Analýza cloudového řešení akademického nástroje pro dolování pravidel z databází

Václav Zeman
Katedra informačního a znalostního inženýrství
Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
vaclav.zeman@vse.cz

Abstrakt: *Webová aplikace EasyMiner je akademický nástroj pro získávání znalostí z malých i velkých dat ve formě asociačních pravidel. Systém využívá prostředí Apache Hadoop a Apache Spark pro zpracování velkých datových zdrojů na výpočetním clusteru MetaCentra sdružení CESNET. Aplikace se skládá z několika mikroslužeb, které vykonávají různé operace z oblasti strojového učení a jako celek tvoří data miningový software fungující jako cloudová webová služba - SaaS.*

Klíčová slova: dolování asociačních pravidel, data mining, machine learning, klasifikace, detekce anomálií, byznys pravidla, MLaaS, SaaS, cloud computing, big data, Hadoop, Spark

Abstract: *EasyMiner is a web service for association rules mining. A new version of this tool uses Apache Hadoop and Apache Spark for big data analysis in the MetaCloud of the CESNET association. The application consists of several services for dataset uploading into a server site, preprocessing, association rules discovery and classification based on associations. All services communicate with each other through REST APIs and form a complex software working as a service in the cloud.*

Keywords: association rules mining, data mining, machine learning, classification, anomaly detection, business rules, MLaaS, SaaS, cloud computing, big data, Hadoop, Spark

1. Úvod

Akademický nástroj EasyMiner[1] je webová služba, vyvíjena převážně na katedře informačního a znalostního inženýrství Vysoké školy ekonomické v Praze, se zaměřením na dolování asociačních pravidel z databází (Agrawal, et al., 1993). Aplikace poskytuje grafické uživatelské rozhraní (viz **Obrázek 1**) a je schopna vykonat všechny nutné operace pro získávání znalostí z dat od nahrávání datových zdrojů přes předzpracování až po samotné dolování a interpretaci výsledků. Nová verze tohoto nástroje dokáže zpracovat i velká data díky nasazení do prostředí Apache Hadoop[2] a Apache Spark[3] a lze ji použít pro akademické účely bez jakýchkoliv omezení s využitím výpočetního clusteru na půdě MetaCentra[4] sdružení CESNET. Mezi nejdůležitější operace, které lze v systému EasyMiner vykonávat, patří:

- Proudové nahrávání datových zdrojů do datového úložiště

[1] *http://www.easyminer.eu*

[2] *http://hadoop.apache.org/*

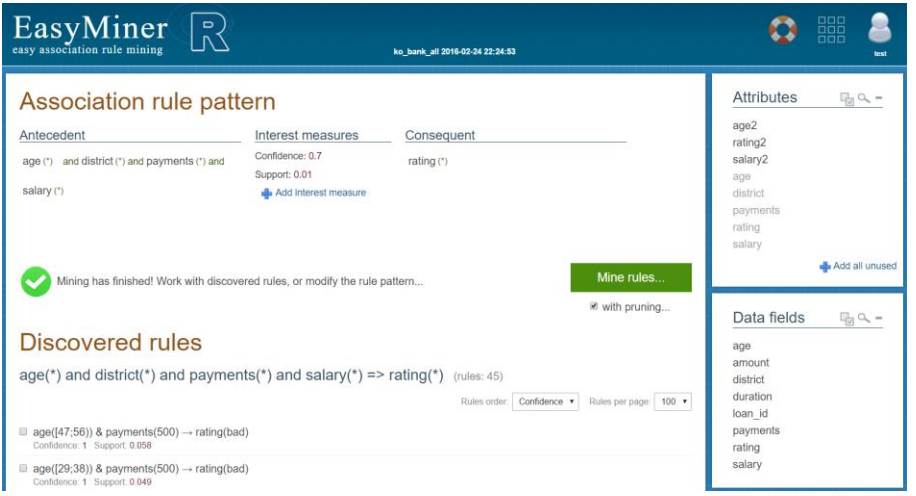
[3] *http://spark.apache.org/*

[4] *https://www.metacentrum.cz/cs/*

- Transformace a předzpracování dat pro účely rychlejšího dolování znalostí
- Dolování asociačních pravidel dle uživatelských požadavků
- Tvorba klasifikačních modelů ze získaných pravidel
- Manipulace s množinou získaných pravidel

Nástroj lze tedy v kontextu cloudových služeb zařadit do kategorie MLaaS (Machine Learning as a Service) (Ribeiro, et al., 2015) a lze jej rovněž použít jako alternativu ke komerčním produktům poskytující nástroje pro strojové učení či dolování znalostí, jako jsou např. BigML.com či Microsoft Azure ML.

Jedním z cílů této práce je poskytnout náhled na funkce a architekturu dolovacího systému EasyMiner a vysvětlit účel použití nástroje pro extrakci znalostí a jejich využití v praxi (viz kapitoly 2, 3 a 4). V dalších kapitolách (5 a 6) jsou popsány jiné přístupy či jiné služby strojového učení sloužící k získávání znalostí z dat a k vykonávání prediktivní analýzy. Na základě této rešerše a popisu existujících služeb typu MLaaS lze lépe vystihnout slabé a silné stránky nástroje EasyMiner a pochopit výhody či nevýhody pravidlového přístupu oproti jiným řešením.



Obrázek 1 Náhled na grafické uživatelské rozhraní nástroje EasyMiner.

2. Pravidla jako reprezentace znalostí

Proces získávání pravidel z databází patří mezi základní úlohy v oblasti dolování dat (angl. data mining), v němž dochází k objevování nových znalostí (Aggarwal & Han, 2014) (Friedman, et al., 2001) (Tan & others, 2006). Pravidla jsou tedy jedním z prostředků pro reprezentaci znalostí a dají se dále v praxi využít pro jiné úlohy, ať už z oboru informatiky či business intelligence. Pravidlo zapisujeme buď ve tvaru IF-THEN nebo jako implikaci představující podmíněný výrok definovaný v matematické logice jako:

$$A \Rightarrow B$$

Toto pravidlo ve tvaru implikace se skládá z levé a pravé strany (antecedent a konsekvent) a dá se interpretovat následovně „pokud platí výrok A, potom platí i výrok B“.

Uvažujme databází klientů nějaké banky poskytující úvěry. V rámci každého úvěru máme podrobná data o klientovi včetně informace o úspěšnosti splácení daného úvěru. Na základě takovýchto dat lze sofistikovanou metodou hledat pravidla napříč celou databází v následující podobě:

$$\text{Plat}(\text{vysoký}) \wedge \text{Bydliště}(\text{Praha}) \Rightarrow \text{Splácí}(\text{dobře})$$

$$\text{Věk}([45;50]) \wedge \text{PočetDětí}(3\text{-}4) \wedge \text{VlastníAuto}(\text{Ano}) \Rightarrow \text{Splácí}(\text{špatně})$$

V oblasti dolování dat se obvykle setkáváme s pojmem dolování asociačních pravidel, což je specifická úloha pro hledání pravidel, u kterých dále definujeme míry zajímavostí, podle kterých určujeme relevanci nalezených pravidel (Agrawal, et al., 1993). Jsou to obvykle:

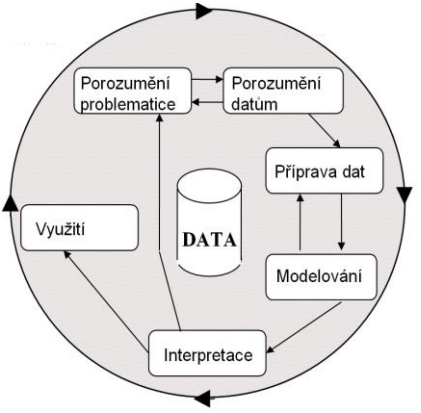
- **spolehlivost** pravidla (angl. confidence), tj. podmíněná pravděpodobnost výskytu pravé strany pravidla dané levou stranou pravidla
- **podpora** pravidla (angl. support), tj. celkový počet záznamů v databázi, které pokrývají dané pravidlo

Pokud tedy vezmeme první pravidlo z příkladu a zjistíme, že má spolehlivost 90% a podporu 5% (např. 50 záznamů z 1000 evidovaných), znamená to, že 90% zákazníků z Prahy a vysokým platem dobře splácí a zároveň takovýchto lidí je celkově 5% ze všech klientů, které banka eviduje. Takovouto znalost může banka použít např. pro marketingové účely či k predikci chování nových zákazníků.

Cílem úlohy pro hledání nových znalostí v datech je objevování právě takových pravidel, která jsou pro nás zajímavá a mají vysokou míru spolehlivosti a podpory (Agrawal, et al., 1994). Z množiny výstupních pravidel lze poté odhalit příslušné znalosti, které nám mohou pomoci řešit různé byznys problémy. Relevantní pravidla si lze představit i jako vzory, které se v databázi vyskytují s vysokou frekvencí. Tyto frekventované vzory můžou pomoci k lepšímu porozumění datům, ať už se jedná o chování zákazníků, charakteristiky produktů apod. Sada pravidel může být dále použita k řešení úloh strojového učení, např. k prediktivní analýze nebo k odhalování anomálií v datech (Bing, et al., 1998) (He, et al., 2005). Výhodou použití pravidlových modelů je snadná čitelnost výsledků, kterým dokáží porozumět i nedatový experti.

3. EasyMiner – nástroj pro získávání pravidel

EasyMiner je webová služba, která se primárně orientuje na hledání asociačních pravidel v databázích a je navržena tak, aby bylo možné s její pomocí řešit kompletní workflow pro dolování dat dle metodiky CRISP-DM (viz **Obrázek 2**) (Wirth & Hipp, 2000). Jednotlivé fáze procesu dolování, definované v této metodice, jsou separovaně implementovány jako samostatné mikroslužby v rámci celého systému EasyMiner.



Obrázek 2 Metodika CRISP-DM pro získávání znalostí z databází.

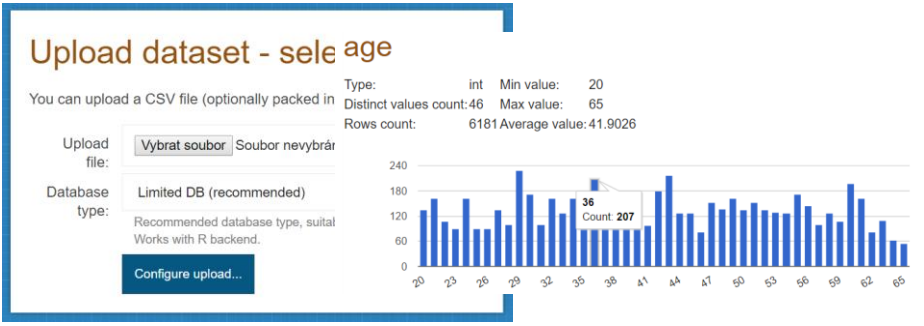
Jedná se tedy o služby pro nahrávání dat (včetně náhledů do základních statistik o vstupních datech), transformaci a předzpracování dat, dolování neboli modelování, interpretaci výsledků a finální využití exportováním výsledného modelu nebo vygenerováním analytické zprávy. Veškeré algoritmy od nahrávání vstupních dat, přes přípravu dat až po samostatné hledání pravidel jsou tedy poskytovány jako služba typu SaaS, která je nasazena na akademickém výpočetním cloudu. Všechny dolovací operace se provádějí na straně serveru resp. množiny serverů a dají se spouštět jak pomocí RESTového API, tak i pomocí grafického uživatelského rozhraní.

3.1 Případ užití: banka poskytující úvěry

Uvažujme příklad z předchozí kapitoly; mějme banku poskytující úvěry, jejíž problémem je vysoké procento špatných úvěrů (zákazníci splácejí úvěry s problémy). Banka se proto rozhodla upravit marketingovou strategii a zaměřit se pouze na takové potenciální klienty, u kterých je vysoká pravděpodobnost, že budou splácet bez problémů. Cílem je vytvořit model, který na základě údajů o žadateli o úvěr dokáže predikovat schopnost splácení potenciálního klienta. Na vstupu máme tabulku o dosavadních zákaznících, kde každý řádek představuje právě jednoho klienta a jednotlivé sloupce představují jeho rysy (např. věk, pohlaví, plat, zaměstnání, výše splátky, bydliště apod.). Dále je v datech zahrnuta informace o tom, zdali klient splácí s problémy nebo bez problémů. Úlohou je nalézt taková pravidla kde konsekventem (neboli pravou stranou pravidla) je rys definující kvalitu úvěru, kterou chceme predikovat. Z těchto pravidel ve finále zkonstruujeme klasifikační model, který nám bude predikovat kvalitu úvěru žadatele na základě jeho rysů.

Nahrání dat

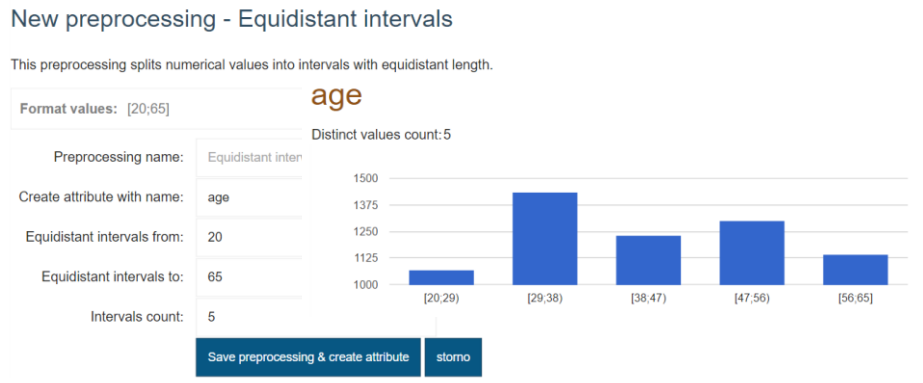
Vstupní data ve formě tabulky lze jednoduchým způsobem nahrát do EasyMineru. Poté si lze jednoduše vizualizovat jednotlivé sloupce v podobě histogramů a u numerických sloupců si nechat zobrazit základní statistické informace jako (minimální hodnota, maximální hodnota, průměr, medián, směrodatná odchylka apod.). Rovněž není problém si zpětně procházet jednotlivé řádky v tabulce.



Obrázek 3 Nahrávání dat a vizualizace atributů v systému EasyMiner

Příprava dat

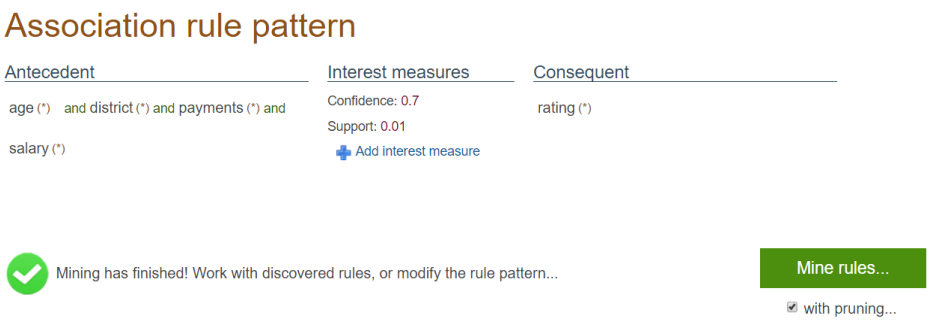
Aby bylo možné použít jednotlivé sloupce k hledání pravidel je nutné je nejdříve připravit. Sloupce lze buď připravit hromadně, nebo individuálně. Při operaci přípravy lze data upravit resp. předzpracovat dle implementovaných metod, jako jsou např. automatická diskretizace numerických sloupců, ruční vytváření intervalů nebo slučování hodnot do vlastních kategorií. Předzpracovaný sloupec je již připraven ke spuštění procesu pro hledání pravidel.



Obrázek 4 Předzpracování atributů v systému EasyMiner

Hledání pravidel

V další fázi si již můžeme namodelovat vzor pro pravidla, která chceme dostávat na výstupu. Lze specifikovat, které sloupce a hodnoty chceme mít na levé a pravé straně pravidla, dále si lze určit minimální prahy měr zajímavostí jako je např. spolehlivost a podpora. Nástroj poté hledá pouze ta pravidla, která přesahují minimální prahy těchto nadefinovaných měr zajímavostí a rovněž se shodují s namodelovaným vzorem. Během tohoto kroku lze rovněž zvolit, zdali chceme pravidla přizpůsobit pro prediktivní analýzu.



Obrázek 5 Modelování vzoru pro dolování asociačních pravidel.

Interpretace a využití pravidel

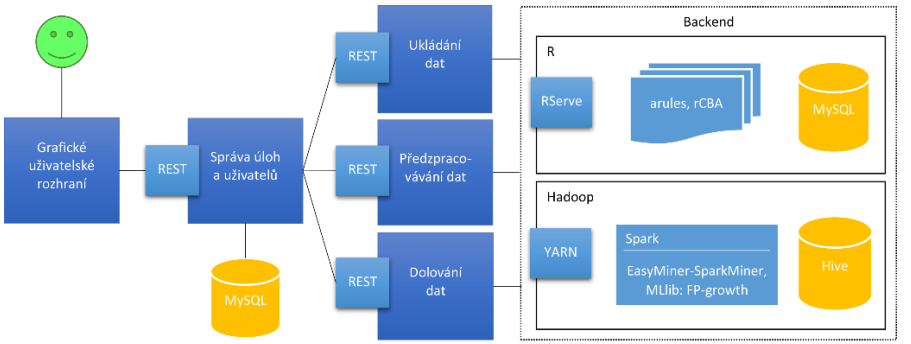
Nástroj nám během krátké doby vrátí všechna pravidla, která splňují nadefinované podmínky. Pravidla jsou vizualizována ve tvaru implikace a zahrnují základní statistické informace o daném pravidlu, tj. spolehlivost pravidla, podpora pravidla apod. S pravidly lze v systému manipulovat, dají se hodnotit, upravovat a exportovat do standardizovaného formátu PMML[5] pro další využití. Pro výše zmíněný případ banky nám systém vrátí všechna relevantní pravidla, která nám implikují kvalitu úvěru na základě různých rysů klientů (viz **Obrázek 6**). Tato pravidla lze stáhnout a použít je k vytvoření klasifikačního modelu predikující kvalitu úvěru žadatelů o úvěr. Tím banka dokáže automaticky posoudit, zdali žadateli úvěr poskytnout či nikoliv.

Obrázek 6 Výsledná pravidla vrácená systémem EasyMiner

[5] *Predictive Model Markup Language*

3.2 Architektura systému EasyMiner

Služba EasyMiner běží na pozadí ve dvou módech: *limited* a *unlimited*. Oba režimy zvládají stejné operace, avšak na pozadí se jinak pracuje s daty a používá se odlišná infrastruktura. Režim *limited* je určen pro malé a středně velké datasety (řádově do stovek MB), které jsou zpracovávány rychle v paměti počítače s použitím jednoho výpočetního uzlu. Pro práci s velkými daty (více než stovky MB) je určen režim *unlimited*, který využívá ke všem datovým operacím výpočetní uzly v distribuovaném Hadoop prostředí MetaCloudu sdružení CESNET. V tomto módu je možné zpracovávat velká data a hledat pravidla paralelně na více počítačích uvnitř velkého výpočetního clusteru (Harper, 2015) (White, 2012). Systém EasyMiner je tedy navržen i pro práci s velkými daty, které nelze standardně zpracovat na jednom výpočetním uzlu.



Obrázek 7 Architektura systému EasyMiner.

4. Byznys problémy a jejich řešení pomocí nástroje EasyMiner

Aplikace EasyMiner převážně slouží k hledání asociačních pravidel, které se dají využít pro řešení mnoha byznys problémů. Tyto problémy jsou obvykle spojovány se základními úlohami strojového učení, které jsou rovněž řešitelné v systému EasyMiner s použitím nalezených pravidel.

4.1 Predikční úlohy

Z objevených pravidel dokáže EasyMiner sestavit klasifikační model, pomocí kterého se dají predikovat hodnoty vybraného sloupce ze vstupních dat na základě ostatních sloupců. K tomu nástroj používá známou metodu CBA (Liu, et al., 2001) (Bing, et al., 1998), která pravidla přizpůsobuje právě k tvorbě klasifikačních modelů.

Uvažujme případ banky z kapitoly 3.1. Na základě metody CBA lze vytvořit z pravidel klasifikační model, který bude predikovat schopnosti splácení klienta na základě jeho rysů. Kvalita úvěru je rozdělena např. do kategorií *špatná* a *dobrá*. Potom lze získat dvě skupiny pravidel, kde první skupina je tvořena takovými rysy zákazníků, u kterých je vyšší pravděpodobnost, že budou splácet s problémy. Druhá skupina naopak tvoří sadu vlastností zákazníků, u kterých je vyšší pravděpodobnost, že budou splácet

bez problémů. Z těchto pravidel se sestaví posloupnost rozhodovacích podmínek, které nám dokáží predikovat kvalitu úvěru potenciálního klienta (viz **Obrázek 6**).

4.2 Explorační úlohy

Nástroj EasyMiner pomáhá k lepšímu porozumění datům na základě nalezených pravidel, které nám dokáží identifikovat jednotlivé vztahy mezi atributy (sloupci) jednotlivých instancí (řádků). Silnou stránkou je schopnost nadefinovat si vzor pravidla a zaměřit se pouze na určitý segment rysů, které uživatele skutečně zajímá. Pokud tedy naleznu pravidlo, které nám říká, že daný rys se v kombinaci s jiným rysem vyskytuje s vysokou spolehlivostí a frekvencí, potom mezi těmito rysy existuje nějaká spojitost, kterou se nám podařilo odhalit (Fürnkranz & Kliegr, 2015). Tuto nově objevenou znalost lze poté využít k modifikaci dosavadního byznys plánu a eliminaci nerentabilních částí podniku.

Uvažujme pravidlo z případu 3.1, které nám vyjadřuje, že pokud je klient z Jihomoravského kraje a zároveň pracuje v lékařství, potom s vysokou pravděpodobností bude splácet s problémy. Takto objevená znalost nám může velmi pomoci k nalezení silných a slabých stránek ve firmě a k identifikaci problémových segmentů.

4.3 Detekce anomálií

Úlohu pro hledání pravidel lze v rámci nástroje EasyMiner přizpůsobit na hledání nefrekventovaných vzorů v datech, které nám pomáhají odhalit anomálie neboli odlehlé části v datech (He, et al., 2005). Jako anomálie je uvažována nějaká instance (řádek) v datech (tabulce), která se výrazně liší v některých rysech od ostatních instancí. To nám může pomoci k odhalení podezřelých klientů či elementů, které se nechovají tak jak je obvyklé. Systém EasyMiner např. na základě případu 3.1 identifikuje následující nefrekventovaný vzor

{Plat(nízký), Půjčka(Vysoká), SplácíBezProblémů(ANO)}

Pokud se takováto kombinace rysů vyskytuje v datech s velmi malou frekvencí, potom ji lze považovat za anomálii. Na základě této znalosti můžeme daného klienta s takovouto množinou rysů blíže zkoumat a hledat příčinu této odchylky.

4.4 Hledání byznys pravidel

Byznys pravidla jsou tzv. tvrzení, která definují či omezují některé aspekty podniku za účelem prosazení podnikové struktury či řízení a ovlivnění chování podniku (Vojíř, 2016). Množina byznys pravidel je tedy jednou z možností pro modelování podniku a lze je využít pro znalostní management. V reprezentaci řady standardů jsou samotná byznys pravidla zaznamenávána v IF-THEN tvaru, které se dají zkonstruovat z asociačních pravidel vrácených systémem EasyMiner (Vojíř, et al., 2014) (Novotný & Průcha, 2013). Pravidla tohoto typu se poté dají použít v dalších systémech pro správu byznys pravidel, jako jsou např. Drools nebo OpenRules (Vojíř, et al., 2013).

Mějme množinu pravidel, které nám dokáží predikovat schopnost splácení daného úvěru nějakého potenciálního zákazníka. Na základě této predikce lze v systému pro správu byznys pravidel namodelovat taková pravidla, která nám definují, jaký alternativní typ půjčky klientovi poskytnout, pokud byla predikována špatná kvalita úvěru. Byznys pravidla si lze tedy představit jako síť pravidel, která na sebe navazují a

pomocí kterých lze automaticky řídit různé byznys procesy. Každou vrstvu těchto pravidel lze vygenerovat v systému EasyMiner, na základě nějakých trénovacích dat, a finálně importovat do systému pro správu byznys pravidel (např. Drools), čímž se nám může velmi zjednodušit a částečně zautomatizovat modelování byznys procesů a jejich řízení.

4.5 Práce s velkými daty

Při řešení výše zmíněných byznys problémů může u reálných firemních dat docházet k problémům spojených s pojmem big data. Big neboli velká data jsou takové soubory dat, které nelze analyzovat standardními výpočetními prostředky a nástroji (Zikopoulos, et al., 2011). Tato data jsou tak veliká, že k jejich zpracování je nutné použít výpočetní cluster neboli sadu vzájemně propojených počítačů, které nám dovolují provádět daný výpočet paralelně. Vznik takto velkých souborů může nastat velmi rychle při pozorování objektů v reálném čase nebo při sledování toků v sítích, např. záznamy hovorů v telekomunikačních sítích apod.

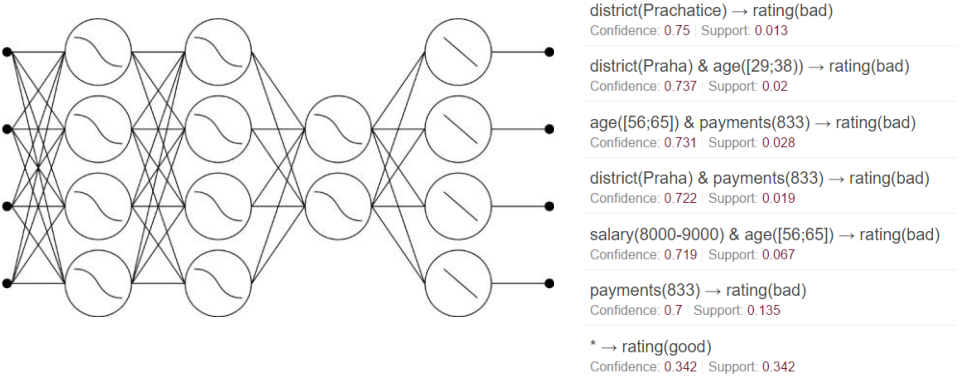
Úloha pro hledání asociačních pravidel je vhodná pro velká data a existuje mnoho stávajících řešení, které dokáží tuto úlohu paralelizovat v rámci nějakého distribuovaného systému (Li, et al., 2008). Nástroj EasyMiner dovoluje uživateli řešit zmíněné byznys problémy i pro velká data díky nasazení do prostředí Apache Hadoop (Zaharia, et al., 2012). V současnosti je nástroj EasyMiner součástí MetaCloudu sdružení CESNET, jenž představuje velkou množinu výpočetních uzlů, na kterých lze provádět hledání pravidel paralelně, a tedy i pro velká data s uspokojivou odezvou. Další výhodou je, že systém EasyMiner používá pro predikční či explorační úlohy právě pravidla, která se dají ve velkých datech hledat velmi efektivně a rychle (Agrawal, et al., 1994) (Fürnkranz, et al., 2012).

5. Výhody a nevýhody pravidlových modelů oproti jiným řešením

Z předchozích kapitol je zřejmé, že nástroj EasyMiner používá k řešení výše zmíněných byznys problémů pravidlový přístup. V této kapitole jsou popsány výhody a nevýhody takovéhoto řešení v porovnání s jinými metodami strojového učení, které definované problémy dokáží řešit odlišnými postupy (Friedman, et al., 2001).

Pro tvorbu klasifikačních modelů schopných cokoliv predikovat existuje celá řada nástrojů a algoritmů. Ať už to jsou běžné přístupy strojového učení, jako jsou rozhodovací stromy, neuronové sítě apod. nebo v dnešní době velmi oblíbená disciplína věnující se hlubokému učení – tzv. **deep learning** (moderní neuronové sítě) (Schmidhuber, 2015), mají modely založené na pravidlech stále svoje pevné místo v oblasti strojového učení (Fürnkranz, et al., 2012). Klasifikační modely, které jsou výstupem hlubokého učení nebo nějakých regresních analýz působí pro běžného uživatele jako černé skříňky, které dokáží „něco“ predikovat, ovšem uživatel přesně neví, jak byl tento predikát odvozen. Uvnitř této černé skříňky je nějaký komplexní matematický model, do kterého se nedá jednoduchým způsobem zasahovat či v něm jednoduše hledat nějaké souvislosti (Tan & others, 2006). Tyto ne snadno interpretovatelné „black box“ modely ovšem svoji funkci plní, zejména ty co jsou výstupem hlubokého učení, a to s velikou mírou přesnosti.

Naopak pravidlové modely, i když nemusejí dosahovat tak dobrých výsledků jako modely hlubokého učení (Kliegr & Kuchař, 2015), mají jednu velikou výhodu a tou je transparentnost kompletního rozhodovacího klasifikátoru sestaveného pouze množinou pravidel (Chorowski, 2012). Tato pravidla jsou snadno čitelná a manipulovatelná; uživatel může lehce porozumět tomu, jak a podle čeho model rozhoduje a případně ho jednoduše rozšířit vlastními pravidly (Kliegr, et al., 2014). Tento fakt může být ve finále pro uživatele podstatnější než jen samotná přesnost klasifikátoru. Mezi hlavními výhodami použitím pravidlového přístupu k tvorbě klasifikátorů v kontrastu s metodami generující „black box" modely patří hlavně srozumitelnost (jednoduchost), manipulovatelnost, rychlost a objektivita.



Obrázek 8 Black box klasifikační model vs pravidlový model.

Srozumitelnost

Hlavní výhodou je srozumitelnost a čitelnost výsledků, které systém EasyMiner uživateli nabízí. Vrácená pravidla jsou snadno čitelná a dají se v nich velmi jednoduše hledat souvislosti a nové znalosti. Při užití pravidel k prediktivní analýze lze velmi snadno nahlédnout do jádra celého modelu a vidět jakým způsobem množina pravidel klasifikuje vstupní data. Naopak modely, které jsou výstupem nějakého hlubokého učení či jednodušší modely používající lineární či pravděpodobnostní funkce jsou bohužel netransparentní a hodí se spíše na klasifikaci takových struktur, které lze jednodušeji převést do matematických modelů (např. audiovizuální soubory, toky v sítích, pozorování pohybu objektů apod.). Naopak pravidlové modely se více hodí pro analýzu reálných dat ve firmách, jako jsou informace o klientech, zaměstnancích, fakturách, produktech apod. (Friedman, et al., 2001).

Dalším důležitým modelem prediktivní analýzy, který se již více podobá pravidlovým strukturám, je rozhodovací strom. Platí, že jakýkoliv rozhodovací strom lze převést na množinu pravidel a obráceně. Množina pravidel je ovšem mnohem jednodušší datová struktura, není tak komplexní a pravidlové modely jsou obecně menší a jednodušší (Fürnkranz, 2011).

Manipulovatelnost

Seznam vrácených pravidel aplikací EasyMiner lze velmi jednoduše redukovat, upravovat či doplňovat. Je tedy možné klasifikační modely modifikovat dle

uživatelských vstupů přidáním vlastních či odebráním existujících pravidel. Většina ostatních klasifikačních modelů touto možností nedisponují (Freitas, 2000). Pro rozhodovací stromy je zásah do jednotné datové struktury mnohem větším problémem než do množiny pravidel; proto může být použití pravidlových modelů výhodnější právě s ohledem na manipulovatelnost.

Rychlost

Pro hledání asociačních pravidel dnes existuje celá řada efektivních metod, které dokáží z ohromného stavového prostoru (tj. všechna pravidla, která můžeme vygenerovat) vybrat malou podmnožinu těch nejzajímavějších pravidel (Agrawal, et al., 1994). Algoritmy pro dolování asociačních pravidel jsou rovněž lehce škálovatelné, tudíž lze celý proces hledání pravidel nasadit do distribuovaného systému stejně jako v případě aplikace EasyMiner, která využívá výpočetní cluster v prostředí Apache Hadoop (Li, et al., 2008). V případě klasifikátorů lze tedy získávat modely mnohem rychleji, a to i pro velká data, než u jiných komplikovanějších metod strojového učení.

Objektivita

Systém EasyMiner dokáže automaticky vytvářet pro numerických atributy intervaly hodnot, čímž se výrazně snižuje počet hodnot v datech, zvyšuje se objektivita modelu a rovněž se touto metodou dají eliminovat anomálie, které mohou vést k přeučení výsledného klasifikačního modelu (Hawkins, 2004). Míra objektivity či specificity modelu se dá redukovat dle prahů měr zajímavostí či omezením počtů pravidel vracených na výstupu dolovacího procesu.

6. Analýza podobných nástrojů fungujících jako SaaS

Nástroj EasyMiner funguje jako webová služba, která dovoluje uživateli nahrávat data na serverovou část, zpracovávat je a hledat v nich zajímavá pravidla. Ze seznamu pravidel lze poté vytvářet modely pro klasifikační úlohy, byznys pravidla či odhalování anomálií. Všechny tyto operace lze provádět přes grafické uživatelské rozhraní přístupné skrze internetový prohlížeč. Na straně serveru jsou data zpracovávána komplexními metodami, které jsou rovněž přizpůsobeny k distribuovaným výpočtům pro paralelní analýzu velkých dat. Samotné uživatelské rozhraní je ovšem velmi jednoduché a intuitivní i pro uživatele, kteří se přímo nezajímají o technickou stránku oblasti strojového učení, ale i přesto chtějí využít takovýchto nástrojů pro získání konkurenčních výhod v rámci svých podnikatelských plánů. Veškeré operace, které lze v systému EasyMiner provádět nejsou dostupné pouze skrze grafické rozhraní, ale i přes RESTová API pro pokročilejší uživatele, kteří si mohou funkce nástroje EasyMiner integrovat do vlastního systému.

EasyMiner je poskládán z několika mikroslužeb, které jsou předmětem akademického vývoje na půdě Vysoké školy ekonomické v Praze a Českého vysokého učení technického v Praze. Jedná se tedy o akademický projekt, který je v nynější době využíván převážně pro akademické účely. Jakékoliv komerční využití a poskytování nástroje v podnikové oblasti je aktuálně analyzováno. S touto analýzou je spojeno i srovnání se současnými jinými nástroji, které jsou schopny poskytovat podobné služby z oblasti strojového učení.

Sada takovýchto služeb spadá do kategorie MLaaS neboli Machine Learning as a Service (Ribeiro, et al., 2015). Jedná se tedy o typy softwaru založených na modelu

cloud computing, které jsou poskytovány jako služba, jejichž typickými příznaky jsou (Bruckner, et al., 2012) (Voříšek & Basl, 2008):

- Dostupnost přes internet
- Pay as you go - uživatel zaplatí pouze za to, co ve skutečnosti využije
- Nasazení do výpočetního cloudu, který se vyznačuje vysokou mírou škálovatelnosti a elasticity. Aplikace by tedy neměla být omezena výší aktuálně přihlášených uživatelů a výpočty by měly probíhat paralelně na více strojích.
- Aktuálnost využívaných služeb
- Schopnost integrovat službu do jiných systémů skrze API
- Spolehlivost, dostupnost a odolnost proti výpadkům
- Bezpečnost nahraných dat

Nástroj EasyMiner má za cíl být rovněž službou typu MLaaS, v níž platí jasná politika a obsahovala by výše zmíněné rysy typické pro aplikace typu SaaS (Baldominos, et al., 2014). Nynější verze[6] nemá zatím nastavena žádná pravidla pro komerční využití a nelze zatím říci, vzhledem k jejímu současnému vývojovému stavu, že by vykazovala rysy veliké spolehlivosti a dostupnosti. Nicméně aktuální verze je přizpůsobena k nasazení do výpočetního cloudu a je schopna obsloužit více uživatelů najednou a to i pro velká data s využitím distribuovaného systému Apache Hadoop (nyní zcela bez omezení pro akademické účely). V následujících odstavcích jsou zmíněny jiné služby (Butler, 2016), které dokáží řešit podobné byznys problémy právě jako nástroj EasyMiner. Na základě popisu ostatních nástrojů lze poté snáze určit, jaké jsou silné a slabé stránky jednotlivých řešení a zdali může systém EasyMiner v některých ohledech konkurovat ostatním komerčním službám.

6.1 Desktopové nástroje

V první řadě je dobré uvést standardní desktopové nástroje, které dokáží výše nadefinované byznys problémy řešit (viz kapitola 4), a které si uživatelé mohou nainstalovat a používat bez omezení na svém počítači. Mezi známé aplikace, které implementují řadu algoritmů strojového učení včetně klasifikátorů a hledání asociačních pravidel jsou:

- RapidMiner
- Weka
- IBM SPSS Modeler
- SAS Enterprise Miner

Tyto nástroje se hodí pro analýzu malých a středně velkých dat a výkon výpočtů závisí na hardwarovém vybavení stroje uživatele. Zpracování dat tedy nelze pro základní verze těchto desktopový nástrojů distribuovat na více výpočetních uzlech, tudíž je i obtížné analyzovat velká data. Velkou nevýhodou desktopových řešení je ovšem nemožnost integrace nabízených funkcí s jiným systémem, např. pomocí webového API.

[6] *EasyMiner v2.4 - listopad 2016*

Mezi výhodami použití webových služeb oproti dektopovým řešením patří:

- Přístup ke všem operacím přes webové API
- Grafické uživatelské rozhraní dostupné online přes internetový prohlížeč
- Není nutné aplikaci instalovat
- Podpora velkých dat a škálování složitých výpočtů

Standardní dektopové nástroje strojového učení jsou určeny spíše pro datové experty, se kterými lze modelovat různé procesy dolování dat od nahrávání, předzpracování, učení až po samotnou evaluaci a práci s výsledným modelem. K dispozici je obvykle velká škála naimplementovaných algoritmů (klasifikace, regrese, shlukování, detekce anomálií apod.) a sada pokročilých vizualizačních nástrojů.

6.2 BigML

Webová služba BigML.com patří v dnešní době mezi špičku v oblasti strojového učení a analýzy dat ve světě. Poskytují komplexní webové služby pro řešení klasifikačních úloh, hledání pravidel, detekci anomálií či shlukování. Rovněž služba nabízí pokročilé metody evaluace, predikce a vizualizace. Aktuálně poskytuje své služby zdarma pro úlohy do 16MB. Pro větší datasety již cena roste s velikostí a počtem paralelních úloh.

Infrastruktura systému BigML využívá cloudových služeb společnosti Amazon; přirozeně tedy dokáže škálovat výpočty a obsluhovat tisíce uživatelů najednou díky distribuované hardwarové infrastruktuře. Služba poskytuje uživateli velmi přívětivé grafické rozhraní, které je dostupné skrze webový prohlížeč. Dále je možnost integrovat operace do vlastního řešení díky existujícímu RESTovému API. Nástroj rovněž dokáže zpracovat datasety v řádech několika TB kde jsou úlohy vykonávány v proudech a v reálném čase.

Aplikace BigML nabízí řešení pro hledání pravidel, avšak neposkytuje takové možnosti jako nástroj EasyMiner. Není zde možnost definování vzoru pravidla, tvorby klasifikačního modelu ze seznamů vrácených pravidel či sestavení byznys pravidel. Co se týče distribuovaných výpočtů, tak EasyMiner běží na známém systému Apache Hadoop a úlohy pro velká data jsou oproti BigML spouštěny dávkově. Výhodou použití Hadoop řešení je snadná rozšiřitelnost a přenositelnost, známá technologie a schopnost analyzovat velká data. Nevýhodou je samozřejmě rychlost, náročnost výpočtu a vytížení celé infrastruktury. Porovnáme-li přístup pro tvorbu klasifikačních modelů nástroje EasyMiner a BigML, nejsou rozdíly příliš markantní. BigML používá rozhodovací stromy, které se dají převést na pravidla, avšak jejich model je mnohem konzistentnější. Naopak EasyMiner vrací pouze množinu pravidel, se kterou se dá lehce manipulovat a lze ji upravovat dle uživatelských požadavků.

Ve zkratce se dá říci, že aplikace EasyMiner poskytuje více možností v oblasti pravidlového přístupu k analýze dat. BigML implementuje pouze základní přístupy pro hledání pravidel a jejich orientace je zaměřena více na tvorbu rozhodovacích stromů v případě prediktivní analýzy. Naopak BigML je silný ve všech ostatních oblastech.

6.3 Google Prediction API

Prediction API[7] je webová služba od Googlu, která poskytuje nástroj pro tvorbu klasifikačních modelů pouze skrze webové API. Pro tuto službu neexistuje žádné oficiální a robustní grafické rozhraní. Soubor k analýze je potřeba nahrát skrze službu *Google Cloud Storage*, ze kterého se poté natrénuje klasifikační model. Velikost datasetu je omezena na 2,5GB, tudíž služba není zaměřena na zpracování velkých dat. Velikou nevýhodou je, že natrénovaný model nelze nijakým způsobem stáhnout či ho analyzovat; predikce se provádí rovněž pouze skrze API. Celý klasifikátor tedy působí jako černá skřínka a nelze nijakým způsobem zkoumat chování daného modelu. Naopak velikou výhodou je schopnost proudového (inkrementálního) učení v reálném čase, což dovoluje měnit uživateli model na základě změn v datech bez nutnosti opakování kompletní trénovací fáze. Služba neposkytuje žádné nástroje pro hledání pravidel či detekci anomálií.

Novým dosud experimentálním produktem od Google je služba *Google Cloud Machine Learning*, která poskytuje API pro tvorbu klasifikačních modelů pomocí frameworku TensorFlow. Tato knihovna nabízí nástroje pro hluboké učení neboli deep learning a slouží hlavně k učení kvalitnějších modelů dle neuronových sítí (Schmidhuber, 2015).

6.4 Amazon Machine Learning

Amazon Machine Learning[8] poskytuje svojí infrastrukturu k tvorbě klasifikačních či regresních modelů. Stejně jako služba od Googlu, i tato předpokládá, že jsou data uložena v cloudu společnosti, která službu poskytuje; v tomto případě je to např. *Amazon S3*, *Amazon Redshift* aj. úložní služby od Amazonu. Komplexní projekt Amazon Machine Learning nabízí webové rozhraní pro grafickou analýzu vstupních dat, nástroje pro transformaci či předzpracování dat a trénovací algoritmy pro tvorbu klasifikačních modelů, které lze použít buď k dávkové, nebo k proudové predikci. Po nahrání vstupních dat se tedy veškeré operace dají dělat pouze na infrastruktuře společnosti Amazon a vlastnosti služby jsou podobné jako v případě nástroje od Googlu. Velikost vstupních datasetů je omezena na 100GB a trénovací fáze může trvat maximálně 7 dní. Služba neposkytuje žádné funkce pro dolování asociačních pravidel a práci s nimi.

6.5 Microsoft Azure Machine Learning

Tato služba nabízí řadu nástrojů a algoritmů z oblasti strojového učení, které lze spouštět přímo na cloudu Microsoft Azure[9]. Jednotlivé úlohy se definují pomocí pipelines, v nichž lze provádět transformace, klasifikace, regrese, shlukování, detekci anomálií aj. operace. Vybírat může uživatel z velké množiny algoritmů, která obsahuje např. rozhodovací stromy, neuronové sítě, pravděpodobnostní modely apod. Pro pokročilé uživatele je nabízena možnost spouštět moduly pomocí skriptovacích jazyků R a Python, a to přímo v cloudu. Služba je tedy určena spíše pro datové experty, kteří již mají základní znalosti o poskytovaných algoritmech a dokáží pracovat se statistickými a analytickými nástroji. K dispozici je velmi přívětivé grafické uživatelské

[7] *https://cloud.google.com/prediction/*

[8] *https://aws.amazon.com/machine-learning/*

[9] *https://azure.microsoft.com/cs-cz/services/machine-learning/*

rozhraní, ve kterém si lze velmi jednoduše definovat komplexní proces pro zpracování dat.

Služba není koncipována pro zpracování velkých dat, ale spíše pro jednoduché použití a přístupnost nabízených algoritmů, avšak některé algoritmy jsou navrženy pro paralelní a rychlé zpracování. Rovněž je možné spustit paralelně několik úloh najednou.

6.6 Srovnání EasyMineru s jinými nástroji

Obecně se nedá říci, který z uvedených nástrojů je nejlepší; každý je silný v různých rovinách. Pro komplexní řešení různých úloh strojového učení jsou k dispozici robustní nástroje Microsoft Azure ML a BigML, které poskytují všechny základní techniky – klasifikace, regrese, shlukování, pravidla, detekce anomálií apod. Oba nástroje disponují velmi přívětivým grafickým rozhraním a BigML navíc možností vyexportovat natrénovaný model pro offline použití. Pro inkrementální trénování nebo využití moderních neuronových sítí (deep learning) je vhodné použít služeb nabízené Googlem. Pro rychlou online klasifikaci a predikci s grafickým rozhraním lze použít Amazon ML.

Nástroj EasyMiner je silný v oblasti dolování asociačních pravidel, ve které nabízí velkou škálu možností. V porovnání s konkurencí nemůže dosahovat takových kvalit, jelikož se jedná o akademický neziskový projekt, avšak společně s BigML jsou jedinými službami (z výše uvedených), které disponují algoritmy pro práci s pravidly jako SaaS. EasyMiner oproti BigML rozšiřuje práci s pravidly o možnosti dolování dle vzorů, tvorby pravidlových modelů a byznys pravidel, hodnocení a manipulaci s pravidly.

7. Závěr

Práce shrnuje základní informace o dolovacím nástroji EasyMiner, který je zaměřen na hledání asociačních pravidel a práci s nimi. Nástroj funguje jako webová služba typu MLaaS, jelikož nabízí komplexní prostředky pro získávání znalostí z databází a řešení základních úloh strojového učení (dle metodiky CRISP-DM) a to skrze webové rozhraní (GUI + API). Aplikace je rovněž schopna zpracovat velká data v prostředí Apache Hadoop díky využití výpočetního clusteru MetaCentra sdružení CESNET a paralelně obsluhovat několik uživatelů najednou.

V práci byly definovány byznys problémy, pro které je vhodné systém EasyMiner použít. Dále bylo ukázáno srovnání použití pravidlového přístupu pro různé úlohy z oblasti strojového učení oproti jiným řešením. Vzrůstající poptávka po provádění data miningu ve firmách zvyšuje i nabídku webových služeb poskytující nástroje pro získávání znalostí či prediktivní a deskriptivní analýzu. I přes velkou sílu online nástrojů typu MLaaS, které nabízejí velikáni jako Google, Amazon či Microsoft, lze i pro menší projekt typu EasyMiner najít takové využití, které jiné služby nenabízejí nebo nabízejí jinou formou nevyhovující cílovému uživateli.

***Poděkování**: Tato práce vznikla za podpory Vysoké školy ekonomické v Praze pod grantem IGA 29/2016 a Fondu rozvoje CESNET pod grantem č. 540/2014.*

Reference

Aggarwal, C. C. & Han, J., 2014: *Frequent pattern mining.* Springer International Publishing: Switzerland

Agrawal, R., Imieliński, T. & Swami, A., 1993: *Mining association rules between sets of items in large databases.* New York, NY, USA, ACM, pp. 207-216

Agrawal, R., Srikant, R. & others, 1994: *Fast algorithms for mining association rules.* San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 487-499

Baldominos, A., Albacete, E., Saez, Y. & Isasi, P., 2014: *A scalable machine learning online service for big data real-time analysis.* San Jose, California, IEEE, pp. 1-8

Bing, L., Wynne, H. & Yiming, M., 1998: *Integrating classification and association rule mining.* New York, NY, AAAI Press

Bruckner, T. a další, 2012: *Tvorba informačních systémů.* Praha: Grada Publishing

Butler, M., 2016: *10+ Machine Learning as a Service Platforms.* [Online] Available at: http://www.butleranalytics.com/10-machine-learning-as-a-service-platforms/ [Přístup získán 4 Červen 2016]

Davis, J., 2016: *Gartner's BI Reboot, Everybody Loves Spark: Big Data Roundup.* [Online] Available at: http://web.archive.org/web/20160426224538/http://www.informationweek.com/big-data/big-data-analytics/gartners-bi-reboot-everybody-loves-spark-big-data-roundup/d/d-id/1324309 [Přístup získán 26 Duben 2016]

Dorard, L., 2015: *Machine Learning Wars: Amazon vs Google vs BigML vs PredicSis.* [Online] Available at: http://web.archive.org/web/20160820153108/http://www.kdnuggets.com/2015/05/machine-learning-wars-amazon-google-bigml-predicsis.html [Přístup získán 20 Srpen 2016]

Fotr, J. a další, 2012. Tvorba strategie a strategické plánováni. *Praha: Grada Publishing.*

Freitas, A. A., 2000: Understanding the crucial differences between classification and discovery of association rules: a position paper. *AcM sIGKDD Explorations Newsletter,* Svazek 2, pp. 65-69.

Friedman, J., Hastie, T. & Tibshirani, R., 2001: *The elements of statistical learning.* Berlin: Springer series in statistics Springer

Fürnkranz, J., 2011: Decision Lists and Decision Trees. V: *Encyclopedia of Machine Learning.* Incorporated: Springer, pp. 261-262

Fürnkranz, J., Gamberger, D. & Lavrač, N., 2012: *Foundations of rule learning.* Incorporated: Springer Science & Business Media

Fürnkranz, J. & Kliegr, T., 2015: *A brief overview of rule learning.* Berlin, Springer International Publishing, pp. 54-69

Harper, J., 201:. *Improving Big Data Analytics with Machine Learning-as-a-Service.* [Online] Available at: https://web.archive.org/web/20160617030331/http://www.dataversity.net/improving-big-data-analytics-with-machine-learning-as-a-service [Přístup získán 17 Červen 2016].

Hashem, I. A. T. a další, 2015: The rise of “big data” on cloud computing: Review and open research issues. *Information Systems,* Svazek 47, pp. 98-115

Hawkins, D. M., 2004: The problem of overfitting. *Journal of chemical information and computer sciences*, Issue 1, pp. 1-12

He, Z., Xu, X., Huang, J. Z. & Deng, S., 2005: FP-outlier: Frequent pattern based outlier detection. *Comput. Sci. Inf. Syst.,* Svazek 2, pp. 103-118

Chorowski, J., 2012: *Learning understandable classifier models*. Doctoral thesis. University of Louisville, Kentuck

Kliegr, T. & Kuchař, J., 2015: *Benchmark of Rule-Based Classifiers in the News Recommendation Task.* Toulouse, France, Springer International Publishing, pp. 130-141.

Kliegr, T., Kuchař, J., Sottara, D. & Vojíř, S., 2014: *Learning business rules with association rule classifiers.* Prague, Czech Republic, Springer International Publishing, pp. 236-250.

Li, H. a další, 2008: PFP: parallel fp-growth for query recommendation*. ACM*, pp. 107-114.

Liu, B., Ma, Y. & Wong, C.-K., 2001: Classification using association rules: weaknesses and enhancements. V: *Data mining for scientific and engineering applications.* US: Springer, pp. 591-605.

Novotný, O. & Průcha, M., 2013: *Automatizovaná extrakce business pravidel se zpětnou vazbou,* Praha: TA ČR

Peddibhotla, G. B., 2015. *Gartner 2015 Hype Cycle: Big Data is Out, Machine Learning is in.* [Online] Available at: http://web.archive.org/web/20160513081309/http://www.kdnuggets.com/2015/08/gartner-2015-hype-cycle-big-data-is-out-machine-learning-is-in.html [Přístup získán 13 Květen 2016]

Piatetsky, G., 2015: *Data Scientists Automated and Unemployed by 2025?.* [Online] Available at: http://web.archive.org/web/20160804113335/http://www.kdnuggets.com/2015/05/data-scientists-automated-2025.html [Přístup získán 4 Srpen 2016].

Ribeiro, M., Grolinger, K. & Capretz, M. A. M., 2015: MLaaS: Machine Learning as a Service. *IEEE 14th International Conference on Machine Learning and Applications,* Miami, USA, pp. 896-902

Schmidhuber, J., 2015: Deep learning in neural networks: An overview. *Neural Networks*, Issue 61, pp. 85-117.

STAMFORD, 2015: *Gartner Survey Shows More Than 75 Percent of Companies Are Investing or Planning to Invest in Big Data in the Next Two Years.* [Online] Available at: https://web.archive.org/web/20160404180842/http://www.gartner.com/newsroom/id/3130817 [Přístup získán 4 Duben 2016].

Tan, P.-N. & others, 2006: *Introduction to data mining.* Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc..

Vojíř, S., 201:. *Učení business rules z výsledků dolování GUHA asociačních pravidel,* Vysoká škola ekonomická v Praze

Vojíř, S., Duben, P. V. & Kliegr, T., 2014: Business rule learning with interactive selection of association rules. *RuleML Challenge,* Svazek 2014

Vojíř, S. a další, 2013: Transforming association rules to business rules: EasyMiner meets Drools. *RuleML (2),* Svazek 1004

Voříšek, J. & Basl, J., 2008: *Principy a modely řízení podnikové informatiky.* Praha: Oeconomica

Wagner, J., 2014: *Machine Learning and Predictive Analytics Foster Growth.* [Online] Available at: http://web.archive.org/web/20160314091443/http://www.programmableweb.com/news/machine-learning-and-predictive-analytics-foster-growth/2014/02/21

[Přístup získán 14 Březen 2016].

Wang, J., Han, J., Lu, Y. & Tzvetkov, P., 2005: TFP: An efficient algorithm for mining top-k frequent closed itemsets. *IEEE Transactions on Knowledge and Data Engineering*, 17(5), pp. 652-663.

White, T., 2012: *Hadoop: The definitive guide.* Book: O'Reilly Media, Inc..

Wirth, R. & Hipp, J., 2000: *CRISP-DM: Towards a standard process model for data mining.* Citeseer, pp. 29-39

Zaharia, M. a další, 2012: Fast and interactive analytics over Hadoop data with Spark. *USENIX Login,* Svazek 37, pp. 45-51

Zikopoulos, P., Eaton, C. & others, 2011: *Understanding big data: Analytics for enterprise class hadoop and streaming data.* IBM: McGraw-Hill Osborne Media.

JEL Classification: C88, D83