Acta Informatica Pragensia, 2018, 7(1), 58–73
DOI: 10.18267/j.aip.114
Original Article

Dolování z otevřených dat o rozpočtech a výdajích

Data Mining from Open Fiscal Data

David Chudán [1], **Vojtěch Svátek** [1], **Jaroslav Kuchař** [1,2], **Stanislav Vojíř** [1]

Abstrakt

Metody dolování z dat jsou aplikovány ve stále větší míře, a to i v doménách, které tradičně nemají tak silnou podporu analytických nástrojů a kde převládá ruční práce analytika. Použití těchto metod v oblasti fiskálních dat umožní jejich hlubší analýzu a může přinést nová zjištění. Nasazení pokročilých metod dolování z dat je jednou z částí projektu OpenBudgets.eu, který se zaměřuje na transparentnost a odpovědnost v oblasti nakládání s veřejnými prostředky. Tento přehledový článek shrnuje některé zkušenosti autorů z tohoto projektu získané při vývoji, implementaci a aplikaci vybraných metod dolování z fiskálních dat. Jedná se zejména o metody detekce anomálií a dolování asociačních pravidel. Tyto metody jsou integrovány do centrální platformy projektu, která je k dispozici pokročilým i běžným uživatelům v případě zájmu o analýzu fiskálních dat. Pilotní analýzy ukázaly, že problémem dataminingové analýzy v této doméně je velký objem nacházených pravidel a různorodý původ jejich vzniku.

Klíčová slova: Dolování z dat, otevřená data, fiskální data, evropský projekt, OpenBudgets.eu.

Abstract

Data mining methods are still more popular, even in domains where there is traditionally limited support by analytical tools and where the analyst´s manual work still prevails. Using these methods in the fiscal domain enables deeper analysis and can bring new findings. The deployment of data mining methods is one part of the OpenBudgets.eu project, which focuses on transparency and accountability in the public funds management. This overview article summarizes selected experiences of the authors of the project from the development, implementation and application of selected data mining methods on mining fiscal data. These methods are integrated into the central platform of the project available for the advanced and common users interested in fiscal data analysis. The pilot analysis showed that the problem of data mining in this domain is the large amount of found rules together with its heterogenous origin.

Keywords: Data mining, Open data, Fiscal data, European project, OpenBudgets.eu.

[1] Department of Information and Knowledge Engineering, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
✉ david.chudan@vse.cz, svatek@vse.cz, stanislav.vojir@vse.cz

[2] Department of Software Engineering, Faculty of Information Technology, Czech Technical University in Prague, Thakurova 9, 160 00 Prague, Czech Republic
✉ jaroslav.kuchar@fit.cvut.cz

1 Úvod

Veřejné instituce na mnoha úrovních, od obcí až např. po Evropskou unii, pravidelně navrhují, předkládají ke schválení a následně čerpají své rozpočty. Jedná se v celkovém součtu o velké objemy dat, zpravidla v podobě vícerozměrných tabulek a na ně navázaných číselníků. Data jsou ve stále vyšší míře vystavována na webu v otevřených formátech, zpravidla jako soubory typu CSV (tj. s hodnotami oddělenými čárkou), případně ve formátu kalkulátoru Excel.

Fiskální, tj. rozpočtová a výdajová, data jsou předmětem zájmu mnoha zainteresovaných stran, od neziskových organizací a občanských sdružení bojujících za transparentnost a efektivní alokaci veřejných výdajů, přes novináře publikující korupční kauzy, až po oficiální kontrolní orgány. Ve většině případů jim však jako analytický nástroj slouží MS Excel (například podle Alloca (2016) byl MS Excel v roce 2016 v komerčních firmách používán pro finanční plánování a rozpočtování v devíti z deseti případů), který je sice velice univerzální, má však značná omezení v pokročilejších analýzách. Uživatel se musí spoléhat především na interaktivní vizualizace a ruční procházení dat. Přestože i takové techniky mohou vést k nalezení zajímavých jevů, mnoho jiných zjištění nutně uniká, protože zkoumaný mnoharozměrný datový prostor je pro ruční analýzu neúnosně rozsáhlý. Nabízí se proto možnost využít techniky automatického dolování z dat (data mining), které byly pro takové úlohy mnohokrát úspěšně využity v jiných oblastech lidské činnosti (např. v medicíně, bankovnictví nebo telekomunikacích).

Jedním z úkolů evropského projektu OpenBudgets.eu (končícího v říjnu 2017) bylo právě prozkoumání, nakolik mohou pokročilé techniky dolování z dat odhalit pravidelnosti nebo naopak mimořádnosti ve struktuře rozpočtů a výdajů veřejných institucí, a tím nepřímo přispět k fiskální efektivitě a transparentnosti. Hlavním cílem tohoto článku je shrnout zkušenosti autorů z aktivit v rámci tohoto projektu. Jednalo se o úpravy stávajících dataminingových nástrojů pro potřeby zpracování fiskálních dat, vývoj nového a inovativního algoritmu (pro detekci anomálií) odpovídajícího požadavkům projektu, a následné provádění samotných analýz. Vzhledem k aktuálnosti tématu předpokládáme, že se v příštích letech bude problematice analýzy fiskálních dat pomocí dataminingových metod věnovat více subjektů v akademické, veřejnosprávní i komerční sféře, a předložené výsledky (spolu s odkazy na další literaturu, zejména na oficiální výstupy projektu OpenBudgets.eu) jim mohou významně usnadnit orientaci. Popisované zkušenosti jsou v článku uváděny nově; předchozí práce Kuchaře a Svátka (2017) a Vojíře et al. (2017a) se věnovaly pouze metodě detekce anomálií, a i pro tu uvedly jen minimální popis vzorového datasetu a analytické úlohy, bez rozboru požadavků potenciálních uživatelů, popisu integrující platformy, interakce s doménovými experty a rozboru obecných zjištění, které jsou hlavním předmětem tohoto článku. Vedlejším cílem článku je pak seznámení odborné veřejnosti s existencí projektu OpenBudgets.eu (představujícího jednu z hlavních celosvětových iniciativ v této oblasti) jako takového, a zejména jeho prakticky využitelných výstupů i nad rámec aktivit spadajících pod oblasti dataminingu. Článek nejprve v kapitole 2 celý projekt stručně představuje, s důrazem na jeho pracovní balíček č. 2, jehož obsahem je identifikace, implementace a v případě potřeby také vývoj metod dolování z dat vhodných pro analýzu fiskálních dat. V kapitole 3 jsou podrobněji představeny ty metody dolování z dat, které rozvíjel a aplikoval tým VŠE: jedná se o dolování asociačních pravidel (včetně prostředku pro kombinování asociačních pravidel s interaktivní vícerozměrnou analytikou typu OLAP) a detekci anomálií. Kapitola 4 obsahuje ukázku výstupů těchto metod. Kapitola 5 diskutuje dosažené výsledky i související teoretické úvahy. Konečně, Kapitola 6 porovnává realizovaný výzkum s obdobnými projekty ve světě, a kapitola 7 shrnuje jeho přínosy a omezení.

