Hodnocení vedoucího závěrečné práce

České vysoké učení technické v Praze **Fakulta informačních technologií**

Student: Bc. Andrej Palička
Vedoucí práce: RNDr. Petr Škoda, CSc.
Název práce: Semi-Supervised Learning of Millions of Astronomical Spectra
Obor: Znalostní inženýrství

Datum vytvoření: 5. 6. 2016

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
1. Náročnost a další komentář k zadání	***1=mimořádně náročné zadání,*** *2=náročnější zadání, 3=průměrně náročné zadání, 4=lehčí, ale ještě dostatečně náročné zadání, 5=nedostatečně náročné zadání*

Popis kritéria:
Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)

Komentář:
Zadání práce je velmi náročné, je částí projektu řešeného na Astronomickém ústavu AVČR v Ondřejově v rámci grantu MŠMT pro podporu aktivit COST "Applications of Artificial Inteligence in Astronomy". Student se musel detailně seznámit se základy astronomické spektroskopie a s nástroji a technologiemi Virtuální observatoře. Musel pochopit principy zpracování a redukce CCD dat při existenci velmi malého poměru signálu ku šumu a napsat adekvátní transformační procedury (s částečnou pomocí dostupných knihoven AstroPy). Dále musel zvládnout práci s masivně paralelním zpracováním několika miliónů spekter pomocí knihoven SPARK v prostředí Hadoop. A to jak na vlastní instalaci, tak později v klusteru METACENTRa CESNETu. Dále použil ondřejovský cloudový systémem VO-CLOUD, na jehož vývoji se již delší dobu podílí.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
2. Splnění zadání	***1=zadání splněno,*** *2=zadání splněno s menšími výhradami, 3=zadání splněno s většími výhradami, 4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Komentář:
Zadání práce bylo přes svoji náročnost splněno v rámci limitovaných dostupných výpočetních prostředků. Původní záměr analyzovat celý archiv přehlídky LAMOST čítající přes 4.5 miliónů spekter narazil na praktické limity použitého Hadoop clusteru Metacentra CESNETu (max cca milión souborů v jednom adresáři, který se nepodařilo technické podpoře vyřešit), proto byl celý výzkum realizován na miliónu objektů (navíc čínskou klasifikační linkou označenou jako hvězdy, což do jisté míry pomohlo najít naše cílové objekty snáze). Volitelný bod zadání - portování celé diplomové práce na systém VO-CLOUD, bohužel selhal, protože výkon a kapacita ondřejovských severů se po pokusné instalaci prostředí Hadoop ukázala příliš nedostatečná.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
3. Rozsah písemné zprávy	***1=splňuje požadavky,*** *2=splňuje požadavky s menšími výhradami, 3=splňuje požadavky s většími výhradami, 4=nesplňuje požadavky*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části.

Komentář:
Práce má 59 číslovaných stran textu, 4 strany příloh asi 15 stran povinných rejstříků a tiráže. Svým rozsahem plně splňuje požadavky na diplomovou práci. Většina textu o strojovém učení je nabita informacemi a vyžaduje pozorné čtení pro pochopení poměrně náročné problematiky. Jednotlivé části jsou celkem vyvážené a neobsahují zbytečné detaily navíc. Bohužel některé velmi komplikované postupy, jako získání a preprocessing spekter diplomant shrnul na několika málo stranách (např. kapitola 3.1), i když věcně vyčerpávajícím způsobem. Pro neodborníky v astronomii se proto může jevit většina práce jako triviální příprava dat. Podle mého názoru při podrobném popisu všech postupů a metod, které student musel použít by se velikost práce podstatně zvětšila.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*80 (B)*

Popis kritéria:
Posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře.

Komentář:

Práce je přehledně členěna vedle úvodu do problematiky a celkového závěru do pěti hlavních rozumně vyvážených částí. Po krátkém úvodu do astroinformatiky a vysvětlení motivace práce (proč bylo zvoleno částečně řízené učení), následuje první kapitola s přehledem metod částečně řízeneho učení. Druhá kapitola je pak krátkým vysvětlením funkce systému Hadoop a Spark. Třetí popisuje vlastní programy napsané diplomantem pro preprocessing i vlastní učení. Zde je zjevná chyba vnoření subkapitoly (číslování 3.2.0.x místo 3.2.1 - záměna \subsection a \subsubsection). Kapitola 4 pak krátce vysvětluje výsledek klíčového kroku - doménové adaptace spektra z jednoho dalekohledu do simulované podoby na druhém, což zajistí dostatek označkovaných vzorů pro učení i v případě, že v daném souboru nejsou.Bohužel, tento novátorský postup není dostatečně vysvětlen a tudíž může být opět považován za triviální. Kapitola 5 je pak popisem jednotlivých experimentů, jejich přesnosti (dokonce ve všech klasických parametrech jako F1 score, recall a precission) a časové náročnosti algoritmů v závislosti na jejich parametrech a velikosti dat. Obsahuje současně i diskusi vhodnosti a spolehlivosti metod pro dané případy. Text je podle mého názoru srozumitelný, ale poměrně stručný, pro větší srozumitelnost by bylo vhodné rozdělit jednotlivé kroky do samostatných kapitol a doprovodit obrázky. Bohužel nejdůležitější výsledek celé práce - seznam cca 350 identifikovaných objektů s emisemi je jen na přiloženém CD. V písemné části je jen extrémně malý vzorek výsledků (přesto zjevně demonstrující správnou funkcionalitu použitých metod). Bohužel velká stručnost je k velké škodě jinak výborné práce.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*95 (A)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 12/2014, článek 3.

Komentář:

Práce je příjemně čitelná, detailně členěná do podkapitol, ilustrovaná množstvím barevných obrázků a grafů. Experimenty jsou popsány řadou tabulek a doplněny ukázkami výsledných spekter. Velmi pěkným prvkem jsou formální zápisy algoritmů v jednotné grafické podobě. Terminologie je vysvětleny. Možná jen měl myslet na čtenáře, kteří nejsou odborníky na strojové učení a vysvětlit některé pojmy (např F1 score). I když je otázka kdo je cílovou skupinou čtenářů práce (pro informatiky jsou tyto pojmy samozřejmé). Práce je psaná pěknou srozumitelnou angličtinou a celý text působí dojmem profesionálního vědeckého dokumentu. Typografická úprava je korektní bez zjevných sirotků a vdov, rovnice jsou řádně číslovány a správně odkazovány v textu. Bohužel jsem kromě nadměrného vnoření sekce (viz výše) nalezl i několik překlepů (resoluting, str. 37, Ondejov's, str.24), nekorektní použití některých slov (např. we shall místo we will, concrete místo particular, či spectras), a nekonzistentní formu zápisu některých názvů - např. VO-CLOUD, Vocloud, VoCloud . To jsou ale jen drobné formální vady.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*95 (A)*

Popis kritéria:
Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:

Myslím, že analýza dostupných pramenů je dostačující, vzhledem k novosti problematiky je většina odkazů elektronických. Student zjevně provedl velmi rozsáhlý průzkum dostupné literatury i elektronických zdrojů. Všech 23 zdrojů bibliografie je v textu řádně odkázáno a autor jasně odlišuje informaci přejatou od vlastních tvrzení. Citace v časopisech a knihách jsou uváděny podle citačních standardů včetně citací elektronických zdrojů, kde je i uveden okamžik čtení či prohlídky daného zdroje.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*95 (A)*

Popis kritéria:
Vyjádřete se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:

Všechen použitý software je typu Open Source a celá práce je i k dispozici pod licencí GPL, i když to není z obsahu CD patrné Výsledkem celé práce by mělo být několik článků v recenzovaných časopisech v rámci řešení grantu MŠMT a dlouhodobější spolupráce na problematice v rámci dalších českých astroinformatických projektů. Získané výsledky jsou již nyní unikátní a předpokládáme je brzy publikovat. Jejich analýza nyní probíhá i po odevzdání práce. Po optimalizaci parametrů se objevily další zajímavé objekty. Chystáme se i na analýzu všech spekter přehlídky, což vyžaduje nalezení vhodné výpočetní platformy (patrně placený výkon na Amazonu)

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
8. Komentář o využitelnosti výsledků	

Popis kritéria:
Uveďte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uveďte možnosti využití výsledků ZP v praxi.

