Oponentní posudek diplomové práce

Autor diplomové práce: Bc. Tomáš Bouček; České vysoké učení technické v Praze, Fakulta stavební

Název diplomové práce: Testování způsobů klasifikace pokrytí území vybraných evropských oblastí

Diplomová práce Tomáše Boučka je zaměřena klasifikaci tříd krajinného pokryvu (land cover) na podkladě družicových dat Sentinel-2. Práce je členěna do deseti kapitol a její celkový rozsah činí 90 stran. Po teoretickém úvodu (kapitola 1) autor popisuje data a software použitý v rámci řešení práce (kapitoly 3 a 4). V dalších dvou kapitolách je pak poskytnut teoretický úvod do klasifikace obrazových dat, přičemž prostor je věnován také metodám validace výsledků klasifikace, a dále pak přehledu současných produktů a služeb mapování krajinného pokryvu v Evropě. Zbývající tři kapitoly jsou věnovány vlastnímu testování klasifikačních přístupů. Rozdělení práce do jednotlivých kapitol je na jednu stranu sice logické, avšak na druhou stranu je počet kapitol s ohledem na celkový rozsah práce poměrně vysoký. Některé kapitoly jsou tak velmi krátké, a bylo by proto rozumnější z nich vytvořit jednu souvislejší kapitolu – například kapitola 6 je de facto tvořena jen jednou tabulkou a má rozsah 1 strany. Obdobně pak kapitola 4 má rozsah pouhých dvou stran. Zpracování teoretické části práce pak demonstruje, že autor má přehled o teorii klasifikace družicových dat včetně metod pro vyhodnocení výsledků klasifikací. V tomto ohledu bych možná vytkl až příliš stručný (respektive obecný) popis charakteristik použitých klasifikátorů. Očekával bych, že budou zdůrazněny především ty jejich vlastnosti, které mají/mohou mít vliv na kvalitu získaných výsledků.

V teoretické části práce jsem narazil na několik chyb a nepřesností. Kapitole 3.1.2 autor například tvrdí, že data Sentinel-2 jsou na úrovni zpracování L1B atmosféricky korigována, což není pravda. O odstavec dále je ale již správně uvedeno, že atmosféricky korigována jsou data na úrovni L2A. V případě popisu klasifikace družicových dat autor dlouhou dobu mluví o „klasifikátoru strojového učení“ aniž by bylo zmíněno který klasifikátor z této široké skupiny algoritmů byl použit (tato informace zazní až po poměrně dlouhé době). Rovněž mi zde chybí zdůvodnění toho, proč byl pro řešení práce vybrán právě klasifikátor Random forest. Autor tento výběr zdůvodňuje „*přesností, s jakou je schopen klasifikaci provést*“. Výsledná přesnost klasifikace je ovšem dána celou řadou faktorů, kdy použití nějakého konkrétního klasifikátoru samo o sobě ještě nezaručuje dobrou přesnost klasifikace. Z hlediska používané terminologie jsem zaznamenal snahu používat za každou cenu české výrazy pro některé odborné termíny (např. „*tvůrčí přesnost*“), se kterou osobně příliš nesouhlasím. Výše uvedené body ovšem nepovažuji za zásadní.

Mnohem zásadnější připomínky mám k vlastnímu praktickému řešení práce. Primárním cílem práce je otestování a vzájemné srovnání výsledků klasifikace získaných pomocí klasifikátorů Maximum Likelihood a Random forest. Při přípravě družicových dat autor zmiňuje problematiku výskytu oblačnosti na použitých scénách. K přípravě masky oblačnosti je přitom používána vektorová maska (.gml), která je součástí datového kontejneru již od úrovně L1C. V případě použití dat na úrovni L2A je ale mnohem vhodnější použití tzv. SCL vrstvy, která vzniká při aplikaci atmosférické korekce, a která obsahuje (mimo jiné) i detekci výskytu oblačnosti. Bylo by proto vhodné uvést důvody, proč tato vrstva nebyla použita. Autor založil celou klasifikaci ve vybraných testovacích oblastech na pouhých dvou scénách: „*jarní*“ a „*letní*“, přičemž jejich výběr je zdůvodněn prakticky jen tím, že byly „*téměř bezoblačné*“. Takovýto výběr mi přijde jako nedostatečný, neboť k odlišení některých tříd krajinného pokryvu na podrobnějších úrovních je mnohdy potřeba pracovat s celou sekvencí scén zachycující všechny stavy, jichž může daný typ krajinného pokryvu v průběhu roku dosáhnout.

K vymezení zájmových tříd krajinného pokryvu autor využívá nomenklaturu Corine Land Cover (CLC). Tento výběr považuji za správný, avšak současně je potřeba mít na paměti i některá omezení, která jsou s použitím dat CLC spojena. Data CLC mají poměrně velkou minimální mapovací jednotku (25 ha), což autor ostatně sám připouští. Důsledkem toho je, že vymezené polygony jednotlivých tříd krajinného pokryvu mohou být poměrně heterogenní. Zajímalo by mě proto, zda před použitím těchto dat pro trénování klasifikátoru proběhla nějaká kontrola vnitřní homogenity použitých polygonů. Autor

práce zcela správně uvádí, že některé velké polygony je před použitím potřeba rozdělit na více menších, avšak toto „podělení“ by mělo probíhat tak, aby výsledné segmenty byly vnitřně co nejvíce homogenní (z hlediska krajinného pokryvu). Z textu práce chápu jen to, že byla řešena problematika tzv. mixed pixels na okrajích jednotlivých polygonů pomocí vnitřního bufferu. Aplikaci tohoto kroku pozitivně oceňuji. Autor dále řeší vzájemnou odlišitelnost jednotlivých tříd, a to zejména na podkladě rozptylogramů. Tento způsob je samozřejmě možný, ale v případě dat s větším počtem příznaků je poměrně neefektivní. Rád bych se proto zeptal, zda autor zná i jiné možnosti jakými lze separabilitu jednotlivých tříd analyzovat.

Autor dále provádí úpravu vstupních referenčních dat jejímž výsledkem je, že v množině trénovacích dat jsou zachovány pouze polygony těch tříd, které lze vzájemně dobře odlišit. To je na jednu stranu pochopitelné, avšak na druhou stranu není nijak ošetřeno (ani jakkoliv analyzováno) kam jsou zaklasifikovány plochy, které jsou v CLC datasetu zařazeny ke třídám, jenž nebyly použity pro trénování. Z tohoto pohledu je pak nutné vnímat i prezentované výsledky úspěšnosti klasifikací, které jsou vyjádřeny pomocí kontingenčních matic. Pokud jsou k trénování i následnému ověření výsledků vybrány cíleně jen plochy patřící ke třídám, které od sebe lze dobře odlišit, pak není divu, že celkové přesnosti vykazují velmi vysoké hodnoty. Pokud bychom ale udělali ověření výsledků například pomocí náhodného výběru zahrnujícího všechny plochy, pak si troufám tvrdit, že výsledky budou o poznání méně optimistické. Rád bych se na tomto místě také zeptal, zda pro ověření přesnosti byl použit alespoň nějaký jiný podvýběr referenčních ploch CLC než jaký byl použit pro trénování klasifikátoru (z textu to není zřejmé). V tomto ohledu totiž působí naprosto nepatřičně zejména část věnovaná výsledkům klasifikace pomocí klasifikátoru Random forest, která by se ve stručnosti dala interpretovat tak, že autor deklaruje schopnost klasifikovat všechny třídy krajinného pokryvu na všech úrovních podrobnosti s přesností 100 %. Napadá mě otázka, zda autor sám těmto výsledkům věří.

Závěr práce je věnován srovnání výsledků obou testovaných přístupů. Jeho rozsah je ovšem poměrně krátký a chybí mi zde jakákoliv hlubší diskuse nad tím, proč bylo dosaženo takových výsledků, jakých bylo dosaženo.

V celkovém souhrnu tedy mohu konstatovat, že autor hlavní cíl své práce splnil, avšak řadu kroků v rámci praktického řešení práce vnímám jako silně diskutabilní. Mnoho věcí mi pak přijde jako nedotažených (zejména analýza získaných výsledků). Vzhledem k těmto skutečnostem navrhuji práci k obhajobě s hodnocením „dobře“, avšak celkové hodnocení práce se bude odvíjet od kvality autorovy obhajoby.

RNDr. Jan Mišurec, PhD.

Praha, 15.6. 2020