

Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Bc. Martin Adam
Název práce Strojové učení pro monitorování počítačových clusterů
Rok odevzdání 2020
Studijní program Informatika **Studijní obor** Umělá inteligence

Autor posudku Mgr. Štěpán Balcar **Role** Oponent
Pracoviště Katedra softwarového inženýrství MFF UK

Text posudku:

Student se v práci zabývá metodami strojového učení v doméně monitorování počítačových clusterů. Cílem je prozkoumat možnosti využití a přínosy technik umělé inteligence pro správu a administraci datacenter. Práce si klade za ideální cíl navržení přístupu, který by mohl odhalovat problémy infrastruktury výpočetního klastru dříve, než se plně projeví.

Prvním krokem bylo zpracování přehledu stávajících monitorovacích systémů. Jsou popsány administrátorské nástroje poskytující informace o strojích na úrovni OS (numerické metriky, analýzy logů). Experimenty byly prováděny na monitorovacím systému MONIT na výpočetním klastru v CERNu. Volba monitorovacího nástroje je dobře zdůvodněna a zasazena do kontextu souvisejících prací.

Součástí práce bylo vytvoření vlastních datasetů. Datasety obsahují měření vybraných metrik a anomálií klastru v průběhu výpočtu, vzniklých monitorováním korektního běhu výpočtu i vyvoláním rozličných chybových stavů. Použitá datová sada je dostatečně heterogenní z hlediska druhu a frekvence výskytu anomálií.

Autor práce použil regresní algoritmy a techniky učení bez učitele a využil jejich větší odolnosti vůči přeučení. Pro redukci počtu dimenzí používá PCA. Pro evaluaci a vzájemné srovnání technik jsou jednotně využity metriky precision a recall. Jako baseline pro regresi je použitý model, který opakuje poslední známý stav. Autor přidal k regresnímu modelu ještě filtrační algoritmus, protože považovat i každou velkou odchylku od regresní předpovědi za anomálii (tři standardní odchylky) vede pořád k vytváření příliš mnoha falešných pozitiv. Random Forest v kombinaci s filtračním algoritmem vylučuje falešná pozitiva a dává pozoruhodně dobré výsledky. Výběr hyperparametrů modelů je korektně zdůvodněn jako výsledek křížové validace. Hyperparametrům filtračních algoritmů se ale pozornost příliš nevěnuje.

Práce je přehledně členěná, rozsahově spíše stručná. Angličtina je na vysoké úrovni a srozumitelná, i když je místy znát, že nejde o autorův rodný jazyk. Seznam použité literatury obsahuje překlady a způsob citování zdrojů není jednotný. Odkazované publikace lze však z obsažených informací bez potíží dohledat, takže se jedná pouze o formální problém. Jména tabulek a obrázků by mohla být stručnější. Pro podrobnější popis obsažených dat by měla sloužit poznámka/anotace. Metriky precision a recall umožnily přehledné numerické porovnání výsledků rozličných modelů.

Součástí není programátorská ani uživatelská dokumentace ani sw analýza problematiky.

Práce je velmi strohá, obzvláště v popisu tvorby datasetů. Přínosné by bylo popsat a zdůvodnit i rozhodnutí, která musela být učiněna při tvorbě datasetů a návrhu experimentů. Zdrojový kód je přehledný, ale obsahuje minimum komentářů.

Práci celkově hodnotím velmi kladně. Cíle práce byly naplněny. Martin Adam ukázal možnost predikovat závady výpočetních klastrů pomocí strojového učení. Práce má vědecký přínos a praktické využití. Výsledky práce byly publikovány na International Conference on Computing in High Energy and Nuclear Physics (CHEP2019).

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Datum 27.1.2020

Podpis