Autor: Bc. Prokop Šilhavý

Název: Using Double Oracle Algorithm for Classification of Adversarial Actions

Posudek vypracoval: Ing. Ondřej Kuželka, Ph.D.

Předložená práce se zabývá problematikou adversariálního strojového učení, kde na jedné straně stojí “obránce” (klasifikátor), jehož úkolem je rozpoznat “příklady”, které vytvořil útočník od “normálních” příkladů, a na druhé straně je útočník, který se snaží klasifikátor “přelstít”. Práce motivuje tento obecný rámec na příkladu rozpoznávání maligního chování v počítačových sítích, kde je cílem klasifikátoru rozpoznat v datech o provozu v síti chování útočníka.

Pro řešení tohoto problému se předložená práce obrací k metodám teorie her. V kapitolách 2 a 3, které následují hned po úvodu jsou velmi pěkným způsobem popsány základy teorie her (včetně některých pokročilých konceptů, které jsou pro tuto práci potřeba): hry v normální formě, nekonečné hry, Nashovo ekvilibrium a metody pro jeho hledání s hlavním důrazem na tzv. double-oracle přístup (jímž se, pokud je mi známo, intenzivně zabývá či zabýval vedoucí diplomové práce).

Ve čtvrté kapitole pak autor formalizuje problém adversariální klasifikace jako problém teorie her. Nejprve tento problém formalizuje jako hru s nenulovým součtem. Jelikož je ale řešení her s nenulovým součtem (hledání jejich Nashových ekvilibrií) výpočetně složité, navrhuje tři různé přístupy, jak tento problém formalizovat jako hru s nulovým součtem. Na základě analýzy těchto tří možných přístupů pak vybírá jeden konkrétní přistup, který je založený na přidání omezující podmínky, která vyžaduje, aby počet falešně pozitivních příkladů byl v průměru menší nebo roven zadanému limitu.

V páté kapitole pak autor popisuje detaily navrženého přístupu a jeho implementaci. V šesté kapitole pak následuje detailní popis výsledků experimentů.

Konečně v sedmé, poslední, kapitole autor shrnuje dosažené výsledky a navrhuje možný budoucí směr dalšího výzkumu.

Téma práce je velmi zajímavé. Ačkoliv nejsem expert v této doméně, můj dojem z této práce je velice pozitivní. Domnívám se že autor dané problematice dobře rozumí. Velmi pozitivně hodnotím to, že autor nepopisuje jen jeden konkrétní přístup, ale i možné alternativy a vysvětluje, proč je jím zvolený přístup vhodnější (rovněž pozitivně hodnotím, že otevřeně popisuje i nevýhody svého přístupu). Tento přístup sice v některých místech práce vede k tomu, že se v ní čtenář ne úplně snadno orientuje (alespoň já jsem si občas při čtení této práce nebyl jistý, co je použitý přístup a co zavržená alternativa). Jinak ale hodnotím práci pozitivně.

Jedna z věcí, která by potenciálně mohla být zlepšena, je to, že (alespoň, jak jsem to z textu pochopil a je možné, že jsem něco přehlédl) autor nezkoumá problém generalizace - benigní příklady jsou zafixované během trénování a už není dále zkoumáno, jak se navržený přístup chová na benigních příkladech mimo trénovací množinu.

Co se prezentace týče, angličtina je ve většině kapitol dobrá, jen v úvodních kapitolách jsou určité a neurčité členy používány, řekl bych, poněkud nestandardně.

Celkově se domnívám, že se jedná o velmi zdařilou práci.

Hodnocení: Práci hodnotím známkou **A** (**výborně**).

Datum a podpis: