Posudek oponentky diplomové práce

Autor práce:	Bc. Adam Novotný
Název práce (EN):	Satellite data analysis using machine learning methods
Název práce (CZ):	Analýza satelitních dat pomocí metod strojového učení
Studijní program:	Aplikace přírodních věd
Studijní obor:	Aplikované matematicko-stochastické metody
Autorka posudku:	doc. RNDr. Lucie Kupková, Ph.D.
Pracoviště:	Katedra aplikované geoinformatiky a kartografie PřF UK Praha
	Albertov 6, 128 00 Praha 2

Cílem diplomové práce Bc. Adama Novotného bylo testovat přesnost klasifikátorů random forest, XGBoost a convolution neural network v prostředí Python 3.6.8 pro detekci vybraných zemědělských plodin z multitemporálních dat družic Sentinel-2 A a B v modelovém území Středočeského kraje. Pro trénování a validaci byla využita data Veřejného registru půdy (LPIS) Státního zemědělského intervenčního fondu. Analyzována byla časová řada 13 snímků Sentinel-2 z února až prosince 2019.

Zvolené téma je velmi aktuální jednak vzhledem k poptávce po praktickém využití družicových dat v různých oblastech včetně zemědělství (např. aktualizace LPIS), dále proto, že je třeba analyzovat narůstající množství volně dostupných dat a zjistit, jaký je jejich potenciál. V neposlední řadě toto téma přispívá ke studiu a rozvoji metod strojového učení, které patří nejen v dálkovém průzkumu Země ke slibným aktuálně velmi využívaným metodám.

Diplomová práce má 5 kapitol, její struktura není zcela klasická. Po úvodní části následuje popis dat (kapitola Data Description), poté rešerše, která je rozdělena do dvou kapitol (Machine Learning a Artificial Neural Networks). V kapitole 4 (Field Crop Classification) pokračuje částečně rešerše (Přehled přístupů ke klasifikaci plodin) a dále jsou popsána konkrétní využitá data a postup jednotlivých klasifikací. Zcela chybí kapitola Diskuse, kapitola Conclusion není číslována. Přestože práce není členěna standardně na kapitoly Rešerše, Data a metody, dá se říci, že je přehledná.

Práce je psána v anglickém jazyce poměrně slušné úrovně, ale občas studentovi chybí slovní zásoba a zkušenosti a některá vyjádření nejsou zcela srozumitelná (například vyjádření: „By choosing the proposed cloudiness levels, we expect the classifier to deal with missing pixel values"; „As can be seen in the figure 4.2, both datasets are quite imbalanced, which are dealt with by the classifiers differently"). Případně vyjádření nejsou zcela přesná („The selected features are sample mean and sample standard deviation for each time and each band" – „sample mean" – průměr čeho? Je myšlen parametr „průměrná spektrální odrazivost"? stejně tak v případě „sample standard deviation"). V některých případech nejsou použity správné termíny („This can be explainable by the visual similarity between the crops" – asi je myšlena „spektrální podobnost", termín „visual similarity" se nepoužívá; „the surrounding pixels which did not belong to the field were excluded" – je myšleno, že byly „odmaskovány"?).

Z formálního hlediska jsou největším nedostatkem nepovedené grafy (obrázky 1.3 a 4.2). Popisky neodpovídají jednotlivým sloupcům. A obrázek 4.2 budí dojem, že filtrovaný dataset obsahoval

v některých případech více dat než dataset původní. Skutečně tomu tak bylo? Z textu to není zcela pochopitelné, ale pokládám to za nepravděpodobné.

Velice podrobně jsou zpracovány rešeršní kapitoly Machine Learning a Artificial Neural Networks a též rešerše k problematice přístupů ke klasifikaci plodin. V postupu předzpracování a analýzy dat oceňuji, že byla náležitá pozornost i diskuse věnována problému oblačnosti v datech. Vhodné bylo také vytvoření filtrovaného datasetu, u něhož se z důvodu větší rozlohy/homogenity jednotlivých „snímků" předpokládá vyšší přesnost klasifikace. Za přínosné považuji i využití vah podílu jednotlivých tříd v datasetu pro snížení rozdílu mezi více a méně zastoupenými třídami a též testování využití rozšířeného datasetu v případě metody CNN.

K předzpracování a analýze dat mám na autora diplomové práce následující dotazy:

1) K analýze byly použity snímky ze 13 termínů v průběhu celého roku – od února do prosince 2019. Jste si jistý tím, že na podzim 2019 nebyla již na některých pozemcích pěstována jiná plodina (ozim) než od jara do začátku podzimu roku 2019? Pokud došlo k mísení spekter různých plodin ve stejných místech/pixelech v průběhu sledovaného období, tak to mohlo významně ovlivnit výsledky klasifikace. Naopak v datasetu chybí snímky z podzimu 2018, kdy již rostly ozimé plodiny, jejichž rozvoj a sklizeň proběhly v roce 2019. Doporučuji pokusně přepočítat výsledky pro dataset z období září 2018 – srpen 2019.

2) V datasetu chybí také snímek ze srpna, který může být důležitý právě pro detekci času sklizně mnoha plodin. Předpokládám, že chyběl, protože nebyla dostupná scéna, která by splňovala kritérium maximální oblačnosti. Je to tak?

3) Ve výčtu použitých pásem Sentinelu-2 je uvedeno, že byla využita 3 pásma s rozlišením 10 m a 6 pásem s rozlišením 20 m. Byla všechna pásma převzorkována do stejného prostorového rozlišení? Jaké prostorové rozlišení měl výsledný dataset?

4) Dataset získaný ze SZIF byl rozdělen na trénovací, validační a testovací v poměru 64:16:20. Proč byl zvolen právě tento poměr? V textu to není uvedeno. Dataset je obvykle dělen na trénovací a validační v poměru 1:1, nebo je podíl validačních dat větší.

5) Ohledně klasifikátorů random forest a XGBoost se chci zeptat, na základě čeho byly vybrány klasifikační příznaky. Není to zmíněno, ale předpokládám, že se jednalo o spektrální příznaky. Byly testovány také texturální příznaky a jejich přínos pro zvýšení přesnosti klasifikací?

I přes některé nejasnosti a možné nestandardní kroky v postupu lze konstatovat, že dosažné výsledky jsou zajímavé a celkové dosažené přesnosti klasifikací i přesnosti dosažené pro některé plodiny jsou dobré. Z diskuse, která mi byla dodána po odevzdání práce, vyplývá, že výsledky jsou srovnatelné a v některých případech i lepší než výsledky uváděné v literatuře. Pokud je to možné, doporučuji kapitolu Diskuse do práce a systému doplnit, například formou errata. I přes výše uvedené připomínky diplomant prokázal, že je schopný výborně pracovat s odbornou literaturou, zvládnout náročné nástroje pro analýzu rozsáhlých datových souborů a vyvodit adekvátní závěry. Diplomovou práci Adama Novotného doporučuji klasifikovat známkou C (dobře).

V Líbeznicích, 24.1.2021 doc. RNDr. Lucie Kupková, Ph.D.