Posudek oponenta habilitační práce

Univerzita	Univerzita Komenského v Bratislavě
Fakulta	Fakulta matematiky, fyziky a informatiky
Habilitační obor	Informatika
Uchazečka	Mgr. Bronislava Brejová, Ph.D.
Pracoviště	Katedra informatiky, Fakulta matematiky, fyziky a informatiky, Univerzita Komenského v Bratislavě
Název práce	Algorithms for Hidden Markov Models with Applications in Bioinformatics
Oponent	doc. RNDr. Tomáš Brázdil, Ph.D.
Pracoviště	Fakulta informatiky, Masarykova univerzita, Brno

Text posudku

Předložená práce shrnuje obsah sedmi vědeckých prací autorky Mgr. Bronislavy Brejové, Ph.D. v oblasti bioinformatiky. Tato oblast ležící na pomezí biologie, informatiky a aplikované matematiky se zabývá vývojem metod pro shromažďování a analýzu rozsáhlých souborů biologických dat. Její součástí je tedy i vývoj matematických modelů dat a jejich následná algoritmická analýza. Jedním z významných modelů, používaných právě v této oblasti, jsou tzv. skryté Markovovy modely (dále HMM), které tvoří jednotící prvek většiny prací v předloženém souboru.

HMM jsou významným modelem v oblasti rozpoznávání vzorů, který je založen na teorii stochastických procesů. Jsou vhodné zejména pro statistické zpracování složitých sekvenčních dat. Proto jejich aplikační doména zasahuje zejména do lingvistiky, zpracování řeči a také do bioinformatiky. Zde se ukazuje být vhodným prostředkem pro modelování sekvencí DNA a proteinů, na které se autorka zaměřuje ve většině svých prací. Její práce vykazují rysy čistě teoretického výzkumu z oblasti algoritmické složitosti a zároveň aplikovaného výzkumu na pomezí biologie.

Text práce je rozčleněn do pěti sekcí. Po velmi krátkém úvodu, který shrnuje obsah práce, následuje druhá sekce, která zavádí základní pojmy z oblasti HMM a shrnuje jejich vztah k bioinformatice. Tato část je psána velmi srozumitelným způsobem s přiměřenou úrovní matematických detailů. Pouze krátká zmínka o conditional random fields působí mírně nepatřičným dojmem, pro pochopení principů CRF je příliš stručná a z hlediska dalšího

textu nepodstatná. Popis aplikací HMM v bioinformatice je také velmi dobře čitelný a srozumitelný i pro neodborníka v dané oblasti. Zde jsou formulovány základní problémy, které je možné úspěšně řešit pomocí HMM. Zejména se jedná o hledání genů (gene finding), predikci topologie proteinů a reprezentaci skupin proteinů pomocí tzv. profilových HMM.

V dalších dvou sekcích, kterým se budu podrobněji věnovat níže, autorka popisuje vlastní vědecký přínos ve dvou podoblastech HMM: algoritmy pro inferenci v HMM a rozšíření stávajících algoritmů o vnější informaci (extrinsic information). Text je standardně zakončen krátkým závěrem.

Jak jsem již předeslal výše, třetí sekce se věnuje přínosu autorky v oblasti inference v HMM. V daném souboru se této oblasti věnují čtyři práce. První z nich [Brejová et al., 2007] se zabývá složitostí problému nejpravděpodobnější anotace (the most probable annotation) sekvence pomocí dané sady značek. Již dříve bylo dokázáno, že tento problém je NP-těžký. Práce autorky ukazuje, že tento problém je NP-těžký i pro fixní HMM (o 34 stavech a 2 značkách) a nelze proto očekávat existenci efektivního algoritmu pro malé fixní HMM obvyklé v dané aplikační oblasti. Článek [Brejová et al., 2007] dále prezentuje rozšíření klasického Viterbiho algoritmu a definuje dostačující podmínky, za nichž tento algoritmus efektivně řeší problém nejpravděpodobnější anotace. Jedná se o netriviální výsledky, které bude možné v budoucnu dále rozvíjet.

Další část třetí sekce je věnována problému přibližného nalezení hranic mezi různě označkovanými úseky dané sekvence. Zde se opět ukazuje, že základní problém, formulovaný Brownem a Truszkowským (2010), je NP-těžký. Autorka diskutuje různé verze tohoto problému a společně se spoluautory ukazuje jejich NP-těžkost [Nánási et al., 2012]. Nicméně v závěru formuluje vlastní dobře motivovanou verzi, která připouští efektivní řešení [Nánási et al., 2010]. Na této části mne zaráží pouze to, že přiložený článek [Nánási et al., 2012] dosud nebyl publikován v recenzovaném médiu. Přiložena byla pouze nerecenzovaná verze z arXiv.org, což mi nepřipadá pro tento typ kolekce vhodné. Poslední článek [Šrámek et al. 2007] prezentovaný ve třetí sekci se zabývá online modifikací Viterbiho algoritmu, který má prokazatelně menší paměťové nároky. Přesněji řečeno, [Šrámek et al. 2007] formálně dokazuje, že v případě dvoustavového HMM je očekávaná prostorová složitost online algoritmu vzhledem k délce sekvence pouze logaritmická oproti lineární složitosti standardního algoritmu. Přestože je tento výsledek omezen pouze na dvoustavový případ, jedná se o zajímavé použití exaktních technik teorie pravděpodobnosti v oblasti, kde tento typ důkazů není příliš obvyklý. V plné obecnosti je efektivita online algoritmu demonstrována pouze empiricky s velmi dobrým výsledkem.

Čtvrtá sekce se věnuje problematice rozšíření stávajících algoritmů pro inferenci a učení HMM o další dodatečné informace. Tento přístup byl samozřejmě hojně zkoumán nejen v souvislosti s aplikacemi HMM v bioinformatice, ale obecně v teorii strojového učení. Autorka nejprve shrnuje některá rozšíření HMM o vnější informaci, poté se věnuje vlastnímu přínosu. Ten spočívá zejména ve vývoji tzv. poradců (advisors), s jejichž pomocí jsou poté upravovány pravděpodobnosti dané HMM modelem. Tyto techniky jsou implementovány v programu ExonHunter, který je vyvíjen týmem autorky a který byl úspěšně aplikován v projektech zabývajících se dekódováním genomu (Schistosoma japonicum, Marssonina brunnea, Echinococcus granulosus). Dále se autorka zabývá složitostí hledání genů s pomocí externích nápověd (hints). Zde opět diskutuje různé verze společně se složitostí příslušných problémů inference v HMM s nápovědami. Zejména se jedná o článek [Kucharík et al., 2011], který se zabývá nápovědami v podobě dané množiny podřetězců anotací, a o článek [Kováč et al., 2009], v němž jsou nápovědy reprezentovány pomocí grafů. Zde mne opět zaráží netriviální množství prostoru, který je věnován článku [Kucharík et al., 2011], který byl prezentován pouze na regionální konferenci. Na druhou stranu se domnívám, že článku [Brejová et al., 2009], který byl publikován v časopise s vysokým impaktním faktorem, je věnováno příliš málo prostoru.

Celkově je práce založena na sedmi publikacích. Z toho dvě byly publikovány v časopisech s relativně vysokým impaktním faktorem (vzhledem k daným oblastem), tři byly publikovány ve sbornících mezinárodních konferencí (rank B dle Core), jedna na národní úrovni a jedna pouze v nerecenzované verzi. Není mi jasné, proč byly poslední dvě jmenované zařazeny do souboru. Nicméně je zřejmé, že Mgr. Bronislava Brejová, Ph.D. prokazuje schopnost vést kvalitní interdisciplinární výzkum v bioinformatice se zajímavými a prakticky velmi užitečnými výsledky. Kromě předložené práce o tom svědčí zejména další výsledky publikované ve špičkových časopisech s vysokým impaktním faktorem (např. Nature, Genome Research, Bioinformatics).

Závěr

Domnívám se, že práce *splňuje* požadavky standardně kladené na habilitační práce v oboru Informatika.

12. ledna 2014 Tomáš Brázdil