Oponentský posudek diplomové práce Andrei Luterové

Klasifikační stromy

Tomáš Fürst

Předložená diplomová práce zkoumá možnosti použití klasifikačních stromů pro predikci výsledků biopsie prostaty u pacientů, kteří byli na preventivním vyšetření v olomoucké Fakultní nemocnici v letech 2006–2012. Téma je součástí rozsáhlejší spolupráce s Urologickou klinikou FNOL. Téma je rozhodně zajímavé, odborná literatura je plná různých statistických přístupů k predikci karcinomu prostaty. Andrea studuje učitelství biologie a matematiky a zaslouží si pochvalu za výběr tématu, které je obtížné, pro ni nové jak fakticky tak matematicky, ale zároveň pěkně zasazené mezi její dva studijní obory.

Celkově soudím, že Andrea udělala hodně pečlivé práce. Jistě se u toho naučila pracovat s klasifikačními stromy v prostředí softwaru Statistica, pochopila mnoho z matematického pozadí této metody a dozvěděla se něco o rakovině prostaty, její diagnostice, léčbě a prekursorech. Tím je velká část smyslu diplomové práce naplněna.

Zpráva, kterou Andrea o tomto svém úsilí podala v podobě předložené diplomové práce, by však mohla být o něco lepší. Podrobnější připomínky budou následovat, v tomto odstavci se pokusím shrnout jen svůj celkový dojem: Práce je příliš obsáhlá, zbytečně a občas nešťastně se pouští do velkých podrobností, kde to není třeba. Samotné těžiště práce – konstrukce a interpretace klasifikačních stromů pro danou datovou sadu – začíná až na straně 60. Z úvodu a závěru není úplně jasné, co, jak a proč se udělalo, k čemu je to dobré a co z toho plyne. Práci by výrazně prospělo celkové zaostření. Je to trochu škoda, zejména proto, že velmi jednoduchý editorský zásah, by celé dílo pozvedl o třídu výš.

Následuje seznam připomínek, na které není třeba reagovat u obhajoby:

1. Tisková kvalita obrázků (logem UP na první straně počínaje) je dost hrozná. Jedním z cílů diplomové práce je i naučit se napsat dobře vypadající odborný text. Softwaru je na to dost.

2. Anglický abstrakt je sice chvályhodný, ale bohužel téměř nesrozumitelný.

3. Prohlášení na straně 4 bohužel deklaruje opačnou implikaci, než je třeba.

4. Úvod je poměrně nejasný, kdybych celou problematiku diagnostiky karcinomu prostaty detailně neznal, měl bych potíže s porozuměním. Je třeba buďto odkázat na nějaké guidelines, nebo skutečně celý postup pořádně popsat. Já bych volil druhou variantu. Uvědomte si, že nespecialista neví, co je punkce, proč se dělá, proč se dělá opakovaně, co se s takto získaným biologickým materiálem dělá dál, atd. Z vašeho textu to nemůže pochopit. Specialista to zase ví dobře a takovýto text nepotřebuje. Z celé práce není jasné, jak se stanoví výsledek biopsie a co to znamená. To je ovšem proměnná, kterou se snažíte predikovat!

5. Celá práce by potřebovala jazykovou korekturu.

6. Místy není zřejmý charakter vašeho textu. V posledním odstavci na straně 8 se pouštíte do spekulací, které jsou předmětem intensivních debat. Není jasné, jestli presentujete fakta (potom chybí reference) nebo svůj názor (potom chybí evidence). Do této debaty bych se určitě nepouštěl v rámci diplomové práce na oboru matematika.

7. Stejně tak první odstavec na straně 9 obsahuje problematická a nepodložená (každopádně necitovaná) tvrzení. Jsou to vaše názory? Jsou podloženy výsledky nějakých studií? Tyto pasáže textu vysloveně škodí a stačilo je prostě vypustit. Co jsou vůbec zdroje celé kapitoly 1?

8. Hned první série obrázků na straně 10 má nevyhovující kvalitu, nejasné popisky a je na ně divně odkázáno v textu. Obrázek 1.3 je potom úplně nesrozumitelný. Je zvláštní, že má každý obrázek dva popisky, jeden nahoře a jeden dole. Hned v prvním případě (obr. 1.1) se potom oba popisky dost podstatně míjejí.

9. Do kapitol 1.1, 1.2 a 2.3 bych se vůbec nepouštěl. Není to práce z urologie. Je mnoho výborných a důvěryhodných zdrojů, kam lze odkázat (wikipedia, European Guidelines, ...). Tyto odkazy v práci bohužel chybějí, místo nich je zařazen polo-odborný text nejasného původu a nevalné kvality, který stejně není v dalších částech práce nijak využit. Doporučuji podívat se na heslo "Prostate-Specific Antigen" na (anglické!) wikipedii.

10. Musím přiznat, že definici dat jsem za svůj dlouhý život ještě neviděl. Poprvé u vás v kapitole 2. Myslím, že práce trochu trpí přejímáním z nešťastných českých zdrojů, které stále něco formálně definují a dělí. Nesouhlasím s dělením dat na kvalitativní a kvantitativní. Každému znaku jde přiřadit číselná hodnota.

11. Je dobré si ověřit, že v klinické praxi (stejně jako téměř kdekoliv jinde) nemají data typicky normální rozdělení. Histogramy na straně 25 ale moc porozumění nepřispívají. Chce to jemnější stratifikaci a lepší popisky. Obrázek vpravo nahoře je vysloveně komický.

12. Velmi mě matou názvy kapitol. Skoro všechny názvy jsou buďto nic neříkající nebo přímo zavádějící. V kapitole "popisná statistika" se testují hypotézy, samotný hlavní výsledek – konstrukce klasifikačního stromu – je v kapitole "tvrzení a hypotézy". Některé kapitoly jsou nelogicky členěny. Opět, jednoduchý editorský zásah by práci výrazně prospěl.

13. Celkově mám problém s logikou textu. Například bych očekával, že napřed uvidím tabulku 9 s jasným popisem toho, proč a jak byla některá data odstraněna. Potom bych chtěl vidět popisné statistiky všech souborů a potom nějakou souhrnnou zprávu o jejich odlišnostech. Zde je příliš mnoho informací poněkud chaoticky rozmístěno. Mnohé informace se navíc duplikují (text i obrázek), všechny čtyři datové soubory jsou velmi podobné, takže některou informaci dostanu celkem osmkrát. Zde by opět méně bylo více.

14. Text kolem strany 35 asi zase trpí přejímáním českých zdrojů. Je to narativ, který má občas smysl a občas ne. Věta "Ve vědě, stejně jako v životě, platí princip efektivity, tedy snaha dosáhnout s minimálními výstupy maximálně možného efektu." je filosoficky zabarvený nesmysl. Věta "Do rovnice zahrnujeme pouze takové proměnné, o nichž víme z teorie nebo empirických zobecnění vyplívajících z analýzy jiných autorů, že jsou pro daný problém relevantní" (citováno včetně hrubé chyby) je podobný nesmysl. Co jsou prosím empirická zobecnění?

15. Ilustrační příklad na straně 43 je dobrý, bez něj jsem konstrukci klasifikačních stromů z textu chápal málo. Na druhou stranu, podrobnost zpracování příkladu je příliš velká. Jakmile je jasný princip, stačí uvést výsledek, není třeba presentovat výpočet se všemi algebraickými detaily (navíc mnohokrát stejný).

U obhajoby bych rád diskutoval o následujících tématech:

1. Z abstraktu, později ani úvodu a nakonec ani celé práce mi vlastně není jasné, co přesně je cílem vaší analýzy. Mohla byste se prosím pokusit mi to vysvětlit alespoň u obhajoby? Kdo a jak může výsledek vaší práce použít v klinické praxi?

2. Úvodní text kapitoly 3 je dobrý a navíc je velmi stěžejní, protože poskytuje zdůvodnění užití klasifikačních stromů. Všechny body na straně 34 by ale zasloužily daleko více pozornosti (na rozdíl od biochemie PSA). U obhajoby bych se rád dozvěděl více. Co říká bod 2? Jak může být nesplněn? Bod 3 podle mne nelze splnit. Proč bych nemohl vysvětlovat výšku člověka velikostí nohy a délkou femuru (korelované prediktory) pomocí lineární regrese? Co mám teda dělat, když taková data mám? Bodu 4 potom nerozumím a zdá se mi, že protiřečí bodu 3. Co znamená bod 5? Bod 6 asi vůbec není předpoklad, to je spíš tvrzení nebo pozorování.

3. U klasifikačních stromů bych rád viděl nějakou míru performance. Na trénovací množině jsem schopen nafitovat strom tak, že v každém listu budou jen uzly se stejnou hodnotou predikované proměnné. Tento over-fitting zřejmě není cílem (dá se udělat i daleko jednodušeji). Cílem je zřejmě vystihnout podstatu, tedy vytvořit pravidlo, které bude dobře klasifikovat i nově přicházející vzorky. To se dobře měří sensitivitou a specificitou. Rád bych viděl, jaká je sensitivita a specificita vašich klasifikačních stromů na testovací sadě dat, tedy na sadě, na které jste strom nekonstruovala. Tohle je klíčová věc každého klasifikačního algoritmu a udivuje mne, že o tom nikde nepíšete. Tento bod bych prosil zpracovat písemně, je to trochu víc práce a u obhajoby to nestihne odpresentovat ústně.

4. Můj obecnější komentář ke klasifikačním stromům je tento: Zřejmě jde o to rozparcelovat trénovací data v D-rozměrném eukleidovském prostoru pomocí **rozumného množství svislých a horizontálních čar** tak, aby byla příslušnost ke skupině (měřena na testovacích datech) co nejlepší. Použití svislých a horizontálních čar mi přijde neuvěřitelně omezující, jakékoliv komplexní hypotézy (třeba případy uvnitř kruhu, kontroly vně) to bude fitovat velmi nešikovně. Předpokládám, že přístup bude trpět prokletím dimenze. Jaký je váš názor?

5. Proč v kapitole 4 postupujete tak, že odebíráte vždy ten nejdůležitější klasifikátor? Tento postup je těžištěm vašeho postupu, ale jeho logika mi uniká.

Celkově soudím, že diplomová práce je zajímavá, obsáhlá a vyžadovala hodně úsilí. K její psané podobě mám sice mnoho výhrad, ale práci rozhodně doporučuji k obhajobě. Předběžně navrhuji hodnocení C.

V Olomouci 10. dubna 2014

Tomáš Fürst