Komentář:
Výsledky práce jsou světově unikátní. Podle mých informací se jedná asi o první úspěšný pokus nasadit strojové učení při hledání konkrétního typu zajímavých objektů podle tvaru jejich spektrálních čar. Seznam nalezených kandidátů bude dále prověřován a objekty detailně analyzovány s následnou publikací v recenzovaném astronomickém časopisu.Velmi pro informatiku revoluční je transformace objektů z jednoho parametrického prostoru (vysokodisperzní spektrum z Ondřejova) do druhého (nízkodisperzní spektrum LAMOST) pomocí fyzikálních principů (konvoluce s instrumentálním profilem) aby byly získány označkované vzory pro strojové učení v situaci kdy jich není dostatek přímo v dané doméně. Toto je jeden z přístupů tzv. doménové adaptace - rapidně se rozvíjejícího oboru, na kterém intenzivně pracuje např. Amazon a Google. Některá nalezená spektra s podivnou emisí na místě kde být nemá, otevírají otázku co je to za exotické objekty, a jako takové potvrzují správnost astroinformatického přístupu při hledání nových astronomických cílů. Samostatně jsou použitelné i některé nově implementované algoritmy pro platformu Spark, které v distribuci chybějí. Ty má student i zveřejněny na GitHubu.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
9. Aktivita a samostatnost studenta v průběhu řešení	9a: ***1=výborná aktivita,*** *2=velmi dobrá aktivita,* *3=průměrná aktivita,* *4=slabší, ale ještě dostatečná aktivita,* *5=nedostatečná aktivita* 9b: ***1=výborná samostatnost,*** *2=velmi dobrá samostatnost,* *3=průměrná samostatnost,* *4=slabší, ale ještě dostatečná samostatnost,* *5=nedostatečná samostatnost*

Popis kritéria:
Posuďte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven (9a). Posuďte schopnost studenta samostatné tvůrčí práce (9b).

Komentář:
Student pracoval velmi samostatně, dohledal si drtivou většinu použité literatury sám a zcela sám si i naprogramoval veškeré programy a skripty. Vedl samostatně komunikaci s technickou podporou METACENTRa a díky jeho asertivitě se podařilo vyřešit většinu identifikovaných nedostatků jejich systému. Pravidelně se účastnil všech konzultací a reagoval operativně na e-mailovou korespondenci, kterou jsme konzultovali dílčí otázky. Spolupracoval i s dalšími členy grantového projektu a aktivně přicházel s řešeními problematiky i před vlastním formálním zadáním diplomové práce. Několikrát byl i na Ondřejově na setkání širšího grantového týmu a udělal si čas na setkání se světovými experty na astroinformatiku a strojové učení, Dr. Rafaelem de Souzou a Dr.Emille Ishidou z Brazílie v době jejich návštěvy Ondřejova. Je třeba zdůraznit, že A. Palička s naším týmem spolupracoval již od doby řešení bakalářské práce , kdy aktivně vystupoval na několika konferencích. Podílel se i na vývoji systému VO-CLOUD, kde odladil většinu preprocessingu později aplikovaného na Hadoopu. I po odevzdání diplomové práce dále vylepšoval algoritmy až k úplné spokojenosti.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*90 (A)*

Popis kritéria:
Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:
Student úspěšně aplikoval některé algoritmy částečně řízeného učení na velký objem spekter z čínské přehlídky oblohy LAMOST aby nalezl zajímavé objekty s emisními čarami, podobné těm, co známe z ondřejovského 2m dalekohledu. To se mu skutečně podařilo a výsledky jsou přesvědčivé s velkým publikačním potenciálem. Samostatně vyřešil velké množství problémů (včetně preprocessingu) a prakticky ozkoušel moderní technologie práce z velkými daty - prostředí Spark-Hadoop. Zároveň ukázal na slabiny tohoto řešení. Vytvořil dobře dokumentovaný software v jazyce Python, používající moderní knihovny AstroPy, SciPy či Pandas a publikoval jej na GitHubu. Prokázal schopnost spolupráce v širším týmu i samostatnost při řešení složitějších problémů. Bohužel toto velké úsilí nedostatečně popsal ve své práci. Pokud bych měl hodnotit písemnou část diplomové práce, klonil bych se k horšímu hodnocení (B). Ale s ohledem na náročnost a unikátní výsledky dávám A.

Podpis vedoucího práce: