Posudok vedúceho na diplomovú prácu Mgr. Jakuba Berana, M.A. „Machine learning on small datasets with large number of features”

Cieľom predkladanej práce bolo navrhnúť a otestovať postupy na strojové učenie klasifikátora na trénovacích dátach, ktoré sú obvyklé v bioinformatických aplikáciách. To sú trénovacie dáta s relatívne malým počtom záznamov ale mnohonásobne vyšším počtom atribútov. Nebolo zámerom vyvíjať nové metódy klasifikácie, ale preskúmať aké kombinácie metód selekcie atribútov a klasifikátorov sú na takúto úlohu vhodné.

Jedným z hlavných problémov s takýmito trénovacími dátami je, že atribúty nie sú nezávislé. Autor navrhol dve štúdie, ktoré skúmali takéto kombinácie na umelých dátach, kde je možné jednoducho vytvoriť skupiny vzájomne korelovaných atribútov a potom overiť, či daná metóda selekcie atribútov dokáže vybrať reprezentatívne atribúty.

Na reálnych dátach nie je možné overiť, či daná metóda vybrala dostatočne reprezentatívne atribúty. Avšak autor navrhol a experimentálne overil, že pomocnou mierou kvality selekcie reprezentatívnych atribútov môže byť stabilita takého výberu, keď sa daná metóda aplikuje na náhodne vybrané podmnožiny dát.

Druhým výsledkom je rigorózny popis algoritmu selekcie atribútov nazývaného Boruta. Tento algoritmus vyberá všetky relevantné atribúty, ale jeho popisy zo známej literatúry sú neúplné, čo autor v texte práce napravil. Je to dôležité preto, že práve tento algoritmus selekcie atribútov je pre skúmaný problém jeden z najlepších.

Tretím výstupom práce sú výsledky rozsiahlych experimentov na umelých i reálnych dátach, ktoré ukazujú, ako úspešné sú kombinácie rôznych metód selekcie atribútov a rôznych klasifikátorov. Obzvláť cenné a čiastočne prekvapivé sú potom závery a odporučenia, ktoré kombinácie metód selekcie atribútov a klasifikácie sú vhodné na takéto dáta. Napríklad selekcia atribútov pomocou náhodného lesa na malých trénovacích dátach nefunguje dobre, ale algoritmus Boruta, ktorý je náhodných lesoch založený, naopak funguje výborne.

Práca je napísaná slušnou angličtinou a minimom gramatických chýb. Autor dokázal do relatívne krátkeho textu dostať výsledky rozsiahlych experimentov a potom ich prehľadne vyhodnotiť a odvodiť odporučenia na využitie vhodných kombinácií metód selekcie atribútu a klasifikátorov.

Autor prácu vypracoval samostatne a používal iba citované zdroje. Všetky kroky i text práce so mnou pravidelne konzultoval. Metodológia, ktorú navrhol je aplikovateľná i na iné metódy selekcie atribútov a klasifikácie, než tie, ktoré sú v práci využité. Autor dokázal prehľadne spracovať výsledky podrobných experimentov a priložené programy umožňujú všetky experimenty a dosiahnuté výsledky jednoducho reprodukovať.

Celkovo je práca Mgr. Jakuba Berana, M.A., na vysokej úrovni. Jeho závery majú okamžité praktické uplatnenie. Preto odporúčam, aby táto práca bola uznaná ako diplomová práca.

Praha, 30.6.2020

RNDr. František Mráz, CSc.
KSVI MFF UK