vijayalakshmi, S., Joseph, D.: Finding of Interesting Rules from Association Mining by Ontology and Page Ranking. International Journal for Modern Trends in Science and Technology (2017) 03 01, pp. 46-50
7. Hahsler, M., Karpienko, R.: Visualizing association rules in hierarchical groups. Journal of Business Economics (2017) 87 3, pp. 317-335.
8. Duben, P. V.: Filtrování zajímavých pravidel v systému EasyMiner. Vysoká škola ekonomická v Praze (2017).

[2] Dle základního uživatelského testování není nezahrnutí měr zajímavosti problémem. V rámci dalšího výzkumu by měly být ověřeny možnosti stanovení priorit pro podobnost antecedentu/konsekventu.

Poděkování: Tento článek vznikl díky podpoře z projektu IGA 29/2016 Vysoké školy ekonomické v Praze.

Annotation:

Rating of (Non)Interestingness of Association Rules using Knowledge Base

Data mining of association rules is one of popular data mining method used to discovering of interesting relationships hidden in data. Association rules are applicable not only for customer behavior analysis, but also for building of classification models and for detection of exceptions. However, algorithms for discovery of association rules have one disadvantage – many results, founded even in simple data mining tasks. During the solving of analytical question, the data mining expert must deal with a big count of rules and choose the interesting of them. This paper introduces a method of selection of (non) interesting association rules according to their similarity to rules previously stored in knowledge base. This post-processing method is implemented in the web data mining system EasyMiner.