Detekce tématu dokumentu

Jan Kandyba[1]

1 Úvod

Příspěvek se zabývá problematikou zpracování přirozeného jazyka, konkrétně klasifikací dokumentů. Klasifikace dokumentů na základě jejich obsahu je v dnešní době velice používaná, aniž bychom si to příliš uvědomovali – například ideálním příkladem je každodenní klasifikace spamových emailů. V tomto příspěvku byla testována efektivnost vybraných klasifikátorů na datech poskytnutých od Ústavu pro jazyk český Akademie věd ČR. Mezi vybrané algoritmy patří Support Vector Classification, lineární diskriminační analýza a aplikace neuronových sítí.

Data od ÚJČ tvoří ručně anotované rozhovory mezi dotazujícími a jazykovými poradci – dotazující zavolá na jazykovou poradnu s nějakým dotazem a poradci se jej snaží zodpovědět. Data byla anotována ručně, aby se vyhnulo problémům z automatického přepisu. Tento dotaz může být různého typu, existuje však několik častých témat, na která se dotazující ptají nejvíce. Tato práce navazuje na projekt NAKI, jehož cílem je vytvořit semi-automatizovaný systém, který usnadní vyhledávání a kategorizaci těchto dotazů. Strom tříd, do kterých poradci ÚJČ tyto dotazy klasifikovali je však poměrně rozsáhlý, a ne pro každou třídu zde bylo dost dat, aby se klasifikátor natrénoval správně. Bylo tedy nutno udělat shlukování menších tříd tak, aby je bylo možné dobře rozpoznat.

2 Experimentální část

Abychom mohli provést automatickou klasifikaci, je nutno data parametrizovat do číselné podoby. Způsobů, jakými lze text parametrizovat, existuje několik. V této práci byla provedena parametrizace pomocí 2 přístupů: TFIDF a doc2vec. TFIDF se skládá ze 2 částí – TF a IDF. TF označuje četnost slova v dokumentu (tato hodnota se často normalizuje na celkový počet slov v dokumentu) a IDF složka reprezentuje „důležitost“ slova – čím častěji se slovo vyskytuje napříč všemi dokumenty, tím méně je důležité, resp. dává menší informaci o tématu daného dokumentu. Doc2vec parametrizace vychází z word2vec parametrizace, která zohledňuje kontext daných slov. Doc2vec navíc přidává číslo dokumentu k dané parametrizaci slov. Bylo zjištěno, že doc2vec parametrizace není na tuto úlohu vhodná zejména z důvodu nízkého množství trénovacích dat. Pro finální výsledky byla tedy použita parametrizace TFIDF.

Po parametrizaci na číselné hodnoty byly provedeny experimenty s různými nastaveními parametrů metod a byly porovnány jejich výsledky. Po prvních experimentech se pohybovala úspěšnost klasifikace všech metod okolo 90 %. Vysoká úspěšnost klasifikace byla zapříčiněna přítomností značně nevyvážené třídy „balast“, která obsahovala řádově více dat, než ostatní třídy – při trénování se tak klasifikátor natrénoval především pro tuto třídu, ale nebyl schopen klasifikovat správně do ostatních tříd. Navíc většina dat pocházela právě z třídy balast, proto byla úspěšnost takto vysoká. Tento problém byl vyřešen tak, že při každém trénování byla náhodně načtena pouze část dat ze třídy balast – tak, aby data byla vzájemně vyvážená.

[1] student bakalářského studijního programu Inženýrská inormatika, obor Systémy pro identifikaci, bezpečnost a komunikaci, e-mail: kandybaj@students.zcu.cz

Po vyvážení tříd klesla úspěšnost klasifikace na necelých 70 %. Dále byly prováděny experimenty za účelem zlepšení přesnosti klasifikátorů. Před parametrizaci dat se přidala lemmatizace, která z původních slov vytvořila základní tvar – například slovo „lepší“ se lemmatizuje na „dobrý“. Lemmatizace mírně zlepšila výsledky, avšak nikterak zásadně. Bylo zjištěno, že největší vliv na správnou klasifikaci má správná volba příznaků – tedy volba parametrů TFIDF. Pokud jsou zvoleny nevhodně, podepíše se to i na úspěšnosti klasifikace.

Vybrané klasifikátory tvoří klasifikace pomocí podpůrných vektorů SVC, neuronové sítě (ANN) a lineární diskriminační analýza (LDA), která je též použita pro redukci dimenze příznaků a následně se klasifikovalo pomocí SVC.

SVC klasifikátor využívá podpůrných vektorů, tedy bodů ležících na okraji každé třídy. Jelikož třídy nejsou lineárně separabilní, využívá SVC jádrový trik, kterým transformuje body do vyššího prostoru, kde již lineárně separabilní jsou.

Princip LDA spočívá v nalezení lineární kombinace příznaků, které oddělují dané třídy, nebo které mají mezi sebou společné. LDA lze využít jako samostatný klasifikátor, nebo jej lze využít pro snížení dimenze příznaků před následnou klasifikací.

Předchozí metody využívají efektivních přístupů, avšak nenabízí velké možnosti volitelných parametrů. Neuronové sítě nabízí značnou svobodu v jejich návrhu – počet vrstev, počet neuronů v jednotlivých vrstvách, volba aktivační funkce atd. Výsledná síť obsahovala 3 vrstvy: vstupní, 1 skrytou s aktivační funkcí *sigmoid* a výstupní s funkcí *softmax*.

Přestože se s aktivační funkcí *sigmoid* pojí problém klesající gradient, osvědčila se jako nejefektivnější aktivační funkce. Propojení mezi vstupní a skrytou vrstvou nebylo úplné, jelikož se mezi ně vložila „vrstva“ v Pythonu označována jako *dropout*, která náhodně vyřadila určité množství neuronů ze vstupní vrstvy – skrytá vrstva tak při trénování vidí pouhou část dat. Neuronová síť se tak nemůže spoléhat na všechna data, což vede ke zlepšení klasifikace.

3 Závěr

Nejlepší výsledky klasifikace lze dosáhnout při správné volbě příznaků. Pokud jsou zvoleny nevhodně, bude špatná i úspěšnost klasifikace. Volba správných příznaků není jednoznačná, výsledky v následující tabulce tak nemusí být nejlepší možné, kterých lze při těchto datech dosáhnout.

Zlepšení výsledků klasifikace může zlepšit předzpracování dat, které nepřímo souvisí se zvolenou parametrizační metodou. Nejvyšší přesnosti dosáhl klasifikátor SVC.

Typ Metody	Průměrné skóre	Minimální skóre	Maximální skóre
SVC	66,88 %	63,37 %	71,04 %
LDA	63,22 %	59,16 %	66,09 %
LDA-SVC	62,85 %	58,66 %	65,59 %
ANN	64,18 %	59,65 %	66,83 %

Tabulka 1: Výsledky klasifikace pro různé metody; použitá metrika: přesnost

Literatura

Hořejš Jiří, Kubát Miroslav, Lažanský Jiří, Mařík Vladimír, Štěpánek Petr, Štěpánková Olga, Zdráhal Zdeněk. "Umělá inteligence", Akademie věd České republiky, 1993.

Adrian A. Hopgood. "Intelligent Systems for Engineers and Scientists", CRC PR INC., 2011.

Zbyněk Zajíc, Lucie Zajícová, Josef V. Psutka, Petr Salajka, Jaromír Novotný, Aleš Pražák, Luděk Müller. "First Insight into the Processing of the Language Consulting Center Data", SPECOM 2018, p. 778-787, Springer, 2018.