

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Aplikace strojového učení při gamaspektrometrické identifikaci hornin
Jméno autora:	Martin Procházka
Typ práce:	bakalářská práce
Fakulta:	Fakulta jaderná a fyzikálně inženýrská (FJFI)
Katedra:	Katedra dozimetrie a aplikace ionizujícího záření
Oponent práce:	Mgr. Dana Majerová, Ph.D.
Pracoviště oponenta práce:	ČVUT v Praze, FJFI, KSI

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání	náročnější
Využití metod strojového učení pro identifikaci vzorků horniny pomocí gamaspektrometrického měření považuji za zajímavé a aktuální téma. Pro studenta bakalářského studia se jedná o náročnější téma vzhledem k tomu, že strojové učení musel zvládnout převážně samostudiem.	

Splnění zadání	splněno
Všechny body zadání považuji za splněné.	

Zvolený postup řešení	vhodný s výhradami
<p>Student pro gamaspektrometrické měření zvolil detektor na bázi CZT kvůli jeho menším rozměrům, avšak doba měření je dlouhá (4 hodiny mi pro předpokládané budoucí měření <i>in situ</i> připadají moc).</p> <p>Výhradu mám především k použití metody podpurných vektorů (support vector machine, SVM): v kapitole 4 (Strojové učení) je diskutována problematika jádrových funkcí a duálního problému, avšak v kapitole 5 student duální problém nevyužívá. Popis jádra RBF je v kontextu výsledků uveden zbytečně, neboť pro klasifikaci vybraných vzorků stačilo jádro lineární.</p>	

Odborná úroveň	průměrná
<p>První tři kapitoly práce a příloha A jsou odborně dobře zpracovány.</p> <p>Kapitola 4 vykazuje určité nedostatky, které pravděpodobně plynou z toho, že problematika strojového učení je pro studenta nová a nedokáže ji popsat s potřebným nadhledem.</p> <p>V kapitole 5 postrádám popis vstupních dat (resp. příznaků), která byla použita pro strojové učení; je zde pouze uveden počet provedených měření od zvolených vzorků hornin a zmíněno, že existuje 4096 příznaků. Dále by bylo vhodné, aby student připravil experimenty s metodami strojového učení pečlivěji: jaké parametry a v jakém rozsahu budou zkoumány funkcí <i>GridSearchCV</i>, zda je nutné předzpracovávat vstupní data (např. škálovat) atp. Použité metody strojového učení jsou popsány spíše z pohledu výsledků, takže některé implementační detaily (např. jak proběhl výběr parametrů pro rozhodovací strom) jsem musela zjišťovat v pythonovském zdrojovém kódu, který jsem si od autora práce vyžádala. V příloze B jsou uvedeny pouze skripty pro přípravu naměřených dat a v kap. 5 čtenář najde ukázky některých kódů týkajících se metod stroj. učení.</p> <p>Z kapitoly 6 (Diskuze) je patrné, že se student dobře orientuje v problematice měření včetně sestavení aparatury a že si uvědomuje některé nedostatky související se strojovým učením i problémy dané sběrem hornin (každá hornina byla nasbírána v jedné lokalitě, přitom obsah uranu se určitě liší napříč lokalitami).</p>	

Formální a jazyková úroveň	průměrná
<p>V práci se občas vyskytují typografické chyby (jednopísmenné předložky/spojky na konci řádku, velké písmeno uprostřed věty, české pomlčky v anglickém abstraktu, nedokončené uvozovky na str. 39, v sekci 4.3.2 by operace transpozice neměla být vysázena kurzívou).</p> <p>Formální chyba je u jednoho odkazu na str. 51: místo čísla kapitoly je uveden text „odkaz pls“.</p>	

Jazyková stránka: v textu se vyskytují překlepy (z těch závažnějších: akezruty namísto aleurity, příkrové namísto příkrovové, Kromtek namísto Kromek); občas přebývá nebo chybí slovo ve větě (na str. 29, 31, 33, 39 to ztěžuje pochopení některých odstavců), ojediněle se vyskytují i gramatické chyby (v kapitole 6).

Nejvíce mi vadí nejednotnost v psaní názvů metod strojového učení (občas česky, občas anglicky, různá velikost písmen) a nekonzistentní zkratky pro metodu k nejbližších sousedů (k -NN, k -NN nebo k -nn).

Výběr zdrojů, korektnost citací

průměrné

Student použil poměrně velké množství zdrojů. Všechny použité zdroje jsou v textu práce citovány.

Mám připomínky ke zdrojům, které souvisí se strojovým učení: většina z [23] až [27] se soustředí na určitý „detail“ strojového učení a neposkytuje všeobecný přehled – pro studenta-začátečníka v dané oblasti by byla vhodnější např. kniha s ISBN 978-1449369415. Namísto odkazu na obrázek (zdroj [25]) bylo vhodné najít konkrétní článek, kde je obrázek využit, např. <https://link.springer.com/content/pdf/10.1007/s10586-022-03564-9.pdf>.

Po formální stránce mám tyto výhrady: seznam použitých zdrojů není seřazen abecedně; online zdroje nemají uvedené datum citování; student nedodržel stejné pořadí/formát psaní jmen a příjmení autorů (porovnejte např. [1], [14] a [16]).

Další komentáře a hodnocení

Student provedl celkem 33 až 35 měření pro každou ze tří vybraných hornin. Tato měření byla časově náročná, a proto bylo potřeba vybrat takové metody strojového učení, kde není potřeba rozsáhlá trénovací množina dat. Naměřená data student zpracoval pomocí tří metod strojového učení (SVM, k -NN a rozhodovací strom).

Nejlepší výsledky dala metoda k -NN. Dle mého názoru by pro danou problematiku měl stačit rozhodovací strom, avšak bylo by potřeba prozkoumat větší množinu parametrů použitých pro *GridSearchCV*. Ještě poznamenám, že u metody SVM je vhodné začít lineárním jádrem a teprve pak případně zkusit jádro RBF (a pokusit se najít jeho vhodné parametry), čili postupovat od jednoduššího ke složitějšímu.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Student úspěšně sestavil aparaturu pro měření spektra záření gama emitovaného danou horninou. Návrh aparatury považuji za dobře promyšlený, jednotlivé komponenty jsou poměrně dostupné, takže by mohla být používána i v praxi při měření *in situ*. Slabinu práce vidím ve zpracování kapitol 4 a 5, které měly více odpovídat tomu, co student skutečně použil ve skriptech pro zpracování dat. Výsledky metod strojového učení bohužel nejsou tak uspokojivé, jak by se dalo čekat. To může být způsobeno malým množstvím dat použitých pro trénování modelů nebo velmi podobnými spektry fylitu a sodalitického trachytu, ale také tím, že student se dostatečně nevěnoval optimalizaci parametrů vybraných modelů (např. rozhodovacího stromu).

Otázka k obhajobě:

- Je nutné dělat cross validaci, když metoda nebyla úspěšná na celkové množině dat? (Týká se sekce 5.3.)

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **C - dobře**.

Datum: 15.8.2024

Podpis:

