

Detekce malwaru je klíčovým aspektem kybernetické bezpečnosti a představuje řadu výzev, zejména ve scénářích uvažujících proud dat, kde dochází k silné změně distribuce a velkému zpoždění mezi obdržáním dat a získáním jich třídy. Změna distribuce je charakterizována přítomností vysoce indikativních, ale rychle se měnících rysů, jako jsou specifické názvy souborů nebo mutexy. Malware však vykazuje také řadu stabilních rysů, jako jsou typy připojení nebo metody zpeněžení, které zůstávají v čase relativně konzistentní. V této práci formalizujeme tento scénář a dále zkoumáme hypotézu, že adaptivní odstranění silně driftujících podmnožin rysů může mít velký vliv na výkonnost algoritmu. V práci prokážeme, že současné metody opravdu vykazují nedostatky spojené s těmito rysy, zejména potom v krátkých obdobích po příchodu nové distribuce. Abychom ověřili hypotézu o zlepšení výkonnosti prostřednictvím adaptivního odstraňování příznaků, předkládáme dvě řešení: jedno založené na detekci změny distribuce pomocí Hellingerovy vzdálenosti a druhé na inkrementálním algoritmu Gaussian Mixture Model. Oba přístupy vyhodnocujeme na reálných datech a na naší syntetické datové sadě a ukazujeme výrazné zlepšení na syntetických datech a slibné výsledky na reálných datech. Kromě toho uvádíme komplexní vysvětlení technik použitých v práci.