2 O projektu OpenBudgets.eu

OpenBudgets.eu byl inovační projekt rámcového programu EU Horizont 2020, který se zaměřoval na podporu transparentnosti a odpovědnosti v oblasti nakládání s veřejnými prostředky. Jeho cílem bylo poskytnout pracovníkům veřejné správy, novinářům, občanským aktivistům a nevládním organizacím softwarové nástroje, data a analytické zprávy („příběhy o datech“), které mohou pomoci v úsilí o dosažení transparentního fungování veřejných agend. Trvání projektu bylo od května 2015 do října 2017.

Projekt byl rozdělen do celkem sedmi pracovních balíčků (workpackages), které se skládají z odevzdaných výstupů (deliverables). Jejich přehled lze najít na stránce http://openbudgets.eu/about/deliverables/.

Projekt byl realizován mezinárodním konsorciem devíti partnerů a koordinován německým institutem Fraunhofer IAIS. Tým VŠE vedl pracovní balíček č. 1, v rámci kterého byly vyvinuty datové struktury pro publikování fiskálních dat na webu. Výsledky této práce byly v r. 2016 publikovány na workshopu o propojených statistických datech (Mynarz et al., 2016). Vedle toho se ovšem tým VŠE výrazným způsobem podílel i na pracovním balíčku č. 2. Předmětem tohoto pracovního balíčku, na kterém kromě týmu VŠE pracovali tři zahraniční partneři, bylo (v oficiálním členění balíčku na pracovní úkoly – „tasks“):

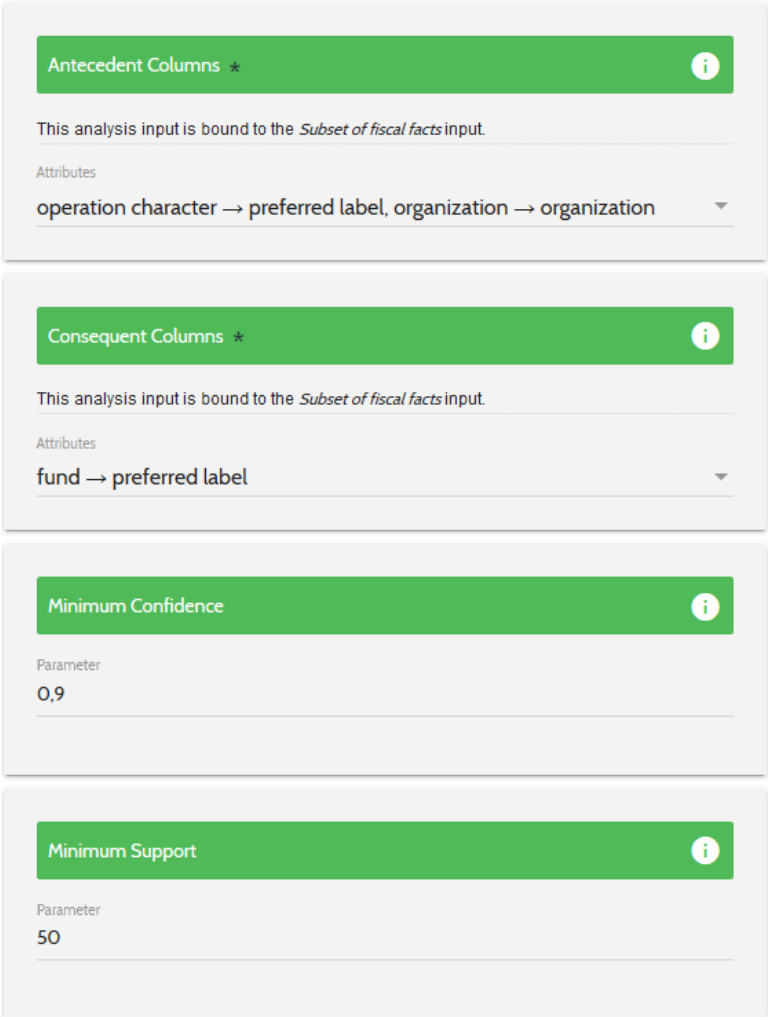
1. Umožnit sběr co největšího množství dat v různých formátech, které používají relevantní subjekty v procesu práce s rozpočtovými daty, a poloautomaticky je převést do formátu RDF (Cyganiak, 2014), ve kterém mohou být data sémanticky popsána a propojena.
2. Vytvořit prostředí pro vylepšování kvality dat již v RDF a jejich obohacování o další relevantní data z veřejně dostupných zdrojů, tak, aby výsledná data byla vhodná pro různé dataminingové techniky.
3. Vytipování vhodných technik a algoritmů pro analýzu agregovaných fiskálních dat, úpravy stávajících algoritmů pro tyto účely, a jejich implementace.
4. Integrace vyvinutých nástrojů do platformy projektu.

Částem 1 a 2, ve kterých zásadní roli sehrál softwarový rámec LinkedPipesETL dlouhodobě vyvíjený na půdě MFF UK Praha,[1] byly věnovány Deliverable 2.1 (Engels et al., 2016a) a Deliverable 2.2 (Klímek et al., 2016).

Části 3 a 4 se opíraly o široké spektrum dataminingových nástrojů, ať už vyvinutých přímo projektovými partnery nebo převzatými od třetích stran. V rámci těchto částí byly postupně vyprodukovány tři shrnující výstupy. V rámci Deliverable 2.3 (Engels et al., 2016b) byla provedena analýza požadavků pro práci s rozpočtovými a fiskálními daty. Výstupem této analýzy bylo identifikování metod dolování z dat, které jsou pro tyto účely vhodné. Z analýzy vyplynulo, že se jedná o následující metody: popisná statistika, shlukování, detekce anomálií, dolování asociačních pravidel a analýza časových řad. V rámci Deliverable 2.4 (Dong et al. 2017) byly tyto metody popsány, dále byly v případě potřeby upraveny pro co nejlepší možnost analýzy dat z fiskální domény, a rovněž je zde popsán nově vyvinutý algoritmus pro detekci anomálií (viz Kapitola 3.2). Poslední částí tohoto balíčku práce je Deliverable 2.5 (Kuchař et al., 2017) zabývající se integrací těchto metod do platformy projektu. Metody jsou prostřednictvím platformy dostupné a je možné je využít pro analýzu jak dat dostupných v platformě, tak vlastních dat, která se do platformy nahrají.

[1] https://linkedpipes.com/

Základním, integrujícím výstupem projektu je platforma dostupná na webové stránce http://openbudgets.eu. Jejím prostřednictvím je možné zpracovávat, zkoumat a analyzovat fiskální data pomocí různorodých nástrojů, zahrnujících jak jednoduché vizualizace, tak i pokročilé analytické nástroje založené na dolování z dat. Dále je na platformě možné najít rozpočtové hry (Journalism++, 2017) a srozumitelný výklad rozpočtování nebo korupčních praktik. Platforma by měla odrážet širší potřeby veřejnosti projevující zájem o porozumění veřejným rozpočtům na místní, národní i evropské úrovni. Obrázek 1 přináší ukázku uživatelského rozhraní, konkrétně menu pro nastavení úlohy dolování asociačních pravidel. Skládá se z částí Antecedent Columns, kam se vkládají atributy, které jsou součástí levé strany asociačního pravidla (antecedent, předpoklad pravidla), z Consequent Columns, kam se vkládají atributy, které jsou součásti pravé strany asociačního pravidla (konsekventu, závěru pravidla), dále dva povinné parametry pro prahové hodnoty spolehlivosti (confidence) vyjadřující podmíněnou pravděpodobnost závěru pokud platí předpoklad a podpory (support), vyjadřující minimální počet objektů splňujících předpoklad i závěr.



Obr. 1. *Nastavení dolování asociačních pravidel v platformě OpenBudgets.eu. Zdroj: Autoři.*

Platforma byla podrobena podrobnému testování pomocí tzv. large-scale trials. Do tohoto testování se zapojili tři představitelé veřejného sektoru, konkrétně zástupci z měst Bonn, Paříž a Soluň. Výsledky tohoto testování jsou shrnuty v Deliverable 7.7 (Orlandi et al., 2017). Z dotazníku, který byl zástupcům předložen, vyplývá, že platformu hodnotí z hlediska použitelnosti a uživatelského rozhraní pozitivně. Pro některé její technicky náročnější součásti,

dolování z dat nevyjímaje, předpokládají nutnost nějaké formy technické podpory. Článek se bude dále podrobněji věnovat technikám dolování asociačních pravidel a detekce anomálií.

3 Aplikované metody dolování z dat a jejich vyhodnocování

V úvodní fázi projektu byla konzultacemi s potenciálními uživateli (zejména projektovými partnery z praxe a členy industriálního poradního grémia) identifikováno široké spektrum požadavků na výstupy z datových analýz fiskálních dat. Mezi více než 40 individuálními požadavky (viz Engels et al., 2016b) figurovaly zejména:

- Analýzy rozpočtových trendů v čase a v závislosti na tematických oblastech
- Automatická identifikace dobrých a špatných příkladů sestavení jednotlivých rozpočtů
- Identifikace anomálně sestavených rozpočtů (pro další, ruční analýzu)
- Identifikace systematických problémů při tvorbě rozpočtů, plynoucích z chybně nastavených obecných pravidel
- Odhalování „černých děr“ v rozpočtech, plynoucích z netransparentní klasifikace výdajů
- Výběr rozpočtů, které má smysl srovnávat (např. obce s porovnatelným počtem obyvatel a/nebo výší HDP)
- Komparativní analýza zvolené dvojice rozpočtů porovnatelných obcí (včetně mezinárodního porovnání)
- Sledování vývoje rozpočtu v průběhu jeho životního cyklu (od předběžného návrhu přes schválení, čerpání, až po certifikaci profinancovaných projektů)
- Porovnávání shody plánovaného rozpočtu a skutečných výdajů, v časovém vývoji (např. zda se přesnost odhadování rozpočtu s časem zlepšuje)
- Porovnání rozpočtů s auditorskými zprávami
- Sledování toku dotací z úrovně EU až na úroveň skutečných příjemců dotace
- Analýza sítí vztahů ekonomických subjektů a včetně fyzických osob ve vazbě na rozpočty (např. odhalování klientelismu)
- Analýza klíčových indikátorů výkonnosti veřejné správy.

Z pohledu dolování z dat lze úlohy mj. rozdělit dle vhodnosti pro tzv. *prediktivní* vs. *deskriptivní* analytiku. Zjednodušeně řečeno, prediktivní analytika na základě minulých případů automaticky předpovídá nebo doporučuje hodnoty určitého atributu pro případy nové, zatímco výstupem deskriptivní analytiky jsou vztahy a hypotézy, jejichž následné zužitkování je záležitostí doménových expertů. Předpokladem prediktivní analytiky je dostupnost minulých příkladů již zaklasifikovaných do odpovídajících kategorií. V tomto ohledu byly zkoumány zejména auditorské zprávy, které by hypoteticky mohly poskytnout příklady rozpočtů klasifikované jako „dobré“ a „špatné“. Ukázalo se však, že informace o kvalitě rozpočtu jsou ve zprávách popsány vágně a bez strojové čitelnosti. Vzhledem k nedostupnosti trénovacích příkladů pro klasifikátory se proto pozornost soustředila na deskriptivní analytiku (sem spadá z výše uvedených požadavků přibližně polovina: zejména identifikace anomálních rozpočtů, systematických problémů a černých děr, výběr porovnatelných rozpočtů, komparativní analýza, vývoj rozpočtu v průběhu životního cyklu). Její výstupy byly průběžně konzultovány s experty na rozpočtovou problematiku v ČR i v zahraničí; závěry z těchto konzultací jsou shrnuty v diskusi.

Metody dolování z dat se liší i podle struktury zpracovávaných dat, a to zejména na tzv. *propozicionální* metody, pracující nad jednou datovou tabulkou, a na metody *multirelační*, pracující nad více tabulkami, popřípadě nad volně strukturovanými soubory relačních faktů

různých logických predikátů. Potenciál pro multirelační analýzy, využívající nativní grafové struktury dat RDF, nabízí úloha analýzy sítí vztahů. Tato úloha sice byla týmem VŠE řešena,[2] ale k datu zpracování tohoto textu dosud nepřinesla použitelné výsledky s ohledem na nutnost zapojení rozsáhlého objemu apriorních znalostí („background knowledge“) potřebných pro zmenšení prostoru prohledávání, který je u multirelačního dolování extrémně velký. V článku se proto omezíme na metody dolování využívající jedinou datovou tabulku, což je dostačující pro naprostou většinu úloh ze seznamu požadavků. Předpokládáme přitom, že tabulka může být vytvořena z grafových dat RDF (již podle potřeby obohacených o externí propojitelná data) pomocí dotazu typu SELECT v dotazovacím jazyce SPARQL standardizovaném konsortiem W3C (Harris, 2013).

Reálně řešené úlohy dolování z dat byly rozděleny mezi zúčastněné partnery z hlediska jejich expertizy a dostupných nástrojů vyvíjených vlastními prostředky. Například řecký partner Open Knowledge Foundation Greece aplikoval nástroje pro shlukování (aplikovatelné mj. na úlohu „Výběr rozpočtů, které má smysl srovnávat…“) a analýzu časových řad (např. „Analýzy rozpočtových trendů v čase…“ nebo „Sledování vývoje rozpočtu v průběhu jeho životního cyklu“). Pro klíčovou úlohu „Identifikace anomálně sestavených rozpočtů…“ byly souběžně vyvíjeny dva komplementární algoritmy využívající odlišných principů, jak týmem VŠE, tak Univerzitou v Bonnu. Některé úlohy, které byly původně směřovány na automatické dolování z dat, byly dodatečně přeformulovány na úlohy řešitelné s pomocí statistických a vizualizačních nástrojů, které sice vyžadují větší objem ruční práce, ale jsou lépe využitelné přímo rozpočtovými experty, protože je jejich výstupům lépe rozumět. V následující části textu se proto zaměřujeme pouze na dvě úlohy řešené týmem VŠE – dolování asociačních pravidel a detekci anomálií). Jak již bylo zmíněno, jedná se o deskriptivní úlohy nad propozicionální reprezentací.

Výstupy analýz byly konzultovány s doménovými experty, kterými byli zástupci municipalit z ČR a Německa,[3] zpravidla podle následujícího scénáře:

1. Doménoví experti identifikovali dle vlastní úvahy „oblasti dat“, kde očekávali výskyt anomálií nebo zajímavých vztahů. Např. v případě municipality německého Bonnu byl očekáván výskyt nadstandardních rozpočtových položek určitých kategorií v souvislosti s 250tým výročím narození L. van Beethovena v r. 2020.
2. Dataminingový tým provedl analýzy dat bez informací o oblastech identifikovaných experty.
3. Experti prošli prvních několik desítek nalezených závěrů, a uvedli, které z nich spadají do oblasti očekávaného výskytu zajímavých vztahů.
4. Experti se dále vyjádřili k tomu, zda některé ze závěrů neodpovídajících oblasti očekávaného výskytu shledávají „zajímavými“, a proč.

3.1 Dolování asociačních pravidel

Asociační pravidla jsou vztahy v datech využívající obvyklou konstrukci „jestliže <předpoklad> potom <závěr>“ (anglicky označovaná jako pravidla „IF-THEN“). U takto

[2] Jedná se o jednu ze součástí doktorské disertační práce V. Zemana (pod vedením V. Svátka) obecně zaměřené na dolování z dat ve formátu RDF.

[3] Interakce autorů článku nad provedenými datovými analýzami probíhaly s českými experty přímo a s německými nepřímo (prostřednictvím německých partnerů). V projektu byla dále analyzována data od řeckých expertů, což ovšem zajišťovali řečtí partneři, s ohledem na nutnost jazykového porozumění dokumentaci datových sad.

vytvořených pravidel nás zejména zajímá, kolik příkladů (datových objektů) splňuje předpoklad pravidla, kolik jeho závěr, a kolik jak předpoklad, tak i závěr. Mezi nejznámější algoritmy pro hledání asociačních pravidel patří algoritmus *apriori* (Agrawal, Strikant, 1994), který byl navržen v souvislosti s tzv. analýzou nákupního košíku, tedy řešením úlohy „jaké položky zákazníci supermarketu obvykle nakupují společně“. V českém prostředí má dlouholetou tradici mírně odlišný přístup k asociačním pravidlům, založený na metodě GUHA. Více o tomto přístupu je možné nalézt například v Rauch a Šimůnek (2014) nebo v Hájek et al. (2010).

Na půdě VŠE byly v posledním desetiletí souběžně rozvíjeny dva komplexní softwarové nástroje pro dolování asociačních pravidel, LISp-Miner a EasyMiner. LISp-Miner, detailně popsaný v monografii Rauch a Šimůnek (2014), je rozsáhlá desktopová aplikace, vyvíjená již od roku 1996, která nabízí kromě dolování asociačních pravidel založených na metodě GUHA celou řadu dalších analytických procedur. EasyMiner, popsaný například v Kliegr et al. (2017), je webová aplikace pro dolování asociačních pravidel a klasifikaci. Oba nástroje se mimo jiné masivně používají při výuce dolování z dat na Fakultě informatiky a statistiky VŠE, a to jak na bakalářském stupni (EasyMiner), tak i na magisterském stupni (LISp-Miner). V rámci projektu OpenBudgets.eu bylo testováno použití obou nástrojů: EasyMineru pro dolování asociačních pravidel obecně a LISp-Mineru pro dolování asociačních pravidel v kombinaci s interaktivní analýzou OLAP, zastřešenou nově vyvinutým nástrojem OLAP Recommender.

OLAP Recommender je webový nástroj[4], který automatizuje proces explorativní analýzy multidimenzionálních dat. Integruje dvě analytické metody, vizualizaci multidimenzionálních dat a dolování GUHA asociačních pravidel, které je zprostředkované nástrojem LISp-Miner. Dolování multidimenzionálních dat je specifické v tom, že se jedná o data agregovaná (při dolování asociačních pravidel se obvykle pracuje s daty transakčními). Problematika dolování agregovaných dat a návrh nástroje OLAP Recommender jsou popsány v Chudán (2015). Systém byl implementován jako praktická část diplomové práce Bohuslava Koukala (Koukal, 2017b). Podrobněji se tomuto nástroji věnuje článek Koukal (2017a).

Nová verze systému EasyMiner, která byla modifikována v rámci řešení projektu a byla následně použita pro analýzu fiskálních dat v rámci projektu OpenBudgets.eu, podporuje kromě asociačních pravidel také detekci anomálií založenou na dolování častých vzorů (viz Kapitola 3.2). Všechny funkce tohoto systému jsou dostupné prostřednictvím REST API, díky kterému byla možná integrace systému do platformy projektu. Data mining prostřednictvím API detailně popsali Vojíř et al. (2017a, 2017b).

3.2 Detekce anomálií (outlier detection)

Detekce anomálií je úloha určení takových položek, událostí nebo pozorování, které neodpovídají očekávanému vzoru v datasetu. Algoritmů a technik pro detekci anomálií existuje celá řada. Například Chandola et al. (2009) tyto techniky a algoritmy rozděluje do několika skupin podle přístupu ke zpracování dat – klasifikační, založené na nejbližších sousedech, shlukovací, statistické, a založené na teorii informace. Další dělení vychází z toho, zda se jedná o učení s učitelem či bez učitele, např. (Aggarwal, 2013).

V rámci projektu OpenBudgets.eu byla týmem VŠE nejprve aplikována jednoduchá metoda vizuální detekce anomálií založená na *poměrových ukazatelích*, která spadá spíše do statisticko-vizualizačních než dataminingových metod. Následně však byla vyvinuta pokročilá metoda detekce anomálií pomocí učení bez učitele založená na tzv. *dolování častých vzorů*. Dalo by se

[4] Dostupný na adrese http://connect-dev.lmcloud.vse.cz/Recommender

říci, že tento přístup pracuje opačným způsobem než standardní algoritmy detekce anomálií, které určují odlehlé hodnoty. Myšlenka tohoto přístupu spočívá v tom, že pokud data obsahují více častých vzorů, je nepravděpodobné, že budou obsahovat anomálie. Přítomnost nebo nepřítomnost častého vzoru je využita pro výpočet celkového skóre anomálií (anomaly score) pro každou instanci dat. Podrobněji o této metodě viz Kuchař a Svátek (2017).

Metoda detekce anomálií pomocí častých vzorů byla implementována do systému EasyMiner v rámci balíčku v jazyce R, nazvaného *fmpoutliers* (Kuchař, 2017). Tento balíček obsahuje kromě výše zmíněné inovativní metody také implementaci šesti existujících algoritmů.

4 Výstupy dolování z dat

V rámci testování jednotlivých metod byly v průběhu roku 2017 použity dva datasety. Prvním byl přehled čerpání prostředků strukturálních fondů EU v rámci České republiky (ESF-CZ-2017-2013) z programového období 2007-2013 (Čerpání, 2017). Tento dataset obsahuje seznam příjemců podpory z fondů EU s uvedením názvů jejich projektů za celé programové období 2007-2013. Skládá se z 107 311 záznamů a 28 atributů, jako například: Název příjemce projektu, Částka hrazená z fondů EU, Datum alokace, Alokovaná částka, Datum průběžné platby, Celková částka proplacená od začátku projektu a Stav. Tento původní dataset byl následně obohacen o data z Administrativního registru ekonomických subjektů (ARES)[5], konkrétně byla přidána data o typu ekonomického subjektu a počtu zaměstnanců.

Druhým datasetem byl ESIF 2014[6], který obsahuje agregovaná data o čerpaných prostředcích z fondů EU pro jednotlivé země Eurozóny. Obsahuje například tyto atributy: Čerpající stát, Kategorie regionu, Celková poskytnutá částka z EU, Kofinancování z národních zdrojů a další.

4.1 Dolování asociačních pravidel

Dolování asociačních pravidel má určité požadavky na podobu dat. Především nepracuje se spojitými numerickými hodnotami. Ty je nutné diskretizovat, tedy převést na intervaly. Před analýzou datasetu Čerpání prostředků z programového období 2007-2013 byly v rámci přípravy dat provedeny následující úpravy:

- Odstraněny zbytečné textové řetězce, například odstraněno číslo operačního programu a ponechán pouze jeho název.
- Odvozen sloupec Rok ze sloupce Datum.
- Přidány poměrové ukazatele jako například poměr mezi alokovanou částkou a částkou skutečně vyplacenou nebo certifikovanou částkou a částkou skutečně vyplacenou.

Tyto poměrové ukazatele byly následně diskretizovány do ekvidistantních intervalů (tedy intervalů o stejné délce), např. [0,05 ; 0,1), [0,1 ; 0,15) ...

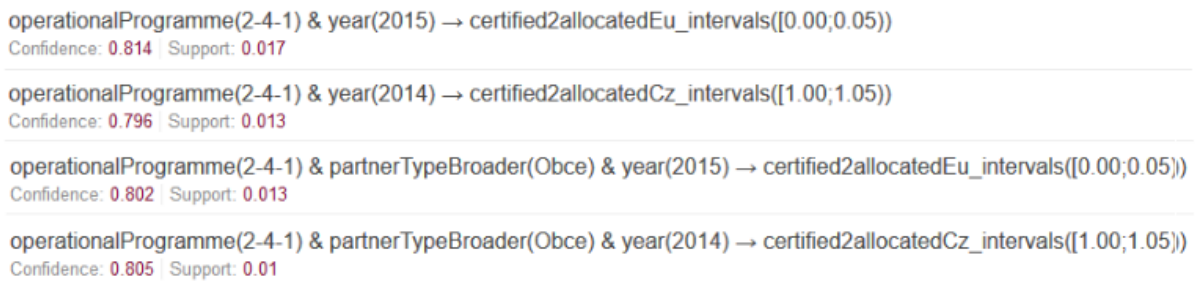
Pro konkretizování úlohy dolování asociačních pravidel byla stanovena následující analytická otázka: *Je možné predikovat poměr mezi alokovanými a vyplacenými penězi s využitím atributů operační program, rok a typ ekonomického subjektu?* Hodnota minimální podpory byla nastavena na 1% (relativní nastavení, 1% z celkového množství záznamů) a hodnota minimální spolehlivosti na 0,7.

[5] http://wwwinfo.mfcr.cz/ares/ares_es.html.cz

[6] https://raw.githubusercontent.com/openbudgets/datasets/master/ESIF/2014/raw/ESIF_FINANCE_DETAILS.csv

Dolování takto nastavené úlohy přineslo 98 asociačních pravidel, od poměrně jednoduchých, obsahující jeden atribut v předpokladu i závěru pravidla, po komplexnější, obsahující větší množství atributů. Obecným problémem při dolování asociačních pravidel je to, že z výsledné množiny získaných pravidel se mnoho z nich překrývá, respektive pravidla s větším množstvím atributů z podstaty výrokové logiky obsahují pravidla kratší. Příkladem může být pravidlo *w & x & y => z* a pravidla *w & x => z* a *x & y => z*. Výsledná množina pravidel potom často obsahuje pouze zlomek pravidel obsahující unikátní informaci o datech. Řešením tohoto problému může být nastavení vyšších prahových hodnot měr zajímavosti (v tom případě ale hrozí, že výsledkem budou pouze velmi obecná pravidla a konkrétnější, zajímavá pravidla do výsledků úlohy neprojdou), ruční procházení pravidel, kdy analytik vybere relevantní pravidla, nebo automatické metody následného zpracování, které mohou výslednou množinu pravidel automaticky redukovat, shlukovat či jinak upravovat (viz např. Kejkula, 2009).

Obrázek 2 ukazuje takový příklad pravidel z úlohy o čerpání prostředků strukturálních fondů EU. Je zřejmé, že první pravidlo je obsaženo ve třetím pravidlu a druhé pravidlo je obsaženo ve čtvrtém pravidlu.



Obr. 2. *Výsledky dolování pravidel z datasetu Čerpání strukturálních fondů EU. Zdroj: Autoři.*

Následují dva konkrétní příklady získaných pravidel:

V rámci programu 2-4-1 Životní prostředí je v roce 2015 pravděpodobnost 81,4 %, že se poměr certifikovaných ku alokovaných prostředkům bude pohybovat v intervalu [0 ; 0,05). Zde se jedná o jiný typ pravidla vyjadřující přímo sílu korelace mezi předpokladem a závěrem. Pravidlo lze číst jako upozornění, že pro program 2-4-1 nelze u většiny projektů ještě předpokládat proběhlou certifikaci uskutečněných výdajů. Na rozdíl od předchozího pravidla je ovšem nelze interpretovat jako odlišnost programu 2-4-1 od jiných programů, protože je možné, že pro jiné programy je míra certifikovanosti projektů rovněž nízká, což by se projevilo formou výskytu obdobných pravidel i pro jiné programy.

V rámci programu 7-1-1 Vzdělávání pro konkurenceschopnost je 3,521krát větší pravděpodobnost, že hodnota poměru certifikovaných prostředků k alokovaných bude v intervalu [0 ; 0,05), než v náhodně vybraném záznamu v datasetu. Vzhledem k tomu, že jde o interval nejnižších hodnot poměru, lze z pravidla usuzovat na možný problematický průběh řešení projektů programu 7-1-1, kvůli kterému se certifikace výrazně opožďuje; může se ale jednat např. i o větší délku projektu a tudíž pomalejší čerpání, nebo jiný faktor.

4.2 Detekce anomálií

Pro detekci anomálií s pomocí poměrových ukazatelů byl použit dataset ESIF 2014, z něhož byly použity atributy Čerpající stát, Fond, Kategorie regionu a Celková částka. Pro potřeby této metody byly dále všechny záznamy agregovány na úroveň jednotlivých zemí. Pro každou zemi byly předpočítány všechny kombinace hodnot na základě příslušného fondu a jeho kategorie. Tyto hodnoty byly použity pro výpočet poměrů.

Příklad:

Celkový součet dotací pro Rakousko je 10 655 136 237 EUR. V rámci fondu EAFRD (Evropský zemědělský fond pro rozvoj venkova) to pro Rakousko činí 7 699 887 667 EUR. Součet všech dotací pro Rakousko je 10 655 136 237 EUR. Poměrový ukazatel pro fond EAFRD tak činí 0,722646. Samotné poměry nejsou pro detekci anomálií vhodné. Pokud je ovšem porovnáme s průměrnou hodnotou, dostáváme daleko zajímavější údaje. Poměrový ukazatel pro fond EADRF pro všechny země Eurozóny kromě Rakouska činí 0,29. Odchylka od průměru pro Rakousko tedy činí přibližně 44 %. Když tuto hodnotu porovnáme, např. pomocí grafové vizualizace, s ostatními odchylkami, dojdeme k závěru, že je výrazně vyšší a je tedy možné tuto hodnotu považovat za anomálii.

Mezi další závěry získané touto metodou můžeme uvést například fakt, že Nizozemí má v rámci fondu EAFRD z hlediska poměrových ukazatelů velmi vysoké výdaje na podkategorii Odborné vzdělávání.

Pro testování metody založené na častých vzorech byl použit obohacený dataset ESF-CZ-2007-2013 (viz Kapitola 4), z něhož byly použity atributy Operační program, Typ ekonomického subjektu a Certifikované prostředky EU. U této metody bohužel není možnost jednoznačně interpretovat konkrétní výstup. Metoda se aplikovala na zmíněný dataset a byla porovnána anomálie postavená na extrémní odchylce hodnoty s anomálií založené na častých vzorech. V tomto případě metoda zohledňuje více deskriptorů instance, ne pouze extrémně vysokou finanční částku. Výstupem je potom instance s uvedeným celkovým „outlier score“ i s rozborem, jak jednotlivé deskriptory k celkovému skóre přispěly.

Anomaly instance (71553): partnerTypeBroader: 36, operationalProgrammeBroader: 107311, amount: 107311

Z tohoto příkladu je zřejmé, že na označení této konkrétní instance jako anomálie měly hlavní podíl atributy Operační program a Celková částka. Atribut Typ ekonomického subjektu měl roli daleko méně významnou. Více viz Kuchař a Svátek (2017).

5 Diskuze

Nástroje pro dolování z fiskálních dat byly experty, a to jak akademickými (vědci působící v oblasti veřejných financí), tak i z praxe (zástupci místní samosprávy z ČR a Německa) hodnoceny jako zajímavé, a některé z výstupů se shodovaly s jejich očekáváními. Současně však byla identifikována některá omezení stávajících dataminingových technologií, která jejich využitelnost pro oblast fiskálních dat (a zřejmě i v dalších oblastech, kde se předpokládá aplikace doménovými experty bez hlubších zkušeností s datovou analytikou) omezují.

V první řadě se, v případě deskriptivních dataminingových metod, jedná o typicky velký rozsah jejich výstupu. Zejména metody dolování asociačních pravidel produkují nad datasety běžné velikosti stovky, i tisíce pravidel, z nichž mnohá jsou si velmi podobná. Jejich ruční procházení je pro experty časově zatěžující, s nevýhodným poměrem zajímavých zjištění oproti nezajímavým. Určitým příslibem je v tomto ohledu důraz, který komunita dolování z dat v poslední době věnuje následnému zpracování pravidel – ta je možno shlukovat, vyhledávat podle zadaných podmínek, případně i filtrovat proti expertním apriorním znalostem („background knowledge“). Na tomto výzkumu se opět podílel i tým VŠE (Kliegr et al., 2010).

Faktorem, který není ani tak problémem datataminingových technologií, jako spíše komplexnějšího procesu zpracování heterogenních dat, je různorodý původ „pravidelností“ či naopak „anomálií“ ve výstupech procesu dolování. Vedle zjištění odrážejících způsoby čerpání rozpočtu z *věcného* hlediska se do výstupů „přimíchávají“ zejména:

- Jevy způsobené *terminologickou* nekompatilibitou datových zdrojů. Jedná se zejména o situace, kdy jsou věcně shodné výdaje v různých rozpočtech klasifikované různými hodnotami číselníků, které na sebe nejsou v potřebné míře správnosti a úplnosti namapované.
- Formální chyby v *původních zdrojích* publikovaných dat. Tradičním příkladem je nepřesné přiřazení IČO firmě, která je příjemcem rozpočtové dotace (např. záměna mateřské a dceřiné firmy).
- Artefakty *předzpracování dat*. Proces převodu dat do podoby propojitelných dat, jejich propojování na externí data, a finální převod do podoby optimalizované pro dolování z dat, zahrnuje řadu komplexních a často (zejména v případě poloautomatického propojování) dokonce heuristických kroků. „Anomálie“, kterou finálně „odhalí“ dataminingový algoritmus, pak nemusí vůbec odpovídat původní podobě dat. Zjištění sice může být užitečné, ale nikoliv pro původní vlastníky dat (či jiné věcně zainteresované), ale spíše pro tvůrce softwarových nástrojů a návrháře datových procesů.

V některých případech je problémem i skutečnost, že důležité údaje jsou sice v datech přítomné, ale pouze v *polostrukturované* podobě (např. v jednotlivých bodech projektové dokumentace, která může být ze strukturovaných rozpočtových dat nepřímo prolinkována). Pro obohacení dat o dostatečně přínosné atributy pak nepostačuje technologie datové integrace, a to ani ve své pokročilé variantě využívající sémanticky popsaná data v RDF. Jedinou možností je pak aplikace metod zpracování přirozeného jazyka, které ovšem podstatně hůře škálují na velké rozsahy dat, které jsou naopak očekávané pro tvorbu věrohodných výstupů v dataminingových přístupech. Faktorem, který může výzkum a praxi v této oblasti posunout vpřed, je tedy pokrok v populárních technologiích označovaných jako např. jako Big Data Analytics a dosud používaných především v korporátním prostředí. Zde je již kombinování strukturovaných a polo či nestrukturovaných dat intenzivně zkoumáno.

6 Porovnání s existujícím výzkumem

Aplikace metod dolování z dat na fiskální datové zdroje je novým tématem, kterému se prakticky žádný z předchozích projektů nevěnoval. Porovnání lze proto udělat spíše jen s nepřímo souvisejícími projekty, zaměřenými na dataminingovou analýzu odlišných typů dat veřejné správy, popřípadě na jednodušší typy analýz fiskálních nebo jiných otevřených dat.

Do první skupiny patří projekty aplikující dolování z dat na data o veřejných zakázkách. Např. v projektu 7. RP EU LOD2 byly (rovněž za účasti týmu VŠE) hledána asociační pravidla predikující počet uchazečů o veřejnou zakázku na základě jejích charakteristik, a také vhodný cenový interval pro nově připravovanou zakázku (Svátek et al., 2014).

Ve druhé skupině, zahrnující analytické aktivity s pomocí databázových statistických, a vizualizačních nástrojů, je projektů podstatně více. Za zmínku stojí především portál OpenSpending,[7] poskytující vizualizační a dotazové techniky nad fiskálními daty, založené primárně na dotazovacím jazyce SQL. Pokročilejší sémanticko-databázové prostředky (jazyk SPARQL) k procházení rozpočtových dat pro čtyři národní a tři „nižší“ rozpočty nabídl projekt popisovaný ve (Vafopoulos et al., 2013). Za další příklad lze uvést projekt TACOD (Tackling Corruption through Open Data, 2017), který využívá otevřených dat pro předcházení korupci.

[7] https://openspending.org/

Jeho výstupem je metodika, která kombinuje informace z legislativy, více než 100 případových studií a 3 800 dotazníků, a je používána pro detekci a prevenci korupce ve 4 zemích EU.

Pokud rešerši rozšíříme o standardní časopisecké a konferenční články týkající se data miningu v sektoru financí, je vhodné zmínit přehledové články o této problematice, shrnující vhodnost jednotlivých metod pro tuto doménu. Příkladem mohou být články Kovalerchuk a Vityaev (2005) nebo Zhang, Zhou (2004). Velmi častým tématem v rámci domény financí je detekce podvodů (fraud detection). I zde existuje množství článků, které shrnují best practices z této oblasti, například Yue et al. (2007), Ngaia et al. (2011), Sharma, Panigrahi (2013) nebo Sagar et al. (2016).

Pokud se zaměříme na dolování asociačních pravidel ve finanční doméně, relevantních článků není mnoho. Je možné zmínit například článek navrhující doporučovací systém pro akciové portfolio, který je založený na asociačních pravidlech (Paranjape-Voditel, Deshpande, 2013) nebo metodu detekce možného bankrotu společnosti s využitím dolování asociačních pravidel v kombinaci s ontologickým modelem vytvořeným na základě finančních reportů (Martin et al., 2011).

Metoda detekce anomálií je doméně financí bližší, této problematice byl například věnován workshop Detekce anomálií v oblasti financí[8] při jedné z největších konferencí věnované data miningu a data science (Knowledge Discovery and Data Mining). Z této oblasti můžeme uvést článek zabývající se detekcí anomálií v big data z oblasti financí (Mohiuddin, Shahadat, 2017) nebo přehledový článek o různých metodách detekce anomálií ve finanční doméně (Mohiuddin et al., 2016).

7 Závěr

Projekt OpenBudgets.eu ukázal, že využití pokročilých analytických metod je možné uplatnit i v doménách tímto směrem dosud prakticky nedotčených. I přes mnohá omezení, která jsou popsaná v Kapitole 5, se domníváme, že tyto metody mají v analýze fiskálních dat potenciál. Zejména metody detekce anomálií, vyvinuté jako poměrně generické, dobře odpovídají velké části požadavků vznášených zainteresovanými stranami. Oproti jednoduchým statistickým metodám umožňují odhalovat i anomálie charakterizované souběhem více dílčích faktorů spíše než prostou odlehlou hodnotou jedné veličiny. Vysoký potenciál mají i metody hledání asociačních pravidel, resp. obecněji, častých vzorů v datech. Zde je však nutno nejprve vyřešit problém velkého počtu překrývajících se pravidel/vzorů, který znesnadňuje práci doménových expertů při vyhodnocování výsledků.

Výběr použitých metod byl ovlivněn jak naformulovanými požadavky, tak i dostupností zdrojových dat, a metod pro jejich zpracování. Budoucí aplikace širšího spektra metod bude záviset zejména na postupném otevírání dalších veřejnosprávních dat ve strojově čitelné struktuře (např. auditorských zpráv s kódovanými výsledky auditu), ale i na rozvoji metod, které mohou dolovat přímo propojená data v síťové struktuře RDF (např. pro potřebu analýz sítí vztahů).

Poděkování

Výzkum byl podpořen z projektu *OpenBudgets.eu* (H2020-645833).

[8] https://sites.google.com/view/kdd-adf-2017

Seznam použité literatury

Agrawal, R., & Srikant, R. (1994). Fast Algoritm for Mining Association Rules in Large Databases. In *Proceedings of 20th International Conference on Very Large Data Bases* (pp. 487-499). San Francisco: Morgan Kaufmann Publisher.

Alloca, S. (2016). CFOs' Top Goal for 2017: Better Analysis and Reporting. CFO: Corporate Finance News and Events. Retrieved November 2, 2017, from http://ww2.cfo.com/analytics/2016/11/cfos-top-goal-2017-better-analysis-reporting/

Cyganiak, R., Wood, D., Lanthaler, M. (2014). RDF 1.1 Concepts and Abstract Syntax. *W3C Recommendation*. Retrieved November 2, 2017, from https://www.w3.org/TR/rdf11-concepts/

Dong, T., Musyaffa, F., Kuchař, J., Vojíř, S., Zeman, V., Chudán, D., Svátek, V., Koupidis, K., Chatzopoulou, A., Bratsas, Ch., Orlandi, F., & Merten, T. (2017). Deliverable 2.4 – Data Mining and Statistical Analytics Techniques. Retrieved November 2, 2017, from http://openbudgets.eu/assets/deliverables/D2.4.pdf

Engels, C., Musyaffa, F., Dong, T., Klímek, J., Mynarz, J., Orlandi, F., & Auer, S. (2016a). Deliverable 2.1 - Tools for Semantic Lifting of Multiformat Budgetary Data. Retrieved November 2, 2017, from http://openbudgets.eu/assets/deliverables/D2.1.pdf

Engels, C., Bratsas, Ch., Koupidis, K., Musyaffa, F., Orlandi, F., Chudán, D., Kuchař, J., Mynarz, J., & Zeman, V. (2016b). Deliverable 2.3 – Requirements for Statistical Analytics and Data Mining. Retrieved November 2, 2017, from http://openbudgets.eu/assets/deliverables/D2.3.pdf

Čerpání. (2017). Evropské strukturální a investiční fondy – Čerpání v období 2007-2013. Retrieved November 2, 2017, from http://dotaceeu.cz/cs/Fondy-EU/Predchozi-programova-obdobi/Programove-obdobi-2007-2013/Cerpani-v-obdobi-2007-2013

Hájek, P., Holeňa, M., & Rauch, J. (2010). The GUHA Method and its Meaning for Data Mining. *Journal of Computer and System Sciences,* 76(1), 34-48. doi: 10.1016/j.jcss.2009.05.2004

Harris, S., & Seaborne, A. (2013). SPARQL 1.1 Query Language. *W3C Recommendation 21 March 2013*. Retrieved November 2, 2017, from https://www.w3.org/TR/sparql11-query/

Chandola, V., Banerjee, A., & Vipin, K. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys,* 41(3), Article 15. doi: 10.1145/1541880.1541882

Journalism++ (2017). The Good, The Band and The Accountant. *GitHub*. Retrieved November 2, 2017, from https://jplusplus.github.io/the-accountant/#/

Kejkula, M. (2009). *Post-processing of association rules by multicriterial clustering method*. (Dissertation thesis). Prague: VŠE-FIS.

Koukal, B., Chudán, D., & Svátek, V. (2017). OLAP Recommender: Supporting Navigation in OLAP Cubes Using Association Rule Mining. In *Data a Znalosti 2017* (pp. 46-50). Plzeň: Západočeská univerzita v Plzni.

Kovalerchuk B., & Vityaev E. (2005). Data Mining for Financial Applications. In Maimon O., Rokach L. (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 1153-1169). Boston: Springer. doi: 10.1007/978-0-387-09823-4_60

Kliegr, T., Chudán, D., Hazucha, A., & Rauch, J. (2010). SEWEBAR-CMS: A System for Postprocessing Data Mining Models. In *Proceedings of the RuleML-2010 Challenge, at the 4th International Web Rule Symposium* (paper 9). Cáchy: CEUR Workshop Proceedings.

Klímek, J., Mynarz, J., Škoda, P., Zbranek, J., & Zeman, V. (2016). Deliverable 2.2 – Data optimisation, enrichment, and preparation for analysis. Retrieved November 2, 2017, from http://openbudgets.eu/assets/deliverables/D2.2.pdf

Kuchař, J. (2017). Fpmoutliers. *Github*. Retrieved November 2, 2017, from https://github.com/jaroslav-kuchar/fpmoutliers

Kuchař, J., Ashenfelter, A., & Kliegr, T. (2017). Outlier (Anomaly) Detection Modelling in PMML. In *Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017* (paper 9). Cáchy: CEUR Workshop Proceedings.

Kuchař, J., & Svátek V. (2017). Spotlighting Anomalies using Frequent Patterns. In *Proceedings of the KDD 2017 Workshop on Anomaly Detection in Finance*. Halifax: PMLR.

Kuchař, J., Vojíř, S., Zeman, V., Mynarz, J., Svátek, V., Koupidis, K., Chatzopoulou, A., Bratsas, Ch., Dong, T., Musyaffa, F., Wang, K., Orlandi, F., & Li, Y. (2017). Deliverable 2.5 - Data Mining Interfaces. Retrieved November 2, 2017, from http://openbudgets.eu/assets/deliverables/D2.5.pdf

Kliegr, T., Kuchař, J., Vojíř, S., & Zeman, V. (2017). EasyMiner – Short History of Research and Current Development. In *Proceedings of the 17th Conference on Information Technologies - Applications and Theory* (pp. 235–239). Cáchy: CEUR Workshop Proceedings.

Martin, A., Manjula, M., & Venkatesan, P. (2011). A Business Intelligence Model to Predict Bankruptcy using Financial Domain Ontology with Association Rule Mining Algorithm. *International Journal of Computer Sciences Issues*, 8(3), 211-218.

Mohiuddin, A., Choudhury, N., & Shahadat, U. (2017). Anomaly Detection in Big Financial Data. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 998-1001). New York: ACM. doi: 10.1145/3110025.3119402

Mohiuddin, A., Mahmood, A., N., & Islam, R. (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, 55(C), 278-288. doi: 10.1016/j.future.2015.01.001

Mynarz, J., Klímek, J., Dudáš, M., Škoda, P., Engels, C., Musyaffa, F. A., & Svátek, V. (2016). Reusable transformations of Data Cube Vocabulary datasets from the fiscal domain. In *Proceedings of the 4th International Workshop on Semantic Statistics co-located with 15th International Semantic Web Conference* (paper 04). Cáchy: CEUR Workshop Proceedings.

Ngaia, E.W.T., Hu, Y., Wong, Y.H., Yijun, Ch., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3), 559-569. doi: 10.1016/j.dss.2010.08.006

Orlandi, F., Dong, T., Karampatakis, S., Hernandez, P, Musyaffa, F., & Liu, H. (2017). Deliverable D7.7 Large-scale trail report including best-practices. Retrieved November 2, 2017, from https://drive.google.com/file/d/0B5ecBIVKmMmeVVYtRHNRMHh3emM/view

Paranjape-Voditel, P, Deshpande, U. (2013). A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing*, 13(2), 1055-1063. doi: 10.1016/j.asoc.2012.09.012

Rauch, J., & Šimůnek, M. (2014). *Dobývání znalostí z databází, LISp-Miner a GUHA*. Praha: Oeconomica.

Svátek V., Mynarz J., Węcel K., Klímek J., Knap T., & Nečaský M. (2014). Linked Open Data for Public Procurement. In Auer S., Bryl V., Tramp S. (Eds.), *Linked Open Data – Creating Knowledge Out of Interlinked Data* (pp. 196-213). Cham: Springer. doi: 10.1007/978-3-319-09846-3_10

Sagar, B. B., Singh, P., & Mallika, S. (2016). Online transaction fraud detection techniques: A review of data mining approaches. In *Proceesings of 3rd International Conference on Computing for Sustainable Global Development*. New York: IEEE.

Sharma, A., & Panigrahi, P.K. (2013). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. Retrieved November 2, 2017, from https://arxiv.org/abs/1309.3944

Tackling Corruption through Open Data. (2017). European Commission – Tackling Corruption through Open Data. Retrieved November 2, 2017, from http://ec.europa.eu/budget/euprojects/node/7645_en?language=en

Vafopoulos, M., Meimaris, M., Álvarez Rodríguez, J. M., Xidias, I., Klonaras, M., & Vafeiadis, G. (2013). Insights in global public spending. In *Proceedings of the 9th International Conference on Semantic Systems* (pp. 135-139). New York: ACM. doi: 10.1145/2506182.2506201

Vojíř, S., Zeman, V., Kuchař, J., & Kliegr, T. (2017a). Using EasyMiner API for Financial Data Analysis in the OpenBudgets.eu Project. In *Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017* (paper 21). Cáchy: CEUR Workshop Proceedings.

Vojíř, S., Zeman, V., Kuchař, J., & Kliegr, T. (2017b). Využití EasyMiner API v projektu OpenBudgets.eu. In *Data a Znalosti 2017* (pp. 56-60). Plzeň: Západočeská univerzita v Plzni.

Yue, D., Wu, X., Wang, Y., & Chu, C.-H. (2007). A Review of Data Mining-Based Financial Fraud Detection Research. In *Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing* (pp. 5519-5522). New York: IEEE. doi: 10.1109/WICOM.2007.1352

Zhang, D., & Zhou, L. (2004) Discovering Golden Nuggets: Data Mining in Financial Application. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, 34(4), 513-522. doi: 10.1109/TSMCC.2004.829279

The article has been reviewed. | **Received:** 22 October 2017 | **Accepted:** 7 December 2017

Academic Editor: Stanislava Mildeova