Automatizace klasifikace evropských projektů pomocí klasifikátoru

Ondřej Zamazal

Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
nám. W. Churchilla 1938/4, 130 67 Praha 3, Česká republika

ondrej.zamazal@vse.cz

Abstrakt. Finanční prostředky Evropské unie do vytváření pracovních míst, do evropské ekonomiky a životního prostředí jsou poskytovány prostřednictvím pěti evropských strukturálních a investičních fondů (ESI fondy). Ačkoliv EU má pro kategorizaci jednotlivých projektů jednotný kategorizační systém, tak jednotlivé členské země EU používají různé své vlastní kategorizační systémy. Některé země již přímo používají nový kategorizační systém, ale mnohé země takto stále ještě nepostupují. V dostupných datasetech o evropských projektech tak zůstává značné množství projektů nezařazeno s ohledem na jednotný kategorizační systém EU. Cílem této práce je vyzkoušení možnosti automatické klasifikace evropských projektů pomocí klasifikátoru. Podpora automatickou klasifikací by podpořila fiskální analýzy.

Klíčová slova: evropský projekt, číselník, strojové učení, klasifikace

1 Úvod

Evropská unie podporuje v jednotlivých členských zemích projekty pro vytváření pracovních míst, zdravou evropskou ekonomiku a dobré životní prostředí prostřednictvím pěti evropských strukturálních a investičních fondů (ESI fondy). Pro kategorizaci těchto evropských projektů je k dispozici jednotná kategorizace pro období 2007 – 2013 a zvlášť pro období 2014 – 2020 s explicitním propojením.[1] Členské země EU však při evidování evropských projektů tuto kategorizaci uvádět nemusí a také tak všechny nečiní. Kategorizace projektů se tak musí dělat ručním způsobem až ex post. Jednotný způsob kategorizace evropských projektů umožňuje přímočaré fiskální analýzy. Motivací této práce je podpořit klasifikaci evropských projektů do kategorizačního systému pro období 2014 – 2020 automatickými prostředky strojového učení, což by umožnilo fiskální analýzy. V tomto příspěvku se zabýváme přístupem strojo-

[1] http://ec.europa.eu/regional_policy/sources/docgener/evaluation/data/categorisation_2014_2020_mapping.xls

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 141-145.

vého učení SVM pro automatickou klasifikaci evropských projektů. V části 2 vysvětlujeme použitá data. V části 3 popisujeme celkový postup přístupu a úvodní experimenty jsou popsány v části 4.

2 Otevřená data evropských projektů

Jednotlivé země zveřejňují informace o evropských projektech na úrovni regionů a celých zemí. Přestože data jsou vystavena odpovědnými orgány, tak jejich formát zůstává poněkud „zavřený". Někdy jsou data vystavena ve formátu pdf, byť samotná data jsou zapsána ve formě tabulek, jindy jsou data vystavena rovnou v příhodnějším tabulkovém formátu. Sběrem a sdílením těchto dat se zabývá organizace „Open Knowledge Foundation Deutschland".[2] Datasety jsou ukládané v různých formátech CSV, XSLX nebo JSON. V rámci zpracování dat se také provádí provázání jednotlivých typů informací (atributů) na jednotný fiskální datový model Open Spending.[3] Kromě vystavených datasetů je k dispozici také jednotný integrovaný dataset [2,3], který lze stáhnout (CSV formát, 879 MB, 2,7 miliónů řádků popisů projektů) a analyzovat vlastními silami nebo za použití připraveného analytického nástroje OS Viewever.[4]

Nový kategorizační systém projektů (podobně jako ten starý) je dostupný v příslušných evropských dokumentech a také v tabulkové podobě z webu „Data for research".[5] Podle informací z tabulkových dat jsme z nich vyextrahovali samostatný RDF [4] číselník obsahující kategorizační systém pro staré odbobí 2007-2013, RDF číselník obsahující klasifikační systém pro nové období 2014-2020 a jejich vzájemná mapování.[6]

3 Postup automatické klasifikace evropských projektů

V rámci prvního přístupu ke klasifikaci projektů jsme pracovali s jednotlivými datasety regionů a zemí samostatně. Datasety obsahují popisy projektů s velmi různorodou mírou záběru (např. některé popisy obsahují jen název projektu, dotace EU, region a termín jiné obsahují také slovní popis projektu ad). Kromě numerického popisu projektů tak bývá k dispozici také slovní popis různého druhu. V našem přístupu se zaměřujeme na využití slov uvedených u jednotlivých projektů. V první fázi jsme dávali dohromady dostupná data. Poloautomaticky jsme našli 20 datasetů z celkových 110, ve kterých se objevuje atribut týkající se informace zařazení do kategorie podle intervenčního kódu. Z těchto datasetů jsme v rámci předzpracování pomocí regulárních

[2] https://www.okfn.de/en/

[3] https://github.com/os-data/eu-structural-funds/blob/master/specifications/fiscal.schema.yaml

[4] http://subsidystories.eu

[5] http://ec.europa.eu/regional_policy/sources/docgener/evaluation/data/categorisation_2014_2020_mapping.xls

[6] viz https://github.com/openbudgets/ v části linksets a Code-lists.

výrazů a přesné podoby intervenčních kódů podle RDF číselníku extrahovali intervenční kód a dále samostatná slova jednotně oddělena mezerami. Takto vzniklá data jsme deduplikovali.

Za účelem „sémantického" propojení slov používáme automatický jazykový překlad všech slov do angličtiny, protože jednak automatický překlad do angličtiny bývá nejlepší a jednak angličtina je používána v popiscích jednotlivých intervenčních kódů v RDF číselníku. Pro automatický jazykový překlad jsme vybrali „Microsoft Translator API".[7] Výběr byl proveden na základě dostupnosti API zdarma (2 milióny přeložených znaků za měsíc) a pokrytí všech požadovaných evropských jazyků.

Jednotlivé projekty tak reprezentujeme pomocí vektorů, které mají tolik složek kolik je v datech unikátních slov (po odstranění stop slov) a binárně sledujeme jejich (ne)výskyt. Přirozené úskalí je velká dimenze vektorů, která by mohla být snížena pomocí shlukování slov nebo například latentní sémantickou analýzou, kde by se identifikovaly koncepty složené ze slov a těmi se indexovaly jednotlivé projekty [1]. V rámci našeho přístupu jsme dimenzi vektorů nesnižovali a spíše jsme se zaměřili na algoritmy, které zvládají práci s daty o vysoké dimenzionalitě. Na základě úvodního testování jsme se rozhodli pro vytvoření SVM (Support Vector Machines) klasifikátoru[8], za použití knihovny LibSVM pro Javu, a jeho lineárního kernelu, který oproti jiným kernelům (RBF kernel, polynomiální kernel) dosahoval nejlepší přesnosti na trénovacích datech.

Úvodní analýzy ukázaly, že získaná data jsou rozložena do tříd (123 intervenčních kódů) značně nerovnoměrně. Celkem 10992 instancí patří do 97 tříd, z nichž 23 tříd má pouze jednu instanci, 43 tříd má alespoň 10 instancí, 34 tříd má alespoň 20 instancí a pouze 20 tříd má více než 100 instancí. Nízké zastoupení velkého množství tříd v trénovacích datech a dominance několika tříd by při automatické klasifikaci vedlo k preferování majoritních tříd a znehodnocení výsledků na reálných datech. V rámci našeho přístupu jsme úskalí nevyváženosti tříd řešili pomocí vzorkování.

