FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student:	Bc. Jakub Švehla
Oponent práce:	Ing. Karel Klouda, Ph.D.
Název práce:	Active Semi-supervised Clustering
Obor:	Znalostní inženýrství

Datum vytvoření: 4. 6. 2018

Hodnotící kritérium:	*Způsob hodnocení – následující škálou 1 až 4:*
1. 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 dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posuďte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Komentář:
Zadání bylo bezezbytku splněno. Bod 2) dokonce na více než 100 %, neb místo implementace alespoň tří metod jich pan Švehla implementoval deset. Splněnost bodu 3) "Propose directions for further improvements of reviewed methods." se posuzuje hůře, ale i ten považuji kapitolou 6 "Future work" za splněnou.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
2. Písemná část práce	*95 (A)*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále 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. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3. Posuďte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Komentář:
Práce je psána velice dobrou angličtinou. V tomto ohledu patří k nejlepším pracem, které jsem na FIT oponoval. Prakticky bezchybné jsou i zápisy matematických vzorců, včetně jejich často problematického začlenění do okolního textu. Autor v rámci rešeršní části musel vtěsnat do cca 20 stránek výtah z více jak 10 vědeckých článků a popsat všechny porovnávané metody a algoritmy. V kapitole dva se mu to podařilo velice dobře a myslím, že aspoň trochu poučený čtenář získá přehled o hlavních myšlenkách algoritmů a rozdílech mezi nimi. V kapitole 3 věnované aktivnímu učení, už se to podařilo méně a trochu mi přijde, že tam autorovi došel dech. Ztrácel jsem se například v používání pojmu neigborhood.
Kapitoly věnované experimentům a implementaci jsou místy trochu stručné, ale opět se dobře čtou a čtenář získá přehled o tom, jak byly experimenty navrženy a jaké byly výsledky. Sekcím 5.4.1, 5.4.2 a 5.4.3 s výsledky experimentů by slušelo, kdyby měly podobnou strukturu (podobně jako přiložené grafy).
Poněkud matoucím dojmem působí závěr (Conclusion), který začíná skoro stejně jako úvod a je v něm nejdříve používán přítomný čas a až poté minulý.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
3. Nepísemná část, přílohy	*90 (A)*

Popis kritéria:
Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce
– opakovatelnost experimentů

Komentář:

Pan Švehla implementoval všechny metody, které v práci popisuje, v jazyce Python jako balíčky s rozhraním doporučeným pro scikit-learn. U některých metod bylo možné použít implementace existující v jiných jazycích (Java, R), ale z dobrých důvodů využity nebyly. Repozitář se zdrojovými kódy je přehledný a vše lze snadno zreplikovat (ne u všeho jsem tak ale učinil). Je trochu škoda, že metody nejsou skoro okomentované, což poněkud snižuje jejich použitelnost.
Experimenty jsou také provedeny poctivě a transparentně. Pro lepší porovnatelnost chování algoritmů semi-supervizovaného clusteringu bych i v případě "number of constraints" použil jejich podíl ze všech možných "constraints", spíše než pevně dané počty (0, 100, 200, ..., 1000), jejichž velikost nereflektuje velikost zkoumaného datasetu. Zajímavé by bylo i simulovat chování při "noisy constraints", tedy (v případě experimentů úmyslně) špatné informaci o tom, že dvojice bodů (ne)patří do stejného clusteru.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Hodnocení výsledků, jejich využitelnost	*95 (A)*

Popis kritéria:
Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Komentář:

Hodnotit kvalitu clusterování je obtížná úloha, která hodně závisí na konkrétních cluterovaných datech, což také potvrzují experimenty v hodnocené práci. Závěr tedy není jednoznačný ani při použití poměrně ukázněných a dobře prozkoumaných šesti datasetů. Přesto je práce velmi dobrou vstupní branou do problematiky a mohla by tak pomoci všem, kteří se v ní potřebují zorientovat.

Hodnotící kritérium:	*Způsob hodnocení – nehodnotí se*
5. Otázky k obhajobě	

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:

1) Proč jste se nepokusil simulovat chování algoritmů i při "noisy constraints", jak popisuji výše?
2) Jaké algoritmy lze použít, pokud mám v datech nečíselné hodnoty (např. "nominal features")?

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Celkové hodnocení	*93 (A)*

Popis kritéria:
Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.

Text hodnocení:

Práci považuji za velmi kvalitní a výborně splňující úkoly ze zadání, proto ji navrhuji hodnotit známkou A (výborně).

Podpis oponenta práce: