Využití EasyMiner API v projektu OpenBudgets.eu

Stanislav Vojíř[1], Václav Zeman[1], Jaroslav Kuchař[1, 2], Tomáš Kliegr[1]

[1]Katedra informačního a znalostního inženýrství, FIS, Vysoká škola ekonomická v Praze
nám. W. Churchilla 1938/4, 13067 Praha 3 – Žižkov
[2]Web Intelligence Research Group, FIT, České vysoké učení technické v Praze
Thákurova 2700/9, 160 00 Praha 6 - Dejvice

{stanislav.vojir|vaclav.zeman|tomas.kliegr}@vse.cz
jaroslav.kuchar@fit.cvut.cz

Abstrakt. V souvislosti s rostoucí popularitou využívání data miningových dat lze registrovat také rostoucí poptávku po možnosti integrace data miningových algoritmů a systémů do komplexnějších, uživatelsky přívětivějších aplikací. Tento příspěvek prezentuje novou verzi systému EasyMiner, integrovanou do softwarového řešení vyvíjeného v rámci evropského projektu OpenBudgets.eu, který je zaměřen na zpřístupňování a analýzy finančních dat samospráv. EasyMiner je webový data miningový systém podporující dolování asociačních pravidel, tvorbu klasifikačních modelů a v současné verzi nově také detekci outlierů. Příslušná funkcionalita je k dispozici nejen prostřednictvím grafického uživatelského rozhraní, ale také prostřednictvím komplexního REST API.

Klíčová slova: asociační pravidla, klasifikace, detekce outlierů, data mining, REST API, EasyMiner, OpenBudgets.eu.

1 Úvod

EasyMiner (*http://easyminer.eu*) je webový data miningový systém dlouhodobě vyvíjený na Katedře informačního a znalostního inženýrství Vysoké školy ekonomické v Praze. Starší verze tohoto systému [1] byly zaměřeny zejména na zpřístupnění dolování asociačních pravidel a posléze tvorbu klasifikačních modelů v přehledném, grafickém uživatelském rozhraní, fungujícím ve všech moderních webových prohlížečích. Tato možnost je stále dostupná, avšak pro možnost automatizovaných analýz většího množství dat a propojitelnost s dalšími systémy bylo nutné celý systém EasyMiner dále rozvíjet.

Tento příspěvek popisuje novou verzi systému EasyMiner, použitou pro analýzu finančních dat v rámci evropského projektu OpenBudgets.eu. Aktuální verze podporuje nejen dolování asociačních pravidel, ale také *detekci outlierů*[1] založenou na dolo-

[1] *Outlier* – z hlediska překladu do češtiny se jedná o „odlehlé hodnoty" – instance dat charakterizované atributy s konkrétními hodnotami, které se odlišují od zbytku datové matice; míru odlišnosti charakterizuje *outlier score.*

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 56-60.

vání častých vzorů (*frequent patterns*). Zároveň jsou všechny funkce EasyMineru dostupné prostřednictvím nového REST API.

Zahraniční komunitě byla tato nová verze systému EasyMiner prezentována v rámci konference RuleML 2017 [2].

2 EasyMiner API

Předchozí verze systému EasyMiner podporovala jednoduché dolování asociačních pravidel v grafickém uživatelském rozhraní. Ačkoliv byla tato varianta vhodná pro koncové uživatele, nebylo možné využívat funkcionality systému EasyMiner v rámci rozsáhlejších projektů. Za tímto účelem bylo vytvořeno nové REST API podporující jednak veškerou původní funkcionalitu systému EasyMiner a zároveň také nově implementované algoritmy pro předzpracování dat a detekci outlierů.

Prostřednictvím daného API je integrována funkcionalita systému EasyMiner do projektu OpenBudgets.eu. Obdobně je možné jej využít také v dalších projektech – pro tvorbu *mashup aplikací*, začlenění do vlastních skriptů atp., což je podpořeno jeho plnou dokumentovaností a open source licencí celého systému EasyMiner.

2.1 Data mining prostřednictvím API

Komplexní REST API pro koncové uživatele je dostupné na adrese *<easyminer-server>/easyminercenter/api*. Tato adresa je rozcestníkem celého API a zároveň je na ní k dispozici také kompletní dokumentace v syntaxi *Swagger*.[2]

Pro využití API musí mít uživatel nejprve vytvořený vlastní unikátní API klíč, který získá prostřednictvím registrace uživatelského účtu. Tento API klíč musí být zasílán ve všech jednotlivých požadavcích na API. Následný postup přípravy dat a dolování pomocí API je obdobný jako při použití grafického rozhraní – viz následující odstavce.

Postup pro dolování asociačních pravidel: 1. nahrání dat ve formátu CSV, **2.** vytvoření *instance mineru*,[3] **3.** předzpracování dat (vytvoření atributů z nahraných dat za využití jednoduchých definic předzpracování), **4.** zadání data miningové úlohy (vzor asociačních pravidel, požadované míry zajímavosti), **5.** spuštění dolování, vyčkání na výsledky a **6.** zpracování výsledků (export ve formátu PMML[4] či JSON).

Postup pro detekci outlierů: Kroky 1.-3. jsou totožné jako u dolování asociačních pravidel. Následující kroky jsou: **4.** definice úlohy detekce outlierů (minimální podpora [support]), **5.** spuštění úlohy detekce outlierů a vyčkání na výsledky, **6.** procházení datových řádků uspořádaných podle *outlier score*.[1]

Předzpracování dat: Pro použití algoritmů založených na dolování frekventovaných vzorů je nutné nejprve předzpracovat (diskretizovat) hodnoty číselných dato-

[2] *Swagger* – framework pro tvorbu dokumentace pro REST API, *https://swagger.io/*

[3] EasyMiner podporuje dvě verze backendů – pro dolování pomocí systému R či pomocí Hadoop serveru. Při vytvoření *instance mineru* dochází k přípravě databáze a inicializaci příslušného backendu.

[4] PMML – XML formát pro záznam data miningových modelů

vých sloupců. Pro řadu případů použití je uživateli vyžadováno také předzpracování hodnot sloupců výčtových. EasyMiner API podporuje všechny základní metody předzpracování dat – převzetí původních hodnot, uživatelsky definované množiny hodnot či intervaly, automaticky vygenerované intervaly (*equidistant, equifrequent, equisized*).

V rámci kroku předzpracování dat musí uživatel předzpracovat všechny datové sloupce, které chce použít pro následnou data miningovou úlohu (tj. vytvořit z nich *atributy*, které se následně nacházejí ve výsledcích). Při dolování asociačních pravidel využívá uživatel zvolené atributy pro definování vzoru hledaných pravidel – má tedy možnost využívat atributy opakovaně v rámci většího množství úloh. V případě úlohy hledání outlierů jsou využity všechny připravené atributy.

2.2 Použité algoritmy

Základní úlohou podporovanou systémem EasyMiner je dolování asociačních pravidel, volitelně s možností využití jejich prořezání za účelem tvorby klasifikačního modelu. Za tímto účelem je v současné době využíván algoritmus *apriori* [3], implementovaný v balíčku *arules* pro systém R. Z hlediska architektury systému EasyMiner je tento algoritmus spouštěn prostřednictvím *R serveru*, jehož funkcionalita je zprostředkována dolovací službou. Pro velké datasety je využíván také *Hadoop server*, konkrétně algoritmus *FP-Growth* [4]. Protvorbu klasifikačních modelů je využívána vlastní implementace algoritmu *CBA*. [5]

Výsledkem dolování jsou asociační pravidla ve tvaru *antecedent* → *konsekvent*, ve kterých mohou být *antecedent* i *konsekvent* tvořeny konkjunkcí atributů s konkrétními hodnotami. V případě tvorby klasifikačních modelů platí je konsekvent omezen na jeden cílový atribut. Použitelnými měrami zajímavosti jsou *konfidence*, *podpora* a *lift*.

V současné době podporuje EasyMiner také detekci outlierů. Za tímto účelem byl Jaroslavem Kuchařem implementován balíček *fpmoutliers* pro systém R [6]. Podporovanými algoritmy jsou *FPCOF*, *FPOF*, *MFPOF*, *WCFPOF*, *WFPOF*, *LFPOF* a nový inovativní přístup *FPI*. [7] Výsledkem detekce outlierů jsou datové řádky z předzpracovaného datového souboru seřazené podle *outlier score*.[1]

3 Využití EasyMineru v systému OpenBudgets.eu

OpenBudgets.eu (H2020-645833, *http://openbudgets.eu*) je evropský projekt zaměřený na zpřístupňování a analýzy finančních a rozpočtových dat samospráv. Součástí tohoto projektu, v rámci pracovního balíčku *WP2 Infrastructure for Data Collection and Mining* byly zkoumány možnosti analýzy finančních dat data miningovými algoritmy a nástroji. Jedním z vybraných nástrojů je také systém EasyMiner. Na základě analýzy požadavků byla do systému EasyMiner implementována podpora dolování outlierů a komplexní REST API.

Všechna data analyzovaná v projektu OpenBudgets.eu jsou dostupná ve formátu RDF.[5] Prostřednictvím nástrojů pro přípravu dat jsou data získaná od samospráv konvertována do formátu RDF, rozšířena o data z veřejně dostupných zdrojů (například z registrů ekonomických subjektů) a posléze zpřístupněna pro analýzy. Pro použití EasyMineru jsou data následně konvertována do CSV nástrojem *LinkedPipes ETL* [8].

Z hlediska integrace do softwarové architektury vyvíjené v projektu OpenBudgets.eu je funkcionalita EasyMineru využívána prostřednictvím API, jehož funkce jsou volány z integračního nástroje *DAM* (*http://github.com/openbudgets/DAM*).

4 Závěr

Data miningový systém EasyMiner je experimentálním akademickým projektem, dostupným pod open source licencí *Apache License, Version 2.0*. Jeho nová verze disponuje nejen grafickým uživatelským rozhraním pro dolování asociačních pravidel, ale podporuje také integraci do dalších projektů prostřednictvím API. Komplexní příklad využití EasyMiner API je dostupný na *http://www.easyminer.eu/api-tutorial*.

Z hlediska reálného nasazení je systém EasyMiner v současné době využíván v rámci data miningové architektury vytvářené v projektu OpenBudgets.eu, zaměřeném na analýzy finančních a rozpočtových dat.

V rámci budoucího vývoje by měly být systém dále rozšířen o podporu dolování nad RDF daty a mělo by dojít k posunu současných limitů týkajících se velikosti použitelných dat.

Literatura

1. Vojíř, S., Zeman, V., Kuchař, J., Kliegr, T.: EasyMiner/R Preview: Towards a Web Interface for Association Rule Learning and Classification in R. In: Proceedings of the RuleML 2015 Challenge, the Special Track on Rule-based Recommender Systems for the Web of Data, the Special Industry Track and the RuleML 2015 Doctoral Consortium hosted by the 9th International Web Rule Symposium (RuleML 2015), Berlin, Germany (2015). http://ceur-ws.org/Vol-1417/paper10.pdf
2. Vojíř, S., Zeman, V., Kuchař, J., Kliegr, T.: 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 hosted by International Joint Conference on Rules and Reasoning 2017 (RuleML+RR 2017), London, UK (2017). http://ceur-ws.org/Vol-1875/paper21.pdf
3. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: SIGMOD (1993) 207-216.
4. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. In: Data Mining and Knowledge Discovery 8 (2004) 53–87.

[5] *RDF* – datový formát pro sémantický web, *https://www.w3.org/RDF/*

5. Kliegr, T., Kuchař, J., Sottara, D., Vojíř, S.: Learning Business Rules with Association Rule Classifiers. In: International Workshop on Rules and Rule Markup Languages for the Semantic Web, Springer (2014) 236–250.
6. Kuchař, J.: jaroslav-kuchar/fpmoutliers. https://github.com/jaroslav-kuchar/fpmoutliers [15. 6. 2017].
7. Kuchař, J,; Svátek, V.: Spotlighting Anomalies using Frequent Patterns. In: Proceedings of the KDD 2017 Workshop on Anomaly Detection in Finance, PMLR, Halifax, Nova Scotia, Canada (2017).
8. Klímek, J., Škoda, P., Nečaský, M.: LinkedPipes ETL: Evolved linked data preparation. In: International Semantic Web Conference, Springer (2016) 95-100.

Poděkování: Tento článek vznikl díky podpoře z projektů OpenBudgets.eu (H2020-645833) a IGA 29/2016 Vysoké školy ekonomické v Praze.

Annotation:

Using EasyMiner API in the OpenBudgets.eu Project

Related to the increasing popularity of data mining there is a growing effort to integrate data mining algorithms and systems into user-friendly applications and information systems. This paper introduces a new version of web-based data mining system EasyMiner and its integration into a software solution developed within the European project OpenBudgets.eu. This project is aimed at publication and analysis of financial data of municipalities. The current version of EasyMiner supports mining of association rules, building of classification models and newly also outlier detection. Its functionality is available not only via a graphical user interface, bus also via REST API. The API can be easily used also from third party applications.