4 Experiment automatické klasifikace evropských projektů

Pro vyzkoušení našeho přístupu jsme provedli experimenty A, B, C a D se získanými daty s různým přístupem k vytvoření trénovacích a testovacích dat. V rámci přípravy trénovacích dat vždy používáme vzorkování tak, aby všechny třídy byly zastoupeny stovkou instancí. K tomu se používá podle potřeby „oversampling" tak, že se náhodně některé instance duplikují a „undersampling" tak, že se některé náhodně vybrané instance smažou. Experiment A pro trénování uvažuje jen třídy s více než 50 instancemi (celkem 21 tříd). Experiment B pro trénování používá všechny třídy s alespoň jednou instancí (97 tříd). Experimenty C a D uvažují všechny třídy pro trénování (123 tříd), kde pro třídy bez instancí jsou použity popisky z RDF číselníku. Pro sestavení testovacích dat v experimentech A, B a C se vybírá náhodně 20 instancí ze tříd, které

[7] https://www.microsoft.com/en-us/translator/translatorapi.aspx

[8] Dále jsme v nástroji Weka zkoušeli logistickou regresi (neúspěšně kvůli příliš vysoké dimenzionalitě dat) a algoritmus pro učení rozhodovacích pravidel JRip (srovnatelná přesnost jako SVM ale příliš pomalé).

mají více než 50 instancí. Experiment D do testovacích dat zařazuje vždy polovinu instancí od každé třídy s více než jednou instancí. Testovací a trénovací množiny jsou vždy disjunktní. Testování bylo spuštěno 3x a výsledná přesnost klasifikace je jejich průměrem viz tabulka 1 spolu s uvedenými charakteristikami jednotlivých experimentů.

Výsledná přesnost pro všechny experimenty vychází relativně vysoká. Nejlépe vypovídající je experiment D, kde trénujeme a testujeme na nejvíce třídách a nejvíce instancích.

Tabulka 1: experimenty (Tr | Ts znamená trénovací | testovací množina)

Experiment	Tr: #tříd (#instancí)	Ts: #tříd (#instancí)	Přesnost
A	21 (2100)	21 (480)	0.851
B	97 (9700)	21 (480)	0.95
C	123 (12300)	21 (480)	0.951
D	123 (12300)	74 (3653)	0.93

5 Závěr

Provedené experimenty ukazují slibné výsledky za situace, kdy máme získaná data rozdělena na disjunktní množiny trénovacích a testovacích dat. Vypovídající schopnost těchto experimentů je však omezená. Testovací data totiž pocházejí z datasetu zemí, ze kterých byla použita data (sice jiná než ta pro testování) pro trénování klasifikátoru a současně platí, že struktura popisu projektů v rámci jednoho datasetu bývá dosti podobná. Závěrem proto musíme konstatovat, že opravdu vypovídající výsledky testování mohou být dosáhnuty jen na testovacích datech datasetů, ze kterých žádné instance nebyly použity pro trénování. To je také záměrem naší plánované práce s integrovaným datasetem evropských projektů.

Literatura

1. Berka, Petr. Dobývání znalostí z databází. Academia, 2003.
2. SubsidyStories.eu. Dataset Descriptions. https://github.com/os-data/eu-structural-funds/blob/master/documentation/subsidyreport%20-%20dataset%20descriptions.pdf
3. SubsidyStories.eu. Methodology & Variables. 2017. https://github.com/os-data/eu-structural-funds/blob/master/documentation/subsidyreport%20-%20methodology.pdf
4. World Wide Web Consortium. RDF 1.1 concepts and abstract syntax. (2014).

Poděkování: Tato práce byla podpořena z projektu EU H2020 č. 645833 OpenBudgets.

Annotation:

Automatic Classification of European projects

European union funding for job creation and a sustainable and healthy European economy and environment is channelled through the 5 European structural and investment funds (ESIF). Although there is European categorization system for EU projects, EU countries apply their own different categorization systems. Some EU countries already apply European categorization system, but many do not. As a result, many projects are not categorized using European categorization system in available datasets. The goal of this work is to examine an option of an automatic classification of European projects using a machine learning classifier. This would enable straightforward fiscal analyses and interlinking categorization systems of EU countries to European one.