Středoškolská

odborná činnost

Strojové učení, dobývání znalostí a detekce anomálií

Autor: Kristýna Němcová

Kraj: Jihomoravský

Obor: 18. Informatika

Brno 2017

Středoškolská

odborná činnost

Strojové učení, dobývání znalostí a detekce anomálií

Machine learning, data mining and outlier detection

Autor: Kristýna Němcová

Škola: Gymnázium Brno, tř. Kpt. Jaroše 14, 658 70 Brno

Kraj: Jihomoravský

Školitel: doc. RNDr. Lubomír Popelínský, Ph.D.

Obor: 18. Informatika

Brno 2017

Čestné prohlášení:

Prohlašuji, že jsem práci na téma „Strojové učení, dobývání znalostí a detekce anomálií“ zpracovala sama pod vedením doc. RNDr. Lubomíra Popelínského, Ph.D. Veškeré prameny a zdroje informací, které jsem použila k sepsání této práce, jsou citovány v poznámkách pod čarou a uvedeny v seznamu použitých pramenů a literatury.

Dále prohlašuji, že tištěná i elektronická verze práce SOČ jsou shodné a nemám závažný důvod proti zpřístupňování této práce v souladu se zákonem č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a změně některých zákonů (autorský zákon) v platném znění.

V Brně dne 17. února 2017 …….....

Kristýna Němcová

Masarykova univerzita, Fakulta informatiky

Katedra teorie programování, Laboratoř dobývání znalostí

Poděkování:

Děkuji vedoucímu mé práce doc. RNDr. Lubomíru Popelínskému, Ph.D. za to, že se ochotně ujal odborného vedení práce a za jeho vstřícný přístup. Dále děkuji RNDr. Karlu Vaculíkovi za cenné připomínky, které mi pomohly práci vypracovat.

Tato práce byla provedena za finanční podpory Jihomoravského kraje.

Anotace:

Práce se zabývá využitím metod detekce anomálií na datech z databáze filmů IMDb. Specializuje se na recenze filmů od uživatelů, konkrétně srovnávání textu recenze a počtu přiřazených hvězdiček. Teoretická část se zabývá strojovým učením a detekcí anomálií. V praktické části byl vytvořen program v jazyce R pro předzpracování dat, klasifikaci a detekci anomálií na textech recenzí. Pro detekci anomálií na datech rozdělených do tříd byly použity programy RapidMiner a Weka. Také jsou zde prezentovány výsledky těchto metod klasifikace a detekce anomálií. Na konci práce jsou shrnuty závěry vycházející z výsledků a jsou nastíněny možnosti budoucího rozšíření programu.

Klíčová slova:

detekce anomálií; strojové učení; dobývání znalostí; filmové recenze; analýza textu; jazyk R

Abstract:

The study deals with usage of outlier detection of data in the Internet Movie Database. It specializes on film reviews written by users, concretely comparing review text and the number of assigned stars. Machine learning and outlier detection is analyzed in theoretical part. In practical part, a program in R language for preprocessing, classification and outlier detection of review text was created. For class outlier detection, there was used a program RapidMiner and Weka. There are also presented results of used methods of classification and outlier detection. Conclusions based on the results are summarized and the possibilities of future expansion of the program are outlined in the rear of this study.

Keywords:

outlier detection; machine learning; data mining; movie reviews; text analysis; R language

Obsah

Úvod

Práce se zabývá detekcí anomálií. Tato oblast úzce souvisí se strojovým učením a dobýváním znalostí. Všechny využívají poznatky z informatiky a statistiky. Strojové učení je užitečné napříč mnohými obory od rozlišení spamů až po predikci vzniku rakoviny či infarktu. Díky detekci anomálií můžeme detekovat neznámé hrozby v počítačových sítích a chyby či podvody při finančních transakcích. Dobývání znalostí poslouží všude tam, kde jsou data a zájem najít zajímavé souvislosti – od knihoven po podnikový management. Dobývání znalostí je proces analýzy dat zahrnující výběr a vyčištění dat, předzpracování, analýzu metodami strojového učení a interpretaci výsledků. Snaží se najít dříve neznámé vztahy a souvislosti v datech.

V úvodní teoretické kapitole rozeberu základní pojmy strojového učení a detekce anomálií. Přednostně se zaměřím na nástroje používané v praktické části, a to metody strojového učení – rozhodovací stromy a náhodné lesy, detekce anomálií – LOF a detekce anomálií na klasifikovaných datech – CODB a RF-OEX.

V praktické části využiji datovou sadu Large Movie Review Dataset, která využívá volně dostupná data z recenzí filmů z webové stránky IMDb, což je on-line databáze informací o filmech, hercích a dalších, psána v anglickém jazyce. Využívám ji právě kvůli veřejně přístupným datům, která při zajímavých problémech nejsou vždy dostupná. Anglický jazyk je také výhodou z důvodu dostupnosti algoritmů předzpracovávajících data v tomto jazyce. Mým záměrem je najít anomální recenze, jejichž hodnocení uživatelem neodpovídá citovému zabarvení textu psaného týmž uživatelem.

Nejdříve pomocí strojového učení ověřím, zda jsem schopna predikovat negativní čí pozitivní hodnocení z textu recenze. Posléze na data použiji metody detekce anomálií. Následně posoudím, z jakého důvodu byly texty hodnoceny jako anomální.

1 Teoretická část

1.1 Strojové učení

Základem inteligence je schopnost učit se. Bez schopnosti poučit se ze zkušeností by živé organismy nebyly schopny se adaptovat. Proto se v rámci umělé inteligence zkoumají metody učení.[1] Strojové učení se často dělí na učení s učitelem, učení bez učitele, učení asociačních pravidel a detekci anomálií.

Učení s učitelem pracuje na základě označených trénovacích dat, kde každý prvek má číselnou hodnotu či třídu, například motorka a auto. Na základě těchto dat algoritmus vytvoří model, který se aplikuje na testovací data. Abychom zjistili, jak bylo klasifikování účinné – porovnáme třídy, které jim přidělil, s těmi původními. Mohou zde být falešně pozitivní a falešně negativní případy. Falešně negativní jsou například motorky, které nejsou klasifikovány jako motorky. Když se auta objeví v množině motorek, jsou falešně pozitivní v množině motorek. Poznamenejme, že některé z těchto příkladů mohou být anomálie.[2] Výstupem učení s učitelem v případě klasifikace je třída, do které daný prvek náleží, a v případě regrese je to odhadnutá číselná hodnota dle zadaných dat.[3]

Učení bez učitele pracuje na základě neoznačených dat. Snaží se je dle rozdílností a podobností rozdělit na množiny – shluky. Uvedu příklad: dítě je schopné vybrat z množiny motorových vozidel motorky, i když neví, co to motorka je. Oddělí je na základě toho, že mají dvě kola, což je nezřetelnější odlišnost od ostatních vozidel. Rozdělená data se sjednocují do shluků na základě podobností.[4]

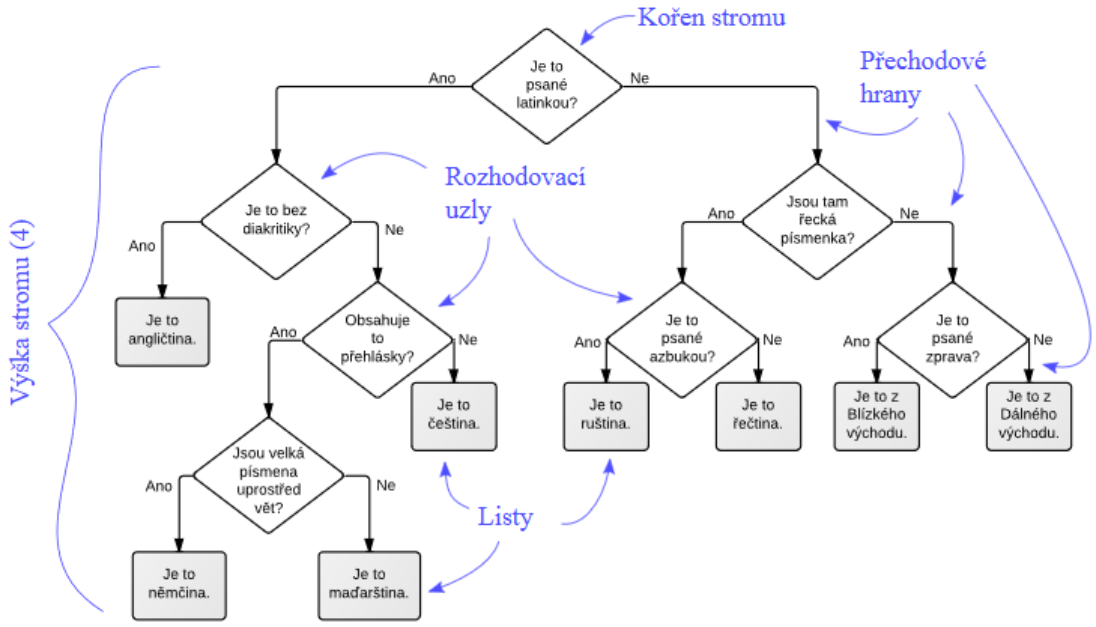
Učení asociačních pravidel hledá v datech často společně se vyskytující jevy, implikace (A⇒B). Asociačním pravidlem při analýze nákupního košíku je například – pokud si zákazník v supermarketu koupí chléb a máslo, tak si koupí i mléko.[5]

1.1.1 Rozhodovací stromy

Rozhodovací stromy náleží do skupiny učení s učitelem. Je možné je využít pro klasifikaci i regresi. Zaměřím se pouze na klasifikaci, ale při regresi se jedná o obdobný

[1] BERKA, Petr. *Dobývání znalostí z databází*. 2003.
[2] KULKARNI, Parag. *Reinforcement and systemic machine learning for decision making*. 2012.
[3] KLASCHKA, Jan; KOTRČ, Emil. Klasifikační a regresní lesy. 2004.
[4] KULKARNI, 2012, op. cit.
[5] BERKA, 2003, op. cit.

proces, při kterém je třída spojitá. Jde o jednoduchý model zobrazitelný jako strom, který nám navíc ukáže, co charakterizuje jednotlivé třídy. V případě často používaného binárního stromu se v každém uzlu strom větví na dvě další cesty. Na označených trénovacích datech si vytvoří vlastní rozhodovací strom s kořenem, kterým jsou celá trénovací data. Vyhodnotí se, na kterých parametrech závisí nejvíce. Tyto parametry se stanou tzv. prediktory, které rozdělují data do větví. Model je zakončen tzv. listy, tedy uzly, které se již nevětví a obsahují označení třídy. Tento strom se poté aplikuje pro predikci na testovacích datech.



Obr. 1: Rozhodovací strom.

Například máme text a chceme zjistit, v jakém je jazyce. Nejprve vezmeme množinu textů v různých jazycích – v tomto případě 8 tříd a strom se naučí, jak je rozlišovat. Nebude například brát v úvahu interpunkci, protože to mají všechny skupiny společné, a zaměří se na jejich rozdílnosti, jako typ písma, diakritika, velká písmena…

Rozhodovací stromy můžeme využít i pro detekci anomálií, a to za předpokladu, že máme data o anomáliích. Z toho plyne, že se jedná o metodu s učitelem. Může se jednat o klasifikovaná data (anomálie – ano/ne) nebo můžeme využít regrese a použít data z jiné metody, jejíž výsledkem je míra anomálnosti jednotlivých prvků. Druhého případu využíváme spíše z hlediska informačního, hledáme, podle čeho se metody rozhodují o anomálnosti dat. Rozhodovací strom je na to vhodný z hlediska jeho jednoduchosti a průhlednosti.

1.1.2 Náhodné lesy

Náhodné lesy náleží do skupiny učení s učitelem. Je možné je využít pro klasifikaci i regresi. Náhodný les je model vytvořený určitým počtem rozhodovacích stromů na základě trénovacích dat. Lesy jako nadstavba rozhodovacích stromů jsou přesnější v klasifikaci a predikci a jsou stabilnější. Ztrácí se zde ovšem jednoduchost a průhlednost stromu. Náhodné lesy patří mezi ansámblové metody (ensemble learners), mezi něž patří kromě náhodných lesů především bagging a boosting.

Bagging (Bootstrap aggregating) využívá soubor stromů tvořených ze souborů vybraných bootstrapově. Bootstrapové výběry se provádí náhodně a s opakováním. To si lze jednoduše představit na příkladu náhodného výběru čísla na základě náhodného výběru jednotlivých číslic. Nejprve náhodně vybereme první číslici, tu ovšem nevyřadíme z výběru pro další číslici. Stejně tak při tvorbě dalšího čísla se neomezíme pouze na číslice, které neobsahovalo číslo předchozí, ale vybíráme ze všech. To je výhoda při malém počtu dat, jelikož se nezmenšuje množina výběru. Můžeme tak i z malých počátečních dat utvořit mnoho trénovacích a testovacích dat.[6] Nevýhodou je, že některá data se mnohokrát duplikují a některá nejsou využita vůbec. Počet těch, která nebudou vybrána, je přibližně 37 %.[7] Z vybraných dat se vytvoří trénovací strom a z nevybraných se stanou testovací data, která následně určují odhad jeho chyby. Při klasifikaci pomocí Baggingu projdou klasifikovaná data každým stromem a výsledkem je většinové hlasování se stejnými váhami těchto stromů. Při regresi jde o prostý aritmetický průměr z výsledků každého stromu.[8]

Boosting pozměňuje váhy jednotlivých pozorování. Od počátečního vektoru se váhy pozorování sníží, pokud byl správně klasifikován dalším modelem, nebo zvýší, pokud došlo k chybné klasifikaci. Díky tomu je pozornost zaměřena na náročnější případy, při kterých dochází ke špatné klasifikaci. Při hlasování mají vyšší váhu modely s nižší chybovostí.[9]

Random Forest využívá stejných principů jako Bagging (Bootstrap aggregating), ale navíc ošetřuje korelaci mezi jednotlivými stromy, která může způsobit nadhodnocení konečných výsledků.[10] Pro každý uzel vybere náhodnou podmnožinu

[6] KOMPRDOVÁ, Klára. *Rozhodovací stromy a lesy.* 2012.
[7] BREIMAN, Leo: Bagging Predictors. 1996.
[8] KOMPRDOVÁ, 2012, op. cit.
[9] KLASCHKA, Jan; KOTRČ, Emil. Klasifikační a regresní lesy. 2004.
[10] KOMPRDOVÁ, 2012, op. cit.

prediktorů a nadále zvažuje větvení pouze na této podmnožině. Random Forest proto dávají v porovnání s ostatními metodami lepší výsledky a náhodný výběr také urychluje výpočty. Výsledkem RF jsou velké neprořezávané stromy. Na samotné kvalitě jednotlivých stromů totiž nezáleží, je zde snaha minimalizovat chybu celého lesa.[11]

1.2 Detekce anomálií

„Anomálie je pozorování, které se natolik odlišuje od ostatních pozorování, až vzbuzuje podezření, že bylo vytvořeno odlišným mechanismem.“ [12]

Anomálie mohou být tří druhů:

a) Bodové – případy, které jsou individuálně velmi odlišné od ostatních.
b) Kontextuální – případy, které jsou v datech běžné, ale když je bereme v kontextu, tak jsou anomální. Například v grafu závislosti teploty na ročním období je hodnota 0 °C zcela normální v lednu, ale v červenci je anomální.
c) Kolektivní – individuálně nejsou hodnocené jako anomálie, ale skupina, ve které se nacházejí, představuje anomální skupinu.

Anomálie buď označíme nebo všem prvkům přiřadíme skóre, které určuje míru jejich anomálnosti.[13]

Například při využití algoritmů pro detekci anomálií založených na vzdálenostech se využívá předpoklad, že na základě velké rozdílnosti anomálií od ostatních případů by v jejich okolí nemělo být mnoho sousedů. Při globálním pohledu si musíme zvolit, jakou vzdálenost považujeme za okolí a poté spočítáme počet případů, které se nacházejí v okolí jednotlivých bodů. Dále si musíme stanovit, jaký počet je mnoho sousedů. Dle této hodnoty rozhodneme, zda se jedná o anomálii. Pokud jsou data v různých oblastech různě hustá, dochází u globálních algoritmů k problému – jako anomálie mohou být označena i zjevně neanomální data. Abychom elegantně obešli tento problém, tak se často používají lokální algoritmy, založené na sousedství. Zvolíme hodnotu ***k***, která určuje, kolik nejbližších sousedů vybereme, a u každého bodu

[11] KLASCHKA, Jan; KOTRČ, Emil. Klasifikační a regresní lesy. 2004.
[12] HAWKINS, D.M. *Identification of Outliers*. 1980.
[13] CHANDOLA, Varun, BANERJEE, Arindam a KUMAR, Vipin. Anomaly detection. [online]. 2009.

zjistíme vzdálenost od jeho ***k*** nejbližších sousedů.[14] Z podobné myšlenky vychází metoda LOF popsaná v následující kapitole.

1.2.1 Local Outlier Factor (LOF)

Algoritmus LOF je založený na vzdálenostech a je lokální. To znamená, že zjišťuje vzdálenosti od ***k*** nejbližších sousedů. Výsledkem je míra anomálnosti jednotlivých prvků, přičemž čím vyšší je číselná hodnota, tím větší je to anomálie. Lze využít pouze pro data, která nejsou rozdělena do tříd. Základem tohoto algoritmu je vzorec:

$$OF_p = \frac{R_p}{\bar{R}_p}$$

Rovnice 1: Výpočet faktoru odlehlosti metodou LOF.

Velikost poloměru okolí prvku podělíme průměrnou velikostí okolí v okolí zkoumaného prvku.[15]

1.2.2 Class Outliers Mining: Distance-Based Approach (CODB)

Algoritmus CODB je založen na vzdálenostech. Tato metoda vyniká oproti ostatním tím, že s ní lze hledat anomálie i v datech, která jsou rozdělena do tříd. Díky tomu můžeme hledat i anomálie, které sice nejsou příliš vzdáleny od své vlastní třídy, ale mají blíže k třídě jiné viz obr. 2.[16] Při použití jiných metod musíme data buď rozdělit do skupin podle tříd nebo úplně přijít o tuto informaci, a to je u většiny klasifikovaných dat nemyslitelné.

Obr. 2: Anomálie daleko od vlastní třídy vs. blízko své třídě.

[14] BREUNIG, Markus M., KRIEGEL, Hans-Peter, NG, Raymond T. a SANDER, Jörg. LOF. [online]. 2000.
[15] VINTROVÁ, Vanda, VINTR, Tomáš, ŘEZANKOVÁ, Hana a ÚRADNÍČEK, Vladimír. Porovnání vybraných algoritmů pro ohodnocení odlehlosti vícerozměrných pozorování. [online]. 2014.
[16] NEZVALOVÁ, Leona, POPELÍNSKÝ, Lubomír, TORGO, Luis a VACULÍK, Karel. Class-Based Outlier Detection: Staying Zombies or Awaiting for Resurrection? 2015.

Class outlier factor se vypočítá součtem tří částí:

$$COF(t) = K \cdot PCL(t; K) + \alpha \cdot \frac{1}{Dev(t)} + \beta \cdot KDist(t)$$

Rovnice 2: Výpočet faktoru odlehlosti metodou CODB.

- Zvolený počet nejbližších sousedů **K** vynásobený pravděpodobností správného označení prvku do třídy, do které náleží, s ohledem na **K** nejbližších sousedů.
- Převrácená hodnota odlišnosti prvku vůči své třídě vynásobená konstantou.
- Hustota okolí prvku (součet vzdáleností ke **K** nejbližším sousedům) vynásobená konstantou.

Čím nižší hodnota COF, tím je to větší anomálie.

1.2.3 Random Forest – Outlier Explanation (RF-OEX)

RF-OEX je metoda vytvořená na Fakultě informatiky Masarykovy univerzity. Je nejnovější z metod pro detekci anomálií v datech rozdělených do tříd. Algoritmus je založen na metodě náhodných lesů. Pro definici podobnosti se totiž využívá klasifikace – příklady jsou si podobné v případě, že je strom v náhodném lese přiřadí do stejné třídy a odlišné v opačném případě. Informace o podobnosti se uchovává v matici blízkosti nabývající hodnoty od 0 do 1, kde hodnoty 1 pro prvky ***i*** a ***j*** nabývá, pokud všechny stromy klasifikovaly tyto dva prvky do stejné třídy.

Samotný faktor odlehlosti se vypočítá součtem tří částí:

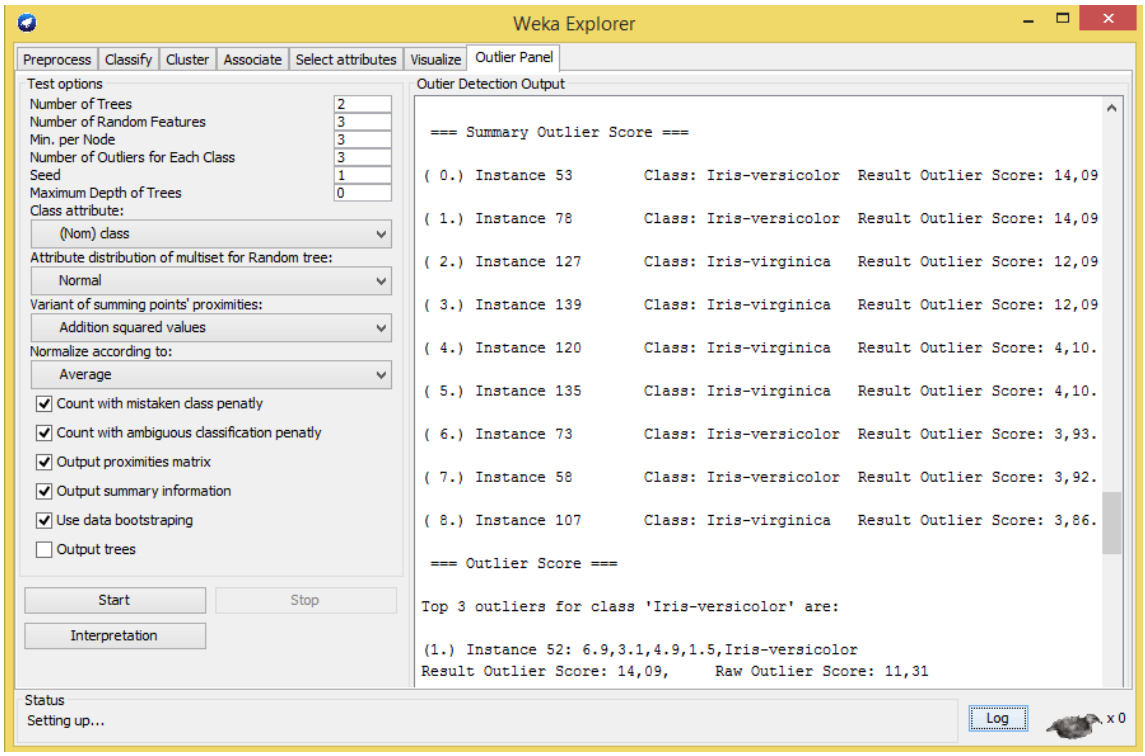
$$FO(p) = FO_1(p) + FO_2(p) + cFO_3(p)$$

Rovnice 3: Výpočet faktoru odlehlosti metodou RF-OEX.

- $FO_1(p)$ – Odlehlost prvku vůči své třídě – normalizovaná obrácená hodnota pro prvek ***p*** sečtené hodnoty z matice blízkosti pro prvky náležející do stejné třídy.
- $FO_2(p)$ – Faktor nesprávného zařazení do třídy – pravděpodobnost nesprávného zařazení prvku *p* do třídy vynásobená konstantou ***c***.
- $cFO_3(p)$ – Faktor odlehlosti nezávisle na třídě – pravděpodobnost odlišnosti od ideálního stavu vynásobená konstantou ***c***.

Čím vyšší hodnota faktoru odlehlosti, tím je to větší anomálie.[17]

RF-OEX je implementován v systému Weka, což je balík s algoritmy strojového učení v programovacím jazyku Java. Prostředí pro detekci anomálií si lze představit na obr. 3.



Obr. 3: Prostředí pro detekci anomálií RF-OEX.

[17] PEKARČÍKOVÁ, Zuzana. *Detekcia odľahlých bodov v klasifikovaných dátach* [online]. Diplomová práce.

2 Praktická část

V praktické části se zaměřím na hledání anomálních recenzí filmů v datové sadě Large Movie Review Dataset. Za anomálii pokládám pozitivní recenzi filmu podle hvězdiček, která má negativní prvky, či naopak. Tedy text recenze nekorespondující s ohodnocením hvězdičkami uživatelem. Převážně budu pracovat v jazyku R, na metodu ECODB použiji jazyk RapidMiner a metoda RF-OEX je implementována ve Wece napsané v Javě. Zmiňované metody lze použít na data, která mají třídy. Na hledání anomálií pouze v jedné třídě v R použiji lofactor a outliers.ranking z balíku DMwR a outlier z balíku outliers. Také použiji metody strojového učení decision trees z balíku rpart a Random Forest z balíku caret.

2.1 Datová sada

V této kapitole stručně popíši, čím se zabývá webová stránka imdb.com, co obsahuje a jak se člení používaná datová sada Large Movie Review Dataset.

2.1.1 IMDb

Užívaná datová sada s recenzemi filmů čerpá z volně přístupné on-line internetové databáze filmů IMDb přístupné z webové stránky imdb.com. Na této stránce se nacházejí informace o filmech, televizních pořadech, událostech a novinkách ze světa filmu a pracovnících z řad herců, režisérů, scénáristů a dalších. Původně anglická databáze vznikla v roce 1990 a v současné době ji vlastní americká společnost Amazon.com.[18]

Po založení účtu mohou uživatelé ohodnotit filmová díla. Hodnotí je pomocí hvězdiček od 1 do 10, kde 10 je nejlepší hodnocení uživatelem. Navíc mají možnost napsat recenzi. Ostatní návštěvníci si poté mohou tyto recenze přečíst viz obr. 4.

[18] *IMDb* [online]. 1990.

Obr. 4: Ukázka recenze v IMDb.

2.1.2 Large Movie Review Dataset

Datová sada Large Movie Review Dataset[19] využívá volně přístupné informace na webu imdb.com. Obsahuje 25 000 trénovacích a 25 000 testovacích recenzí rozdělených na polovinu podle počtu hvězdiček na negativní a pozitivní. Pozitivní jsou hodnoty 7–10, negativní 1–4. Dále je zde 50 000 nespecifikovaných dat určených pro učení bez učitele.

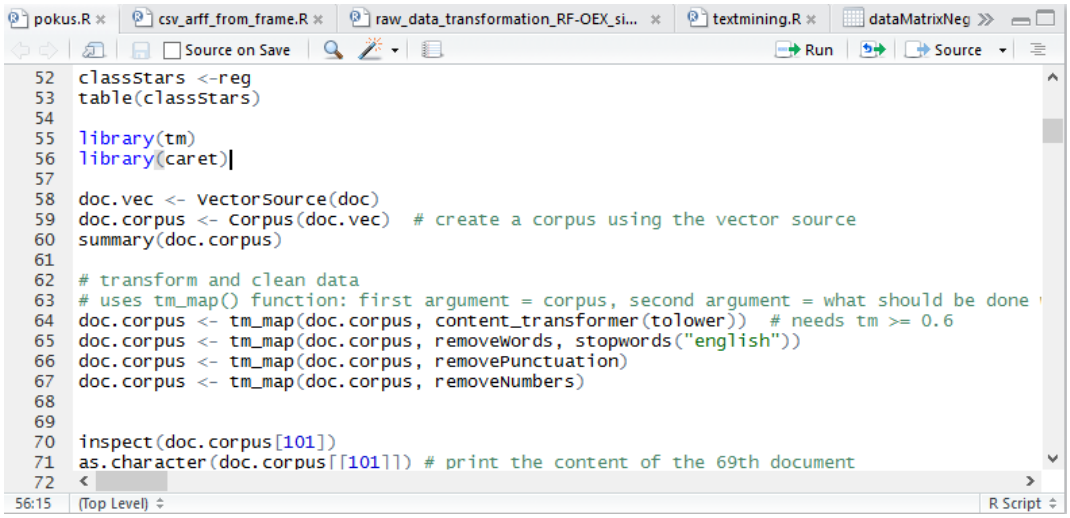
Data jsou členěna do složek train a test. Ty obsahují podsložky neg a pos. V těchto složkách se nacházejí textové soubory obsahující text recenze a jejich název je ve formátu [[id]_[rating].txt]. Kde id je jednoznačný identifikátor a rating je počet hvězdiček, které recenzent zvolil při psaní recenze. Složky train a test dále obsahují textové dokumenty s url odkazujícími na stránku idmb.com s recenzemi filmu, ke kterému přísluší daná recenze. Můžeme tedy identifikovat recenze k určitému filmu. K jednomu filmu se zde nachází maximálně 30 recenzí.[20]

[19] Ke stažení na: http://ai.stanford.edu/~amaas/data/sentiment/
[20] MAAS, Andrew; E. DALY, Raymond; PHAM, Peter T.; HUANG Dan; NG, Andrew Y.; POTTS, Christopher. Learning Word Vectors for Sentiment Analysis. [online]. 2011.

2.2 Předzpracování dat

Psaný text obsahuje spoustu nadbytečných informací, a proto jsem i já musela použitá data upravit do formátu srozumitelného pro počítač. Data jsem upravila pomocí programovacího jazyka R a použila jsem tyto metody:

- převod velkých písmen na malá
- odstranění slov, která se v daném jazyce vyskytují často a nenesou žádný význam, tzv. stopwords např. the, and, is, are apod.
- odstranění interpunkčních znamének
- odstranění číslic
- převod původních slov na kmeny slov



Obr. 5: Ukázka vytvořeného programu pro předzpracování textu.

Takto upravený dokument jsem uložila do tabulky, kde řádky jsou recenze a sloupce slova. Tabulka zaznamenává četnost jednotlivých slov v recenzích viz obr. 6. Celkový počet všech slov je přibližně 80 000. Algoritmy nejsou tak efektivní na datech s vysokým počtem dimenzí (sloupců) a slova, která se vyskytují menšinově, nemají významnou výpovědní hodnotu. Proto jsem počet proměnných snížila na 68 nejfrekventovanějších slov.

	act	actor	actual	also	back	bad	best	better	can
12502	0	0	0	1	0	0	0	0	0
12504	0	0	0	0	0	0	1	0	0
12505	0	0	1	0	0	0	0	0	1
12507	0	2	0	0	0	0	0	0	0
12508	0	0	0	0	0	0	0	0	2
12509	0	0	0	0	0	0	1	0	1

Obr. 6: Ukázka tabulky četnosti slov.

2.3 Klasifikace pozitivních a negativních recenzí

2.3.1 Rozhodovací stromy

Rozhodovacím stromům jsem nejdříve předložila data i s jednoznačným identifikátorem filmu, ke kterému daná recenze přísluší. Výsledky byly velice slibné:

```
  Confusion Matrix and Statistics

          Reference
Prediction  neg  pos
       neg 2882  264
       pos  868 3486

               Accuracy : 0.8491
                 95% CI : (0.8408, 0.8571)
```

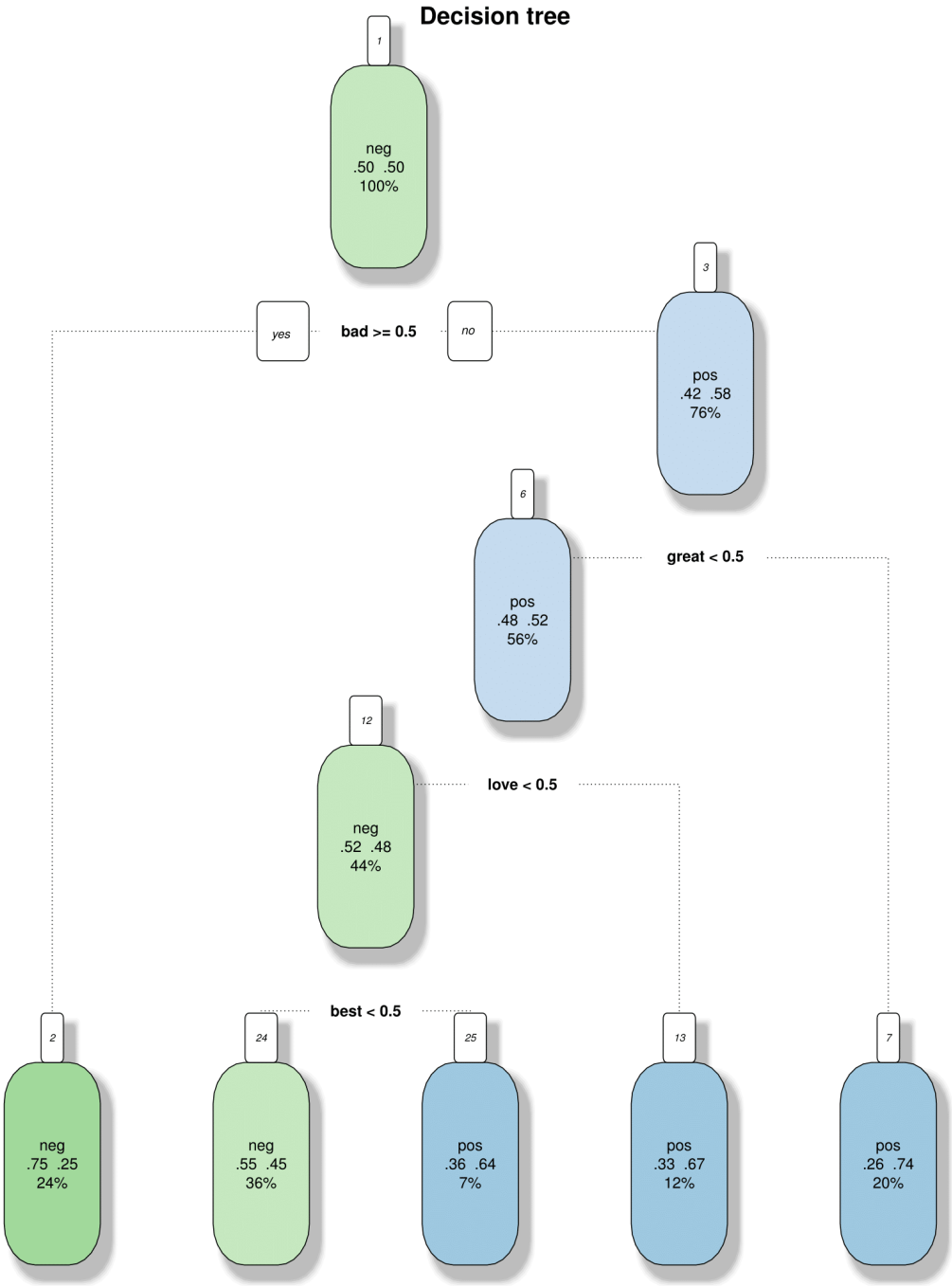
Tedy téměř 85% celková správnost. Druhá hodnota se týká intervalu spolehlivosti pro celkovou správnost. Po prozkoumání modelu tohoto stromu jsem si povšimla, že se o pozitivnosti, či negativnosti rozhoduje téměř výhradně podle ID filmu. Je pravda, že podle obsazených herců, režisérů atd. můžeme predikovat úspěšnost daného filmu. Nicméně pokud se tento film někomu nelíbí a napíše o něm negativní recenzi včetně hodnocení hvězdičkami, tak tento pak nezapadá do definice anomálie, jak jsem si ji zvolila. Nehledám vybočující recenze o brilantních filmech či naopak propadácích. Proto jsem ID filmu odstranila a dostala výsledky:

```
Confusion Matrix and Statistics

          Reference
Prediction  neg  pos
       neg 2712 1715
       pos  788 2035

               Accuracy : 0.6806
                 95% CI : (0.671, 0.6901)
```

Celková správnost se zhoršila téměř o 20 % na 68 %. Samotný strom ovšem vypadá uspokojivě:



Obr. 7: Strom pro klasifikaci filmových recenzí.

Ze zobrazení stromu lze vyvodit, jakým způsobem klasifikuje data. Nejprve zjistí, zda recenze obsahuje slovo bad a zhruba čtvrtinu dat rovnou klasifikuje jako negativní. U zbytku se rozhoduje podle slov great, love a best.

2.3.2 Náhodné lesy

Náhodným lesům jsem předkládala data již bez ID filmu. Jeho výsledky byly oproti rozhodovacím stromům lepší:

```
Confusion Matrix and Statistics

          Reference
Prediction  neg  pos
       neg 9984 5150
       pos 2516 7350

               Accuracy : 0.7133
                 95% CI : (0.704, 0.7225)
```

Z toho plyne, že mohu predikovat hodnocení z recenze s celkovou správností 71 %. Je dobře, že mohu úspěšně klasifikovat data, protože mohu opravdu detekovat anomálie, které jsem si stanovila. V datech je vzorec, který je může zařadit do pozitivní či negativní třídy, a já se zajímám o data, která do tohoto vzorce nezapadají. Kdyby nebylo možno predikovat hodnocení, nalezla bych jiný typ anomálií – recenze, které se výrazně vymykají normálu.

2.4 Detekce anomálií

2.4.1 RF-OEX

RF-OEX je dle mého názoru nejúspěšnější metoda pro hledání anomálií. Nalézám recenze, které by klidně mohly mít i opačné hodnocení. Ve dvaceti nejodlehlejších se nachází například:

******* Faktor Odlehlosti: 9,40

- pozitivní recenze s velmi špatným hodnocením herců

  Tsui Hark's visual artistry is at its peek in this movie. **Unfortunately the terrible acting by Ekin Cheng and especially Cecilia Cheung (I felt the urge to strangle her while watching this, it's that bad :) made it difficult to watch at times.**

  This movie is a real breakthrough in the visual department. When I first saw this, my jaw dropped repeatedly and I thought to myself that I've never seen

anything remotely like it but this is how it should be done in order to do full justice to the mythical world of Chinese historical kung-fu novels! Without a doubt this is one of the best-looking Chinese historical kung-fu epic ever made.

But alas, Tsui Hark hasn't improved much in the writing department, and the story and dialog are rather juvenile (his apparent obsession with the silly and overly-long depiction of the evil guys didn't help either). To make it worse, this movie is very badly cast. They decided to use the "hot" popular Hong Kong idols as lead characters, but unfortunately both Ekin Cheng and especially Cecilia Cheung are totally unsuited for historical kung-fu dramas because they lack the nobility and mystique that such characters are supposed to embody. Adam Cheng Siu-Chow and Brigitte Lin in the 1983 version are infinitely better.

I wish that someday Zhang Yi-Mou and Tsui Hark can join forces and produce a movie that has the visual artistry of Tsui but with the maturity and story-telling poetry of Zhang...

******* FO: 9,28

- pozitivní recenze, kde autor sám uvádí, že je to opravdu špatný film

Chaulk this one up to being a guilty pleasure, I knew it's a bad film. It has all the characteristics of one. Yet there's just something about it that makes me feel compelled to watch it from time to time (preferably with beer in hand). I'm even willing to overlook the absolutely horrid ending (which, I do have to say, I hate) I guess I like it because it has a fun atmosphere about it and some pretty cool kills.

********** FO: 8,22

- pozitivní recenze na opravdu špatný horor. Kdyby byl brán jako horor, tak by měl dostat nejhorší hodnocení, ale autor to bere jako zábavný pořad a tomu dal hodnocení nejlepší

> People are seeing it as a typical horror movie that is set out to scare us and prevent us from getting some sleep. Which if it was trying to do then it would deservedly get a 1/10.
>
> The general view on this movie is that it has bad acting, a simple script that a 10 year old could produce and that it cant be taken seriously and people are rating it low because of this.

********** FO: 10,46

- krátká pozitivní recenze poukazující na množství lidí žijících v chudobě

> Kurosawa is a proved humanitarian. This movie is totally about people living in poverty. You will see nothing but angry in this movie. It makes you feel bad but still worth. All those who's too comfortable with materialization should spend 2.5 hours with this movie.

******** FO: 9,68

- pozitivní recenze filmu, který má velmi nestabilní kvalitu vtipů a písniček

> The plot wanders around for a while but we are distracted by an unending string of jokes ranging from hilarious to dull. To break up the detached plot and jokes they gave us some silly musical sequences, which much like the jokes, range from entertaining to a quick trip to the fridge.

********** FO: 10,21

- pozitivní recenze na film, ve kterém mohou být souvislosti pro lidi matoucí a hloupé

> its inspired from Bad boys... but it has Indian Masala to it... people think it might be confusing and stupid...

******* FO: 9,57

- velmi krátká pozitivní recenze, která zmiňuje špatnou scénu

> I really liked this movie despite one scene that was pretty bad (the one when Samantha and Nick are flirting in the hotel).

*** FO: 6,11

- negativní recenze obhajující herecké kvality herce před živým publikem

> Arthur Askey's great skill as a comic was in the way he communicated with his public. His juvenile jokes, silly songs and daft dances went down well because he was able to engage folk and draw them into his off the wall world. A lack of a live audience was a distinct disadvantage to him, and he was never completely comfortable in films.

**** FO: 5,98

- negativní recenze, která rozebírá, že s lepším úvodem a závěrem by to byl vcelku dobrý film

> With a better beginning and conclusion, this weird story would be a good low budget slasher movie.

* FO: 6,57

- negativní recenze s ironickým úvodem

> If utterly facile, regressive, self-indulgent, anti-establishment, anti-civilisation juvenilia appeals to you, then this is the ideal film.

2.4.2 ECODB

Metoda na těchto datech celkově nevykazuje vysokou úspěšnost a je zde uvedena pro srovnání s metodou RF-OEX. Přesto jsem v prvních dvaceti nejanomálnějších recenzích nalezla:

**

- recenzi psanou přímo hercem, který ve filmu hrál nějakou vedlejší roli

> As time goes on, I am counting myself lucky that my name is in no way connected to this film.

****

- recenzi, kde autor pouze vyjmenovává názvy televizních pořadů – srovnává realitu a filmovou fikci

****

- kladnou recenzi poukazující na nepřiměřený věkový rozdíl filmových milenců

**

- několik ironických recenzí např.

> Oh just what I needed, another movie about 19th century England. Which is pretty much like regular England,only nobody's vandalising football stadiums.

Ve vybraných anomáliích se nenacházejí dlouhé recenze jako v LOF, ani příliš krátké.

2.4.3 Lofactor (DMwR)

Z důvodu zohlednění pouze jedné skupiny dat jsem v množině negativních dat nenalezla recenze směřující ke kladnému hodnocení, nýbrž recenze, které se výrazně vymykají normálu. Jedná se tedy o jiný typ anomálií než v předchozích příkladech. V prvních deseti nejanomálnějších recenzích s různými hodnotami parametru nejbližšího okolí (3 a 10) jsem objevila:

***

- recenzi na počítačovou hru. Byla negativní, ale použitá slovní zásoba byla výrazně odlišná od recenzí filmů

> Well, maybe the PC version of this game was impressive. Maybe. I just finished playing the PS2 version and it's pretty much a complete mess.

*

- recenzi s obsahem

> FAIL. I'd love to give this crap a 0. Yes, I registered just to rate this garbage. I want to go back in time

> and cut my wrist. Heres some copy and paste to take up 10 lines.

Tedy, že film je opravdu hrozný, zopakováno desetkrát, aby zaplnil místo.

****

- recenzi na udílení Oscarů, kde bylo nadměrně zmiňováno slovo best (Best Actor, Best Film…), což není pro negativní recenze typické

> There's no such thing as a " minor " Oscar and just because the award is for Best Animated Short or Best Costume Design they're as well deserved as Best Picture or Best Director .

***

- recenzi, kde jen v prvním odstavci je použito 14krát slovo like. V celé recenzi je to 34krát

> "It's **like** hard to **like** describe just how **like** exciting it is **like** to make a relationship **like** drama **like** with all the **like** pornographic scenes thrown **like** in for **like** good measure **like,** and to stir up like contro- **like** -versy and make us more **like** money and **like** stuff." - Ellen, the lost quote.

Další anomální negativní recenze převážně zdlouhavě vyprávěly děj filmu a vlastní hodnocení bylo velice krátké nebo skryté v sarkasmech.

V množině pozitivních recenzí nebyly výsledky většinou tak výrazné. Většina z celkem třinácti nejvíce anomálních recenzí s různými hodnotami parametru nejbližšího okolí (3 a 10) byla velmi dlouhá.Přesto se zde nachází nádherné anomálie v podobě:

****

- recenze, která uváděla mnoho záporů

> In addition to that, I want to reassure any reader of this that in spite of all the negative things I have just written that this is still mostly good...

```
********
```

- recenze na dle recenzenta tzv. trash film, tzn. odpad

```
I'm glad it allowed as I've heard the sequels are
equally outrageous.
```

```
********
```

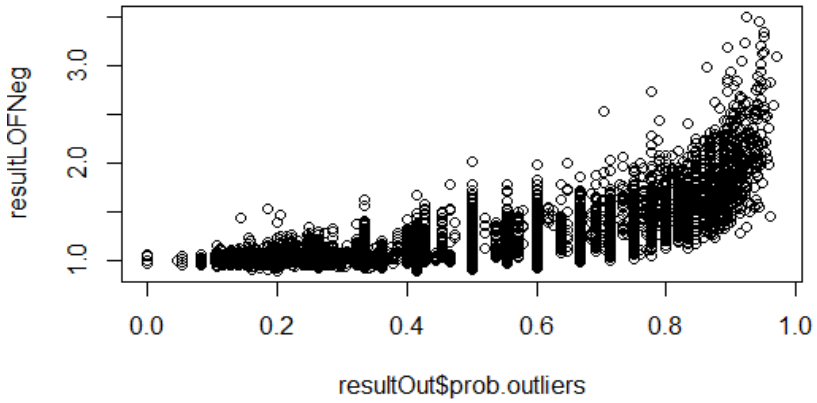
- recenze na dle recenzenta hloupý film

```
The plot intellect is about as light as feather down.
```

```
Pure classic silliness at its best.
```

Tato metoda většinou jako anomální označila recenze, které jsou velmi dlouhé. To se projevuje hlavně na množině pozitivních recenzí.

Pro detekci pouze na jedné množině dat jsem použila ještě další metody. Uvádím graf porovnání výsledků metody LOF a Outlier.ranking z balíku DMwR na množině negativních dat. Zobrazuje rozdílnost výsledků těchto metod. Kdyby určovaly míru anomálnosti shodně, body by byly v jedné přímce.



Obr. 8: Graf porovnání výsledků metod detekce anomálií.

Závěr

Při detekci anomálií v pozitivních a negativních recenzích se nečekaně jako nejspolehlivější ukázala metoda RF-OEX vyvinutá na Masarykově univerzitě. Zdaleka převýšila výsledek všech ostatních metod. Výsledné anomálie mají charakteristiku anomálií, jakou jsem si na počátku představovala – pozitivní se vyjadřují negativně a naopak. Jako anomální byly většinou označeny pozitivní recenze, které byly také více odlehlé. Nacházely se zde recenze, které kritizovaly některé aspekty filmů a některé se dokonce vyjádřily, že celý film je špatný. Negativní jsou často ironické, ale některé recenze filmy obhajují. Metoda RF-OEX se projevila jako nejúspěšnější metoda pro detekci anomálií na použitých datech.

Od detekce anomálií metodou LOF jsem mnoho neočekávala, protože hledá jiný typ anomálií, přesto se ukázala být dobrá. V negativních recenzích byly spíše recenze atypické a v pozitivních dlouhé, ale našly se zde tři recenze směřující k zápornému hodnocení – tedy anomálie, které jsem si představovala. Celkově největší zklamání přinesla metoda ECODB. Očekávala jsem, že bude mít nejlepší výsledky, jelikož zohledňuje data, která mají třídu, a navíc je všeobecně velmi rozšířená.

Do budoucna bych chtěla využít informace o morfologii a syntaxi textu, které mohou výrazně přispět ke komplexnímu zpracování textu. Také bych chtěla zohlednit index čitelnosti textu (readability), který hodnotí, zda je text dobře čitelný pro člověka. Dále se domnívám, že zpětné zařazení ID filmu s optimální nižší váhou by mohlo pozitivně ovlivnit jak klasifikaci, tak i detekci anomálií.

Seznam použitých pramenů

1. BERKA, Petr. *Dobývání znalostí z databází*. Praha: Academia; 2003. ISBN 8020010629.
2. BREIMAN, Leo: Bagging Predictors. *Machine Learning 24*, 1996, 123-140.
3. BREUNIG, Markus M.; KRIEGEL, Hans-Peter; NG, Raymond T.; SANDER, Jörg. LOF. *ACM SIGMOD Record* [online]. 2000, **29**(2), 93-104 [cit. 2017-02-12]. DOI: 10.1145/335191.335388. ISSN 01635808. Dostupné z: http://portal.acm.org/citation.cfm?doid=335191.335388
4. CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection. *ACM Computing Surveys* [online]. 2009, **41**(3), 1-58 . DOI: 10.1145/1541880.1541882. ISSN 03600300. Dostupné z: http://portal.acm.org/citation.cfm?doid=1541880.1541882
5. HAWKINS, D.M. *Identification of Outliers*. Dordrecht: Springer Netherlands, 1980. ISBN 9789401539944.
6. *IMDb* [online]. 1990 [cit. 2017-02-12]. Dostupné z: http://www.imdb.com
7. KLASCHKA, Jan; KOTRČ, Emil. Klasifikační a regresní lesy. *ROBUST.: sborník prací zimní školy JČMF: [Jednota českých matematiků a fyziků]*. Praha: Jednota českých matematiků a fyziků, 2004, **13.**, 177-184. ISSN 80-7015-972-3.
8. KOMPRDOVÁ, Klára. *Rozhodovací stromy a lesy*. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-785-7.
9. KULKARNI, Parag. *Reinforcement and systemic machine learning for decision making*. Hoboken, NJ: Wiley, c2012. ISBN 9780470919996.
10. MAAS, Andrew; E. DALY, Raymond; PHAM, Peter T.; HUANG Dan; NG, Andrew Y.; POTTS, Christopher. Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* [online]. Portland, Oregon, USA: Association for Computational Linguistics, 2011, 142-150.
    Dostupné z: http://www.aclweb.org/anthology/P11-1015

12. NEZVALOVÁ, Leona; POPELÍNSKÝ, Lubomír; TORGO, Luis; VACULÍK, Karel. Class-Based Outlier Detection: Staying Zombies or Awaiting for Resurrection? In Elisa Fromont, Tijl De Bie, Matthijs van Leeuwen. Advances in Intelligent Data Analysis XIV – 14th International Symposium, IDA 2015. Springer, 2015. s. 193-204, 12 s., ISBN 978-3-319-24464-8.

13. PEKARČÍKOVÁ, Zuzana. *Detekcia odľahlých bodov v klasifikovaných dátach* [online]. Brno, 2013. Diplomová práce. Masarykova univerzita. Vedoucí práce doc. RNDr. Lubomír Popelínský, Ph.D. [cit. 2017-02-12]. Dostupné z: http://is.muni.cz/th/207719/fi_m/diplomova_praca_pekarcikova.pdf

14. VINTROVÁ, Vanda; VINTR, Tomáš; ŘEZANKOVÁ, Hana; ÚRADNÍČEK, Vladimír. Porovnání vybraných algoritmů pro ohodnocení odlehlosti vícerozměrných pozorování. *Informační bulletin České statistické společnosti* [online]. 2014, **25**(1), 1-12 Dostupné z: http://www.statspol.cz/cs/wp- content/uploads/2014/4/IB_1_2014.pdf

Seznam obrázků

Seznam rovnic

Seznam symbolů, veličin a zkratek

SOČ	Středoškolská odborná činnost
IMDb	Internet Movie Database
LOF	Local Outlier Factor
CODB	Class Outliers Mining: Distance-Based Approach
COF	Class outlier factor
RF-OEX	Random Forest – Outlier Explanation
URL	Uniform Resource Locator
ID	jednoznačný identifikátor
FO	Faktor odlehlosti