



Posudek oponenta závěrečné práce

Oponent práce: Ing. Daniel Vašata, Ph.D.
Student: Bc. Maroš Kramár
Název práce: Implementace metriky pro větvení v Isolation Forest
Obor / specializace: Znalostní inženýrství
Vytvořeno dne: 5. června 2023

Hodnotící kritéria

1. Splnění zadání

- [1] zadání splněno
- ▶ [2] zadání splněno s menšími výhradami
- [3] zadání splněno s většími výhradami
- [4] zadání nesplněno

Zadání bylo splněno. Myslím ale, že by část zaměřená na vlastní zkoumání možností, jak zkombinovat dělicí kritéria v libovolném směru (non-axis-parallel splits) s chytřejšími způsoby nalezení prahu pro dělení (split guiding), mohla být více rozpracována. Studentovy návrhy v části 4.3 mi přijdou na diplomovou práci poněkud jednoduché (především je triviální studentovo pozorování parametrické redundance v dělicím kritériu prvního návrhu na str. 34).

2. Písemná část práce

80/100 (B)

Práce je logicky strukturovaná a po jazykové stránce v pořádku. Zdroje jsou relevantní a správně citované. Po obsahové a typografické stránce mám k práci několik výtek. Práce je občas psána formou vyprávění o tom, jak autor postupoval (např. v části 4.3). Autor občas používá autorský plurál (str. 40), občas používá singulár (str. 41). Z matematického pohledu jsou v práci neoptimální výrazy a formule. Například vztah v části 5.8.4 obsahuje hvězdičku (symbol konvoluce) pro násobení a proměnnou "depth" psanou kurzívou (což není správně, protože to je stejné jako součin jednopísmenných veličin d , e , p , t , a h). Vztah (3.5) obsahuje opět hvězdičku, ale především tam je v argumentu exponenciály logaritmus $0|X^k|$, což je dle mého názoru součin nuly s něčím a tedy nula, což je mimo definiční obor logaritmu. První vztah v části 4.4 obsahuje dle textu definici nadroviny. Uvedená rovnice tomu ale neodpovídá, protože to je pouze definice funkce f . Teprve její rovnost 0 může definovat nadrovinu. V části 4.4 se píše o projekci na nadrovinu. To je ale dle mého názoru nesprávně, protože projekce bývá do nějakého směru - aby výsledkem byl jednorozměrný objekt, který jde právě porovnávat a dělit podle nějaké prahové hodnoty.

3. Nepísemná část, přílohy

97 /100 (A)

Nepísemnou částí práce byla především netriviální implementace zvoleného Fair Cut Forest algoritmu do frameworku H2O a dále realizace experimentů porovnávajících tuto implementaci s existujícími algoritmy na bázi Isolation Forrest, které již v H2O implementovány byly. Všechny provedené experimenty jsou v práci srozumitelně popsány. V příloze práce jsou pak k dispozici všechny potřebné zdrojové kódy, které jsou přehledné a na velmi dobré úrovni.

4. Hodnocení výsledků, jejich využitelnost

96 /100 (A)

Teoretickým výsledkem práce je přehledný popis detekce anomálií pomocí Isolation Forrest a jeho pokročilejších variant. Praktickým výsledkem práce je potom netriviální implementace Fair Cut Forest algoritmu do frameworku H2O a provedené experimenty zkoumající jednak rozšíření variant Isolation Forrest a také porovnání této implementace v H2O s již existujícími obdobnými algoritmy. Výsledek je 100% využitelný, protože je (nebo jistě brzy bude) součástí platformy H2O.

Celkové hodnocení

90 /100 (A)

Text práce má drobné nedostatky ale implementační výsledky jsou na výborné úrovni s reálným přínosem. Proto navrhuji hodnocení stupněm A.

Otázky k obhajobě

V části 5.8.1 píšete, že při testech vyšla implementace gradFindSplit stejně rychlá jako exhaustive search a, že to může být jiné na velkých datasetech. Proč jste tedy nevzal velký dataset a neotestoval jste, že je toto tvrzení pravdivé?

Instrukce

Splnění zadání

Posudte, 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, posudte 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.

Písemná část práce

Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posudte, 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. Posudte správnost používání formálních zápisů obsažených v práci. Posudte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posudte, 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.

Nepísemná část, přílohy

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.

Hodnocení výsledků, jejich využitelnost

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.

Celkové hodnocení

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.