Hodnocení vedoucího závěrečné práce

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

Student: Bc. Ksenia Shakurova
Vedoucí práce: RNDr. Petr Škoda, CSc.
Název práce: Unsupervised Learning and Outlier Detection in Large Archives of Astronomical Spectra
Obor: Znalostní inženýrství

Datum vytvoření: 3. 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". Studentka se musela detailně seznámit se základy astronomické spektroskopie a s nástroji a technologiemi Virtuální observatoře. Dále musela zvládnout práci s masivně paralelním zpracováním několika miliónů spekter pomocí knihoven SPARK v prostředí Hadoop klusteru CESNETu i práci s ondřejovským cloudovým systémem VO-CLOUD. Kromě rešerše dostupné (i velmi nové) literatury bylo třeba uskutečnit velké množství experimentů a vyhodnotit jejich výsledky na menších vzorcích z ondřejovského 2m Perkova dalekohledu, než bylo přistoupeno k vlastní analýze miliónu spekter projektu LAMOST. Vedle nutnosti psaní programu pro SPARK v prostředí CESNETu musela ještě studentka napsat mnoha analytických a visualizačních programů v Ipython notebooku s použitím knihoven SciPy a Matplotlib.

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. Můj původní záměr analyzovat celý archiv přehlídky LAMOST čítající přes 4.5 miliónů spekter narazil na praktické limity použitého Hadoop clusteru Metacentra CESNETu (problém max cca miliónu souborů v jednom adresáři - přesněji 2^20, který se nepodařilo technické podpoře vyřešit), dále limitace paměti všech 24 strojů v clusteru a zastaralé knihovny nedovolily realizovat některé nastíněné experimenty. Toto ale není rozhodně vina studentky, ale spíše ukazuje na přílišný optimismus představ o práci s Big Data konfrontovaný s finančně únosnou realitou (proto jsme museli zavrhnout možnost použití Amazonu). Nicméně i dílčí výsledky analýzy přehlídky LAMOST prokázaly, že použité metody opravdu fungují a nalezené outliery jsou opravdu exotická spektra, která si zasluhují následnou detailní vědeckou analýzu. Podrobné experimenty na malém souboru spekter z ondřejovského dalekohledu prokázaly použitelnost strojového učení např. pro identifikaci spekter nov a supernov, či grupování spekter stejné hvězdy přes rozdílné instrumentální artefakty.

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 64 číslovaných stran textu, 3 strany příloh asi 15 stran povinných rejstříků a tiráže. Svým rozsahem plně splňuje požadavky na diplomovou práci. Většina textu je nabita informacemi a vyžaduje pozorné čtení pro pochopení poměrně nestandardní problematiky. Popis experimentů je doplněn názornými obrázky a výsledky detailně diskutovány. Na závěr je pak uvedena dodatečná analýza vhodnosti daných algoritmů pro jednotlivé typy dat a z toho plynoucí doporučení pro další výzkum.

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

4. Věcná a logická úroveň práce	95 (A)

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 tří hlavních částí. První obsahuje minimální přehled o astronomické spektroskopii a popis použitých dat, podrobný popis moderních metod klastrovacích algoritmů, algoritmů redukce dimenze a algoritmů hledání outlierů, jakož i přehled metod pro verifikaci správnosti klastrování. Druhá, stručnější, část pak krátce shrnuje implementační detaily a použité knihovny a popis Hadoop instalace Metacentra. Nejrozsáhlejší, třetí část, pak popisuje řadu experimentů s jednotlivými metodami na obou datových archivech (ondřejovském a části LAMOST přehlídky) a podrobně diskutuje použitelnost a přesnost těchto metod pro dané případy. V rozumné míře ja pak uvedeno několik příkladů nalezených outlierů i ukázky klasterovaných dat, jasně prokazujících funkčnost zvolených metod.
Velmi cennou součástí práce jsou i stovky MB příloh na CD, kde je vedle kompletní sady zdrojových kódů i vyčerpávající seznam nalezených outlierů včetně jejich grafické vizualizace, tabulky přesnosti (confusion tables) i výstupy zpracování ze SPARKu.
Já osobně jako astronom bych ale ocenil i podrobnější popis předzpracování fyzikální informace (normalizace kontinua, regridding vlnových délek, oříznutí pozorovaného rozsahu na oblast dobrého signálu), což ale jistě bude v budoucím článku napraveno, a chápu že pro informatickou podstatu to není podstatné.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*90 (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 i přes vysoký podíl řádně anotovaných matematických výrazů, detailně členěná do podkapitol, ilustrovaná množstvím barevných obrázků a grafů. Experimenty jsou popsány řadou tabulek a doplněny ukázkami výsledných spekter. Terminologie je korektně použita a zkratky řádně v textu vysvětleny. Je psaná pěknou srozumitelnou angličtinou a nenalezl jsem viditelné překlepy, což svědčí o použití spell-checkeru. Celý text působí dojmem profesionálního vědeckého dokumentu. Typografická úprava je korektní bez zjevných sirotků a vdov, matematické symboly správně reflektují skloněné a stojaté písmo, rovnice jsou řádně číslovány a správně odkazovány v textu stejně jako dopředné i zpětné odkazy na subkapitoly. Bohužel, celkový dojem trochu kazí porušení tiskového zrcadla u některých dlouhých tabulek a včleněných obrázků. I když chápu autorčinu snahu o čitelnost a eliminaci zbytečně prázdných stránek s úzkou tabulkou na ležato, stejně jako zachycení maximální logiky při čitelnosti obrázků a grafů, je toto značně rušivé a v případě publikace v časopise těžko přijatelné řešení, za které jsem nucen snížit počet bodů.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*94 (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. Studentka zjevně provedla velmi rozsáhlý průzkum dostupné literatury i elektronických zdrojů. Všech 42 zdrojů bibliografie je v textu řádně odkázáno a autorka jasně odlišuje informaci přejatou od vlastních tvrzení.
Citace v časopisech a knihách jsou uváděny podle citačních standardů, i když někde není uveden vydavatel. . Drobný problém vidím i u citací elektronických zdrojů, kde není v drtivé většině uveden okamžik čtení či prohlídky daného zdroje (retrived on, accessed, Cited apod.), přestože podle mých zkušeností na to není závazný standard - jen jakási doporučení.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*100 (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, vzhledem ke speciálním řešením skriptů a kusu kódu, se ale nepředpokládá veřejné použití kódu bez vazby na konkrétní data vyžadující stejně spolupráci s autorkou. Hlavní použití vidím spíše v pokračování dalších experimentů s datovými soubory stále uloženými v METACENTRu a dalším zobecňování vizualizačních skriptů ve vazbě na VO-CLOUD. 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, ale hlavně se chystáme 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:

Zejména analýza přehlídky LAMOST odhalila v outlierech řadu velmi nezvyklých objektů s emisními čarami, které zjevně nejsou hvězdy (jak uvádí výstup automatické čínské klasifikační linky), ale spíše podivné kvazary či rafinované instrumentální artefakty. Práce je proto unikátní a její publikování je velmi žádoucí. Pokud je mi známo, je to zatím asi jediná komplexní studie použití strojového učení na hledání nových nezvyklých astronomických objektů ve velkých spektrálních přehlídkách. Další vylepšení algoritmů a zejména zpracování všech dnes dostupných spekter přehlídky LAMOST i s SDSS (každá celkem asi 5 mil. spekter) je unikátním vědeckým projektem s vysokým potenciálem dalšího návazného výzkumu.

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:

Studentka se na počátku řešení důkladně seznámila (i během pozorování s 2m dalekohledem v Ondřejově) s principy pořizování, redukce a analýzy astronomických spekter. Během celé doby řešení se pravidelně účastnila konzultací každé 2-3 týdny a pružně odpovídala na e-mailovou komunikaci, přes niž probíhala většina diskusí o dílčích výsledcích. Několikrát se zúčastnila i setkání s dalšími spolupracovníky na projektu. Během řešení projektu spolupracovala i se studentem Paličkou, s nímž řešila problémy s funkcí systému na METACENTRu i předzpracování dat. Samostatně komunikovala s technickou podporou METACENTRa. Podobně byla i v kontaktu s autorem systému VO-CLOUD J. Kozou.Většinu problémů s astronomickou problematikou si nejprve sama nastudovala a až pak se ptala mě na spíše menší detaily. Předvedla profesionální přístup k řešení náročné odborné problematiky.
Oceňuji, že se studentka i po odevzdání práce pokoušela dále metody optimalizovat a vyprodukovala další exotické objekty a jejich vizualizace pro detailní analýzu prováděnou v rámci grantu.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*96 (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:

Celá práce byla velmi náročná a vyžadovala porozumění astronomické spektroskopii, vyřešení spousty problémů s poměrně velkým objemem dat, orientaci v aktuálních metodách strojového učení i spolupráci s dalšími účastníky řešeného astroinformatického projektu. Výsledky jsou světově unikátní a studentka přislíbila další spolupráci i účast na psaní recenzované publikace.
Studentka prokázala, že se umí zorientovat v náročné odborné problematice a poradí si i s velmi náročným problémem i s programováním distribuovaných aplikací na systému SPARK-HADOOP. I úroveň zpracování práce považuji až na drobné nedostatky s formátováním tabulek za velmi vysokou. Proto dávám celkovou známku A.

Podpis vedoucího práce: