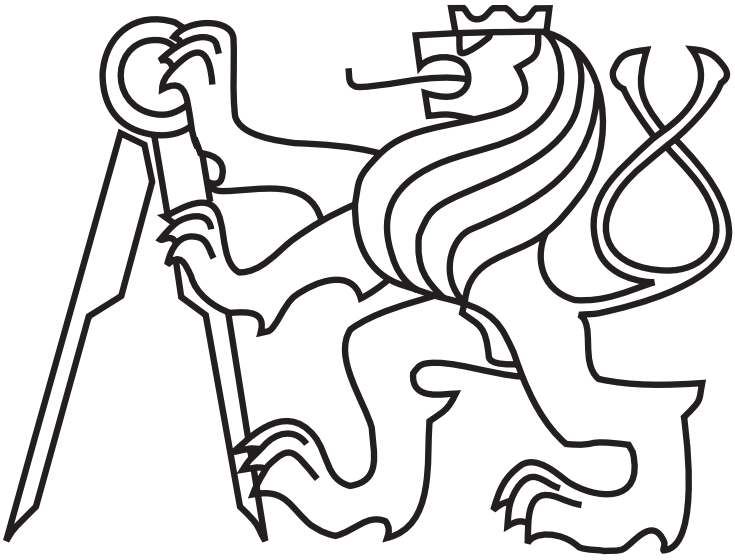
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta elektrotechnická

Bakalářská práce



Adam Ficenec

Detekce cheaterů ve hrách

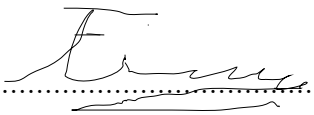
Katedra řídicí techniky

Vedoucí práce: **Ing. Ondřej Pluskal**

Prohlášení

Prohlašuji, že jsem svou bakalářskou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v přiloženém seznamu.

V Praze dne 3.1.2014 .....



Czech Technical University in Prague
Faculty of Electrical Engineering

Department of Control Engineering

BACHELOR PROJECT ASSIGNMENT

Student: **Adam Ficenec**

Study programme: Cybernetics and Robotics
Specialisation: Systems and Control

Title of Bachelor Project: **Cheating Detection in Online Games**

Guidelines:

1. Research the state of the art of cheater detection.
2. Analyse online game dataset from game industry.
3. Create a cheater model.
4. Evaluate the model.

Bibliography/Sources:

[1] S. F. Yeung , John C. S. Lui , Jiangchuan Liu , Jeff Yan, "Detecting cheaters for multiplayer games: Theory, design and implementation"
[2] Ruck Thawonmas, Yoshitaka Kashifuji, and Kuan-Ta Chen, "Detection of MMORPG Bots Based on Behavior Analysis,"
[3] Chapel, L.; Botvich, D.; Malone, D., "Probabilistic approaches to cheating detection in online games,"
[4] Muhammad Aurangzeb Ahmad , Brian Keegan , Jaideep Srivastava , Dmitri Williams , Noshir Contractor, "Mining for Gold Farmers: Automatic Detection of Deviant Players in MMOGS"
[5] Ah Reum Kanga, Jiyoung Wooa, Juyong Parkb, Huy Kang Kima, , Online game bot detection based on party-play log analysis"

Bachelor Project Supervisor: Ing. Ondřej Pluskal

Valid until the winter semester 2014/2015

prof. Ing. Michael Šebek, DrSc.
Head of Department

L.S.

prof. Ing. Pavel Ripka, CSc.
Dean

Prague, September 26, 2013

Poděkování

Velmi rád bych poděkoval vedoucímu své práce, ing. Ondřeji Pluskalovi za cenné rady v průběhu tvorby celé práce a jeho aktivní a přátelský přístup. Dále bych rád poděkoval své přítelkyni a rodině za psychickou podporu a toleranci, díky které jsem byl schopen studovat a psát tuto práci v pohodové atmosféře a klidu.

Abstrakt

Cílem mé práce je udělat analýzu metod používaných pro detekci podvodníků v online hrách a na základě toho připravit software pro jejich detekci a to konkrétně ve hře Pool Live Tour od společnosti Geewa a.s. Po výběru vhodné metody identifikace cheaterů pro tuto hru se ji pokusím aplikovat na reálná data, které byla poskytnuta společností Geewa a.s. Data obsahují informace z jednoho měsíce o všech hráčích včetně těch, kteří byli ze hry vykázáni. Na základě těchto informací se pokusím připravit co nejpřesnější algoritmus schopný odhalit podezřelé herní anomálie a tím i skutečné podvodníky.

Abstrakt

The goal of this work is to analyse state of the art methods of cheater detection in online games, and to prepare a software for their detection in specific online game, namely Pool Live Tour by Geewa a.s. company, based on the previous research. Chosen methods will be tested on real data provided by Geewa a.s. These data consist of logfiles of each active player throughout one month of playing, including players that were banned for their behavior. Based on the given data, I will try to create the most effective algorithm capable of revealing suspicious gaming anomalies and thus real cheaters.

Obsah

1 Cíl práce **1**

2 Úvod **1**

3 Úvod do problematiky **5**

3.1 Naivní detekce 6

3.2 Detekce pomocí učení bez učitele 7

3.3 Detekce pomocí učení s učitelem 7

3.4 Kombinace unsupervised a supervised metod 8

4 Popis problému **10**

4.1 Geewa Pool Live Tour 10

4.2 Herní data 11

4.3 Extrakce dat 13

5 Analýza dat a implementace existujicích algoritmů pro detekci cheaterů 14

5.1 WEKA 14

5.2 Hodnocení výsledků 15

5.3 Detekce na základě učení bez učitele 16

5.3.1 DBSCAN 17

5.4 Detekce na základě učení s učitelem 17

5.4.1 Naive Bayes 18

5.4.2 IBl 19

5.4.3 AD tree 19

6 Implementace vlastních metod **21**

6.1 Analýza dat a vytvoření modelu cheatera 21

6.2 Vytvoření vlastního kódu 22

6.3 Aplikace kódu na reálná data 24

6.3.1 Porovnání celkového množství prodeje a nákupu tág 24

6.4 Analýza prodeje a nákupu stejných tág v časovém okně 25

7 Výsledky úspěšnosti jednotlivých algoritmů **26**

7.1 Shrnutí uplatnění existujících metod 27

7.2 Shrnutí implementace vlastního kódu 28

8 Závěr **30**

Seznam obrázků

1 Profit online her od roku 2006 do roku 2012 2

2 Lobby hry 10

3 Příkladový logfile hráče, neobsahuje ID 11

4 Vizualizace nerozlišitelnosti cheaterů od hráčů 16

5 DBSCAN algoritmus 17

6 Příklad vizualizace AD stromu 20

7 Rozložení prodeje tág v rámci získaného měsíce 22

Seznam tabulek

1 Formát jednotlivého řádku a jeho popis 12

2 Tabulka atributů 13

3 Šablona chybové matice pro výpočet precission a recall hodnot 15

4 Tabulka podezřelých hráčů z podvodného prodeje 24

5 Výsledky algoritmu DBSCAN, parametry viz. 5.3.1 26

6 Výsledky algoritmu Naive Bayes 26

7 Výsledky algoritmu IBL 26

8 Výsledky algoritmu AD rozhodovacího stromu 27

9 Tabulka podezřelých hráčů z prodeje stejných tág, pro nižší prahy 28

10 Tabulka podezřelých hráčů z prodeje stejných tág, pro vyšší prahy 28

11 Obsah CD 34

12 Tabulka možných akcí v logfile souboru 35

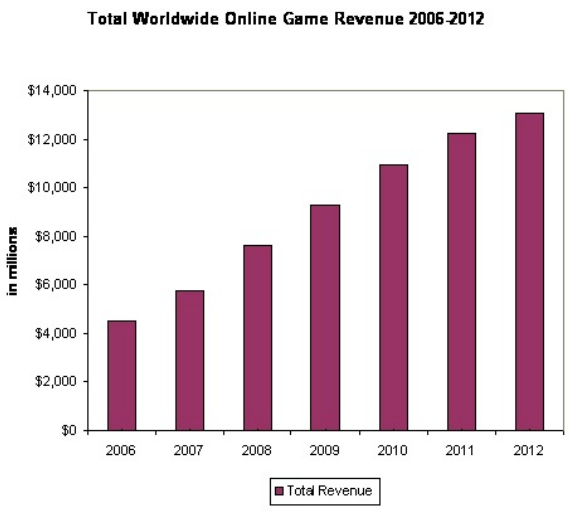
1 Cíl práce

Cílem této práce je představit problematiku podvodníků v online hrách, rozebrat nejmodernější metody určené k jejich detekci a poté se tyto metody pokusit uplatnit na reálná data v rámci představené online hry. Další část práce se bude zabývat analýzou těchto dat a nalezení modelu, pomocí kterého by bylo možné podvodníka identifikovat. Tento model bude následně zpracován a poslouží jako stavební kámen pro konstrukci vlastní metody určené k detekci cheatera ve vybrané online hře. Výsledky úspěšnosti existujících i vlastních metod budou v závěru práce shrnuty a analyzovány pro jejich budoucí vylepšení jak v globálním měřítku, tak v rámci zmíněné hry.

2 Úvod

Herní průmysl se v posledních desetiletí velice rychle rozrůstá a na trh přinesl velké množství obchodních příležitostí. Mezi hlavní větve pak patří online hraní, které se neustále rozšiřuje. Vzhledem k tomu, že se jedná o zábavní průmysl založený na interakci velkého množství uživatelů, kteří se snaží dosáhnout nějakého cíle či výhry většinou nějakou formou kompetice s ostatními hráči, je tato odnož velice populární a proto herní společnosti vykazují vysoké zisky [1] jak je vidět i z přiloženého grafu.[1] Firmy jsou tak schopné investovat nemalé částky do projektů i s poměrně velikým riskem neúspěchu [2]. Základ úspěchu pak stojí jak na vhodně vybraným tématem jako v každém jiném zábavním průmyslu, tak v kvalitě jejího provedení, neboť díky popularitě této branže existuje velké množství konkurence. Hry proto musí být bezproblémové, stabilní a vzhledem k jejich povaze velice odolné vůči hackerům, kteří nerespektují herní pravidla a podvodnými způsoby se snaží hru oklamat ku vlastnímu prospěchu. Tyto „cheateři“ mají na hru neblahý a nezanedbatelný dopad, vzhledem k tomu, že mohou způsobit reálnou finanční škodu, získat nestandardní výhody oproti běžným hráčům a obecně kazit zábavu těm čestným, což v řadě případů

[1] Převzato z http://www.dfcint.com/game_article/may07article.html



Obrázek 1: Profit online her od roku 2006 do roku 2012

může vést k poklesu hrajících, či zájemců o hru a tím i velikým finančním újmám pro daného vydavatele. Proto by ochrana před těmito útoky měla být jednou z prioritních problémů všech kvalitních online her.

Typy herních podvodů (cheatů) se liší podle her, jejich žánru i způsobu provedení a proto jich je veliké množství. Mezi nejčastěji používané útoky patří nabourání se do databáze uživatelů a změna jejich dat nebo jejich pouhé odcizení, především informace o kreditních kartách, heslech a podobně. Tato metoda je dnes ovšem náročná díky lepšímu zabezpečení datových serverů, přesto může nastat i zkušeným developerům [3]. Podobnou, rozmáhající se metodou je tzv. "phishing". Tento trend je dnes velice populární a využívá důvěry uživatelů. Hacker rozešle emaily, tvářící se jako oficiální pošta z dané hry vybízející ke změně či udání hesla nebo čísla kreditní karty na jeho stránkách, které opět kopírují oficiální vzhled hry i její adresu (liší se například v pár znacích, kterých si uživatel nevšimne). Tyto metody řadím mezi externí, neboť přímo neovlivňují chod hry, ale zneužívají soukromé informace uživatelů a proto se jimi nadále zabývat nebudu.

Útoky, které přímo ovlivňují herní prostředí jsou také rozmanité, ovšem jejich nalezení a potrestání je reálněji proveditelné než u těch externích. Patří sem například tzv. aim-boti v

FPS hrách (first person shooter), kteří hráči zajišťují nadstandardní přesnost, dále aplikace způsobující rychlejší pohyb po herní mapě, algoritmy či skripty, které hrají hru za hráče a vykonávají jednoduchou, stále se opakující činnost, která hráči může přinést různé výhody. Například těžba zlata ze stejného, obnovujícího se místa přinese hráči vysoké výdělky bez jakékoliv práce. Firmy se tak těmto podvodníkům snaží bránit vydáváním aktualizací či softwaru, který jsou schopen známé podvody detekovat, ovšem i tyto aplikace jsou náchylně vůči hackerům, neboť jsou instalované na straně klienta. Proto může být jejich efektivita nízká, protože stejně jako u antivirového softwaru, obě strany se neustále přizpůsobují, aby byli o krok napřed před druhou. Moje práce se proto bude zaměřovat na řešení, využívající pouze serverovou komunikaci mezi klientem a hrou, tedy na řešení, která jsou vůči výše zmíněným útokům imunní, protože nejsou klientovi k dispozici. Tato řešení využívají tzv. logfile, který je ukládán pro každého hráče, z důvodu zachování dat při výpadku serveru a pro detekci herních chyb a problémů. Na základě informací poskytnutých těmito dokumenty je možné vytvořit efektivní program pro detekci podvodníků v hrách. Efektivita programu je ovšem svázána s dostupnými informacemi, které jsou k dispozici v daném dokumentu, proto by měl mít tento soubor co nejobsáhlejší. Vzhledem k povaze metody, která detekuje podvodníky na základě herních informací, bývá snadno modifikovatelné pro různé typy her i typů podvodů. Cenou je ovšem přístup programátora k osobním datům, proto bývají tyto herní zápisy komunikace mezi hráčem a serverem omezeny na nejnutnější minimum, což komplikuje tvorbu kvalitních detektorů cheaterů.

Struktura práce bude následující.Nejprve stručně představím moderní metody používané pro detekci cheaterů v online hrách a jejich principy, které budou v následujících částech více rozvedeny. V druhé kapitole se budu věnovat popisu online hry, na kterou vybrané metody uplatním. Kapitola se dále bude věnovat struktuře získaných dat ze hry, jejich extrakci a roztřídění pro budoucí analýzu. V další kapitole budou na základě analýzy dat vybrány a popsány jednotlivé algoritmy, které budou použity pro detekci cheaterů v představené hře. Čtvrtá část se bude zabývat vytvořením a popisem vlastní metody pro detekci na základě rozboru dostupných dat. Výsledky všech experimentů budou sepsány a diskutovány v následující kapitole. Závěr práce pak stručně shrne dosažené výsledky a

rozebere možná zlepšení pro případné budoucí navázání.

3 Úvod do problematiky

V dnešní době existuje mnoho možných řešení detekce cheaterů s různou efektivitou a složitostí. Protože existují různé typy podvodů, škála možných řešení je veliká. Například pro hledání botů (algoritmus hrající za hráče) a účtů, které byly ukradnuty (hacker se nějaký způsobem zmocní hráčova přístupového jména i hesla), se může použít identifikace hráčů na základě jejich časového vyžití [4]. Hledání ukradnutých účtů a detekce botů je založena na unikátnosti časového rozložení aktivního/neaktivního hraní, díky které se dá s vysokou pravděpodobností rozeznat, zda li je herní postava ovládána skutečným hráčem a majitelem účtu. Náhodná shoda tohoto rozložení pro únosce účtu i majitele je vysoce nepravděpodobná (podobně jako například otisk prstu). Boti jsou pak rozpoznáváni podle příliš vysoké aktivity, které by skutečný hráč nebyl schopný dosáhnout. Tato řešení však často bývají "šitá"na míru dané hře a dají se špatně aplikovat pro jiné aplikace. Proto zde budou uvedena pouze nejuniverzálnější a obecně nejvýhodnější metody vhodné detekci hráčů pro online hry. Jejich principy budou vysvětleny v následujících sekcích a budou vybrány metody, které budou nejlépe uplatnitelné pro řešení této práce. Proto při rozboru vynechám například ruční detekci založenou na kontrole uživatelů, kteří byli nahlášeni samotnými hráči jako podezřelé. Je totiž víceméně založena na odpovědnosti uživatelů, což je neovlivnitelný faktor, který může velmi zmanipulovat výslednou efektivitu detekce.
Obecně můžeme rozdělit tyto metody podle způsobu zpracování dostupných informací a jejich využití pro nalezení podvodu. Dále jsem vybíral pouze metody které mohu uplatnit vzhledem k dostupným informacím v podobě herních záznamů. Existuje totiž velké množství metod, které vyžadují okamžitou serverovou komunikaci s klientem pro úspěšnou detekci. Tyto metody nemohu využít a proto dále nebudou představeny. Názvy vycházejí z obecného pojmenování strojového učení, protože se tato tematika s naším problémem v mnoha směrech shoduje.

Metody jsou následující:

- Naivní
- Učení bez učitele (Unsupervised)
- Učení s učitelem (Supervised)
- Kombinace předchozích dvou metod (Semisupervised)

Toto rozdělení je generalizující, ovšem vzhledem k podobnosti metod v jednotlivých skupinách a rozdílnosti s ostatními je vhodné. Nyní jednotlivé typy metod stručně rozeberu, poukážu na jejich výhody a nevýhody. Po představení online hry, pro kterou budu problém řešit a jejích rysů pak vyberu nejvhodnější metodu pro řešení zadání mé práce a pokusím se jí aplikovat na reálná data které jsem dostal k dispozici.

3.1 Naivní detekce

Naivní detekcí jsou nejjednodušší možné metody detekce cheaterů. Jejich princip spočívá čistě na základě určení uskutečnitelného nějakého prahu/limitu pro danou akci zvládnutelné běžným hraním, popřípadě množství vykonaných akcí za dobu. Může se jednat o přesnost, rychlost, zvládnuté úkoly za časový úsek, doby aktivního hraní atd. Pokud hráč překoná daný limit, je označen za možného cheatera a dále sledován, popřípadě rovnou potrestán. Jde tedy v podstatě o jednoduchý filtr. Tato metoda může mít různou úroveň efektivity a užití. Její nevýhodou je nutnost seznámení se s daty a přípravě vhodného filtru, což může být časově náročné a zdlouhavé. Její výhoda pak tkví v jednoduchosti a široké aplikovatelnosti i modifikovatelnosti, ovšem může být nepoužitelná pro sofistikovanější podvody, které se snaží imitovat reálné chování.

3.2 Detekce pomocí učení bez učitele

Metoda učení bez učitele je obecné označení pro algoritmy, které se snaží najít skryté struktury nebo shluky v neoznačkovaných datech. Existuje velké množství těchto metod, jako je vyhledávání na základě rozložení dat podle gaussovy křivky [5], k-means clustering třídící data do předem určeného počtu skupin na základě průměrné vzdálenosti mezi blízkými body [6], nebo vyhledáván datových shluků na základě hustoty (viz. DBSCAN obrázek 5). Vzhledem k povaze problému s detekcí cheaterů, která má za úkol vyhledávat méně běžné hodnoty a abnormální data je pro náš účel použitelná pouze poslední zmíněná metoda třídící na základě hustoty dat, protože je schopná rozlišit datový hluk, který může představovat rysy cheatujícího hráče. Nevýhodou těchto postupů je navíc případ, ve kterém podvádějící hráči svým množstvím vytvoří shluk natolik veliký, že ho program vyhodnotí jako regulérní, protože rozlišit podvodníky a normální hráče nadále není možné. To se může stát v případě nějakého na provedení jednoduchého podvodu , který se v komunitě rychle rozšíří a nestačí se včas zamezit jeho zneužití, což bývá v online hrách docela běžný případ [7].

3.3 Detekce pomocí učení s učitelem

Metody, které obecně používají předem nasbíraná a vyhodnocená data pro naučení se rozpoznávání vzorců v dané oblasti. Algoritmy založené na tomto principu jsou dnes hojně využívány v mnoha oborech robotiky a informatiky, jako je například rozpoznávání písma, objektů, ale i pro detekci anomálií. Mezi nejrozšířenější z těchto metod patří, díky její relativně jednoduché aplikaci, různé formy Bayesovského klasifikátoru, který je postaven na Bayesově teorému, popisující vzájemnou souvislost jevů s podmíněnou pravděpodobností [8]. Metoda je založena na učení z trénovacího vzorku, čímž jsou nastaveny pravděpodobnostní výskyty dané akce. Při překročení nastaveného prahu dochází k detekci podvodu. Tato metoda se dá uplatnit na široké pole problémů, a pokud je správně použita bývá velice spolehlivá. Nicméně v případě nekvalitní trénovací množiny, výsledky mohou být ve-

lice nepřesné, proto bývá vhodné použít co nejobsáhlejší trénovací množinu, obsahující co největší počet typických vzorů, nebo předem vybrat co nejvhodnější trénovací množinu, což bývá neefektivní.Mezi další rozšířené metody klasifikace cheaterů je použití rozhodovacích stromů[9]. Jejich princip je založen na třídění dat/hráčů pomocí specifických dotazů vycházejících z dostupných dat. Hráči jsou tak velmi efektivně zařazeni do podobných skupin i při vysokém množství zohledněných vlastností. Výhodou je relativní jednoduchost při tvorbě, nevýhodou pak naopak výpočetní složitost při vyšším množství porovnávaných atributů. Další nevýhodou této metody a obecně jakéhokoli algoritmu založeného na učení s učitelem, je špatné vyhodnocení neznámých jevů. Pokud trénovací množina neobsahuje nějaký neznámý jev, který v testovací množině nastane, klasifikátor ho kvůli nerozpoznání jakéhokoli vzoru může špatně zařadit. Tato situace je vcelku běžná, pokud se zabýváme detekcí různých anomálií, které mohou být v online hrách poměrně běžné a mohou poukazovat na možného cheatera.

3.4 Kombinace unsupervised a supervised metod

Identifikace podvodníků se, jak již bylo zmíněno, detekce anomálií z velké části podobá (ze své podstaty: podvody vedou k nadprůměrným, popřípadě anomálním výsledkům, jinak postrádají smysl) a proto se metody obou vzájemně téměř překrývají a proto má práce čerpá především z metod, které se pro ně používají. Mezi robustnější a zároveň nejspolehlivější detekci anomálií pak patří využití moderních algoritmů pro detekci pomocí kombinace předchozích dvou skupin, tedy pomocí supervised i unsupervised metod - proto název semi-supervised metody. Tento typ řešení se snažit využít to nejlepší z předchozích skupin, a zároveň se snaží eliminovat jejich nedostatky. Unsupervised detekce totiž ztrácejí na efektivitě v případě příliš vysokého výskytu anomálií, neboť je klasifikuje jako běžná data. Tato analogie jde uplatnit i na poctivé hráče a cheatery. Některé metody, které by to byly schopné řešit, jako například k-means clustering na druhou stranu potřebují předem

známý počet shluků. To velice limituje využitelnost tohoto přístupu, protože bychom mohli pracovat pouze s předem známým počtem anomálií a v takovém případě je vhodnější rovnou použít algoritmy založené na učení se z trénovací metody - tedy supervised metody. Ty na druhé straně selhávají, pokud trénovací metoda neobsahuje všechny známé anomálie, nebo pro náš účel, všechna vzorová data podvodů. Semi-supervised anomaly detection je jednou z metod která těží z obou metod nejlepší výsledky [10], je však závislá na efektivitě předchozích metod pro daný problém, který teprve zlepšuje.

Toto byly některé z nejlepších metod, které je možné využít pro detekci cheaterů v online hrách. Vybral jsem je díky jejich modifikovatelnosti a úspěšnosti.Jejich stručným rozborem se budu věnovat v další části své práce a na základě jejich vlastností vyberu vhodný postup pro řešení mé práce.

4 Popis problému

Tato práce se bude věnovat jedné specifické online hře kulečníku a mým cílem bude pokusit se uplatnit různé algoritmy pro automatickou detekci podvádějících hráčů a snažit se najít řešení pro efektivnější způsob jejich nalezení. Práce bude vycházet z poskytnutých reálných dat nasbíraných za jeden měsíc. Nejprve stručně představím vybranou hru, tak abych pokryl všechny oblasti, ze kterých mám k dispozici herní data.

4.1 Geewa Pool Live Tour

Hra, které se budu věnovat se jmenuje Pool Live Tour a patří herní společnosti Geewa, věnující se online flash hrám pro více hráčů. Hra je vcelku jednoduchá. Po registraci se hráč ocitá v lobby, kde si může vybrat z několika možností (viz obrázek 2). Hra proti

Obrázek 2: Lobby hry

kamarádovi, hra proti náhodně vybranému soupeři, nákup herních mincí, nákup nového tága, hra v turnaji nebo trénink proti počítači. Před každou hrou proti cizímu hráči je před startem zápasu nutno vsadit určitý obnos, který je úměrný postupu ve hře a který v případě výhry hráč vyhrává spolu se soupeřovou sázkou. Při prohře je to naopak. Hráčovi se zvyšují možné sázky postupně podle počtu vyhraných her nad náhodnými oponenty. Pokud uživateli dojdou herní mince, může si je za reálnou měnu dokoupit aby mohl pokračovat

ve hraní, nebo počkat do následujícího dne aby dostal denní bonus za přihlášení. Samotná hra pak sleduje pravidla skutečného kulečníku a jeho variace, které jsou k dispozici opět na základě hráčovy herní úrovně. K dispozici má každý různé druhy tága, které se liší vlastnostmi jako je přesnost, síla a úroveň spinu kterého lze úderem dosáhnout. Většina těchto tág se dá koupit za herní peníze, některé jsou ovšem k dispozici za speciální měnu, kterou lze získat pouze převodem peněz, nikoli hraním hry. Tágo lze koupit během hry i v lobby, kde lze koupit i další kosmetické úpravy jako vzhled avatara, efekty při výhře apod. Turnaje se pak skládají z vyšší základní sázky a několika kol, kde se výše výhry odvíjí od umístění. Data, která jsem měl k dispozici se skládala z výše zmíněných informací.

4.2 Herní data

Geewa a.s. poskytla data, která pokrývala aktivitu všech aktivních hráčů během jednoho měsíce (březen až duben 2013). Data se skládala ze zhruba 25gB a obsahovala herní záznamy 271 143 hráčů, roztříděných podle jejich ID , tedy název každého souboru představoval ID daného hráče a obsah jeho aktivitu za poslední měsíc (viz obrázek 3). Geewa Pool Live Tour logfile obsahoval server/klient komunikaci, která zahrnovala

Obrázek 3: Příkladový logfile hráče, neobsahuje ID

informace o verzi hry, navázání spojení se serverem, pohyb v lobby, koupi mincí a tág, dále výběr hry, sázky a její start, údery hráče s určitým tágem, výběr tága, informace o opuštění hry, prohře/výhře/odplatě a také zda byl uživateli účet zakázán (ban). Logfile neobsahoval konverzaci hráčů, přesnost hráčových úderů v jakémkoliv smyslu - k dispozici byla pouze informace, že hráč vypálil daným tágem. Pokud byl uživatelův účet zabaven, soubor bohužel neobsahoval informace o důvodu tohoto trestu. Jednotlivé řádky souboru se drželi relativně přesného vzoru až na pár výjimek (především místy vynechaný serverový čas), které připisuji lidské chybě při kódování hry. Příklad jednotlivého řádku (viz obrázek 3) přepsáného do tabulky a jeho obecné formátování naleznete v následující tabulce (viz tabulka 1).

123456789	2013-04-16 18:18:58.117000000	2013-04-16 18:19:18.940000000	shot	game-fb	match,play-win	5875,1	””v””:””3.9.24636.7””
ID hráče	Skutečný čas	Serverový čas	Název akce	Rozhraní hry	Podtyp akce	Hodnoty akce	Generická data

Tabulka 1: Formát jednotlivého řádku a jeho popis

Generická data, která obsahovala data jako vygenerované id jednotlivého zápasu, verzi hry, typ použitého tága a jsou vždy zapsány ve složených závorkách, v práci nebyla využita. ID bylo zadáno náhodně, kvůli zachování hráčské anonymity. Analýzou těchto dat byl zjištěn význam jednotlivých akcí a ty pak byly následně sepsány.V příloze je tabulka shromážděných akcí a jejich stručné vysvětlení(viz tabulka akcí v příloze 12). Tato tabulka obsahuje všechny akce, která se v souborech vyskytovala, ale neobsahuje již úplně všechny parametry jednotlivých akcí. Některé nebyly zahrnuty z důvodu přehlednosti, jiné z důvodu nepodstatnosti pro budoucí práci. Formát zápisu akcí byl kvůli své mírné neuspořádanosti upraven aby byl v rámci tabulky co nejpřehlednější a přesto zachoval původní smysl. Formát názvů některých akcí byl proto také upraven a jejich drobné variace přesunuty pod hodnoty sekundárních atributů(např. shop,shop-sell, shop-sell-confirm byly původně jednotlivé akce).

4.3 Extrakce dat

Kromě souborů s daty byly také k dispozici předem zpracované soubory obsahující různé informace o jednotlivých hráčích, které by mohli souviset s charakteristikou případné detekce podvodníků. V podstatě se jednalo o extrakci všech relevantních akcí a jejich zpracování. Vynechány byly akce související s připojováním ke hře, pohyb v herním prostředí (opouštění a navštěvování lobby, zapnutí zvuku atd) a chat. Informace byly uloženy v .csv souborech, které bylo potřeba upravit pro kompatibilitu v programu WEKA, který byl použit pro analýzu dat. Po zpracování a sloučení jednotlivých souborů byla získána data obsahující 61 atributů. Ta sloužila jako základ pro uplatnění algoritmů pro detekcí cheaterů. Seznam těchto atributů je zobrazený v následující tabulce (viz tabulka 2). Vzhledem k podobnostem některých atributů byly tyto atributy vyřazeny z následující analýzy. Jedná se o složky gender-female, neboť obsahují stejná data jako gender-male, pouze s prohozenými číslicemi. To samé se týka i Friend-lose a Play-and-lose (shoda s Friend-win, Play-and-win). Tyto atributy jsou již obsaženy v jiných a jejich existence by pouze komplikovala vyhodnocování jednotlivých algoritmů, popřípadě dobu jejich výpočtu. Dalšími vyřazenými atributy byla procentuální herní doba v rámci celého dne (tedy norm-match-start-hour atributy). Tyto atributy nemají pro tuto práci význam, neboť pouze znázorňují rozložení právě hrajících uživatelů během celého dne. Zůstalo tak 34 atributů, které byly použity pro testování jednotlivých algoritmů.

Atribut	Popis	Atribut	Popis
Id	ID hráče	Age	Věk hráče
Gender-unknown	Neznámé pohlaví	Gender-male	Muž
Gender-female	Žena	Match-start	Počet zápasů
Score	Skóre	Play-and-win	Poměr výher proti náhodnému oponentovi
Friend-win	Poměr výher nad přáteli	Friend-lose	Poměr proher nad přáteli
Play-and-lose	Poměr proher proti náhodnému hráči	Opponent-gallery	Počet prohlédnutých sad tág protihráčů
Opponent-card	Počet prohlédnutých karet hráčů	Owner-card	Počet prohlédnutí vlastní karty
Shot	Počet celkových šťouchů	Total-playtime	Celkový odehraný čas v hodinách
Trophies	Počet trofejí	Played-at-lvl-X	Počet odehraných her na daném levelu (1 - 15)
Maximum-level	Nejvyšší dosažený level	Coins	Finální počet mincí
Norm-match-start-hour-X	Poměr odehrané doby v danou hodinu(0-24)	Registration-bonus	Finální výše bonusů
Shop-buy	Finální počet nakoupených tág	Shop-sell	Finální počet prodaných tág

Tabulka 2: Tabulka atributů

5 Analýza dat a implementace existujicích algoritmů pro detekci cheaterů

Zbylé atributy byly analyzovány pomocí programu WEKA (viz kapitola 5.1), který byl použit i pro implementaci kódů určených k detekci cheaterů. Vytvořený soubor, který obsahoval extrahované informace o všech uživatelích byl v tomto programu nadále analyzován a zpracován.

5.1 WEKA

Program WEKA patří mezi nejrozšířenější aplikace používané pro analýzu velkého množství dat a jejich klasifikaci pomocí zabudovaných algoritmů strojového učení. Aplikace obsahuje velké množství modifikovatelných algoritmů, které využívají učení s učitelem i bez něj, čehož bylo využito pro testování detekce cheaterů na základě jejich odlišnosti od ostatních uživatelů. Díky snadné implementaci pomocí knihoven je i dobře modifikovatelný v JAVA prostředí, které bylo pro práci použito.

Vzhledem k tomu, že extrahovaná data (viz tabulka 2) obsahovala informaci o tom, zdali byl hráč zabanován (atribut ban, 1 = ano, 0 = ne), byla k dispozici možnost uplatnit na tato data metody strojového učení s učitelem i bez něj a porovnání jejich účinnosti, kde správně klasifikovaná hodnota banu sloužila jako základ ověření efektivity použitých algoritmů. Poslední metodou, která byla uplatněna na data byl vlastní kód snažící se nalézt cheatery na základě analýzy logfile souborů a vytvoření modelu cheatera. Tato kapitola stručně vysvětluje princip již existujících metod, které byly použity, včetně jejich výhod a nevýhod. Tabulky výsledků všech algoritmů, jak existujících tak vlastních, jsou pak shromážděny v kapitole 7, kde je diskutována i jejich úspěšnost.

5.2 Hodnocení výsledků

Samotná efektivita jednotlivých algoritmů byla zkoumána pomocí hodnot precission a recall, vycházející z chybové matice. Pokud podle struktury chybové matice označíme správně klasifikované cheatery jako true positive, tedy tp , nesprávně klasifikované cheatery jako false positive fn , správně klasifikované poctivé hráče jako true negative tn a nesprávně klasifikované poctivé hráče jako false negative fn,pro hodnotu precission a recall pak platí následující vztah:

$$Precission = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

Precission, v češtině přesnost pak vyjadřuje pravděpodobnost jevu, že náhodně vybraný prvek je správně klasifikovaný hráč. Recall pak udává pravděpodobnost s jakou hráč není podvodník. Následující tabulka představuje šablonu chybové matice, podle které budou počítány hodnoty precission a recall v kapitole 7.

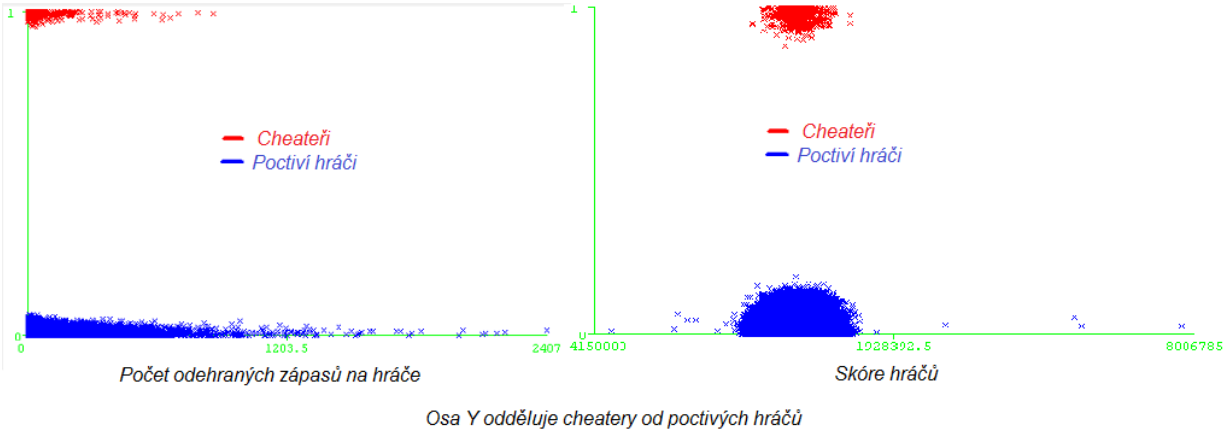
Klasifikace:Non-cheater	Klasifikace:Cheater
Správně identifikovaní cheateři (tp)	Chybně klasifikovaní cheateři jako poctiví hráči (fp)
Špatně identifikovaní hrači jako cheateři (fn)	Správně identifikovaní poctivý hráči (tn)

Tabulka 3: Šablona chybové matice pro výpočet precission a recall hodnot

Tento systém hodnocení je vhodnější než procentuální vyjádření úspěšné klasifikace ku neúspěšné, vzhledem k tomu, že výsledná přesnost je zavádějící, neboť dosahuje hodnot často kolem 99 procent. To je způsobené nadměrně vysokým poměrem nepostihnutých hráčů vzhledem k cheaterům (270 520:622) a jejich úspěšné identifikaci. Nejsledovanější hodnotou je proto poměr špatně klasifikovaných cheaterů ku správně klasifikovaným, vzhledem k jejich malému množství.

5.3 Detekce na základě učení bez učitele

Před samotným uplatněním algoritmů jsem identifikoval, že odhalení cheateři ve většině atributech nevybočují z běžného rozložení průměrného hráče. Dokonce se zdálo, že v mnoha případech patří přímo do průměrných hodnot. V následujícím obrázku 4 je uvedeno několik příkladů. Osa Y odděluje hráče od cheaterů, pro lepší přehlednost. Protože unsupervised



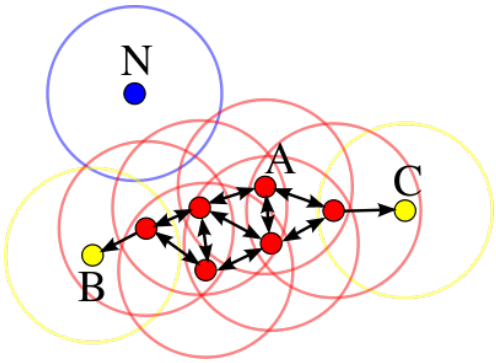
Obrázek 4: Vizualizace nerozlišitelnosti cheaterů od hráčů

metody jsou založené na shlukování dat (clustering) na základě vzájemné podobnosti a odlišitelnosti od ostaních skupin, byla tato metoda testována pouze okrajově, protože cheateři při svém nízkém množství zabíraly širokou škálu hodnot a nezdály se být shlukovány do skupiny. Vzhledem k většímu množství atributů a jejich prolínání, které není možné zobrazit v grafu, tato skutečnost nutně neznamená, že uplatnění algoritmů učení bez učitele metod nemá smysl. Ovšem dalším problémem může být také vetší množství atributů, které způsobují úpadek efektivity těchto metod. Tato chyba je způsobena fenoménem nazývaným curse of dimensionality [11].Vzhledem k těmto vlastnostem dat byl proto vybrán pouze algoritmus DBSCAN, který tyto problémy alespoň částečně adresuje.

5.3.1 DBSCAN

DBSCAN, neboli density based spatial clustering of applications with noise, patří mezi nejpoužívanější shlukovací metody. Algoritmus je založen na vzájemné vzdálenosti jednotlivých bodů od určitého centra hustoty pro daný prostor (viz obrázek 5).

Pro algoritmus jsou důležité dvě hodnoty. Hodnota p udávající kolik je potřeba bodů k



Obrázek 5: DBSCAN algoritmus

vytvoření centra clusteru a hodnota ε, která udává minimální vzdálenost bodů, od které se již nejedná o shluk. Princip je pak jednoduchý: pokud je alespoň počet bodů p od sebe navzájem vzdálených na vzdálenost ε, je vytvořen shluk, kde body, které leží od více bodů ve vzdálenosti ε tvoří jádro shluku a body, které leží pouze od jednoho bodu v této vzdálenosti se stávají okrajovými body. Pokud bod nepatří do žádné skupiny, je označen jako šum. Na obrázku (5) A značí body jádra,B a C okrajové body a N značí šum. Algoritmus byl vybrán díky své schopnosti nacházet shluky i v rámci větších celků a schopnosti rozeznat šum (instancím dat, která vybočují z normálu). Tabulka výsledků při různých hodnotách p a ε je k dispozici v kapitole 5.

5.4 Detekce na základě učení s učitelem

Díky získaným informacím o tom, zda byl hráč označen jako cheater či nikoliv(atribut ban), bylo možné uplatnit tyto metody, kde atribut ban sloužil jako klasifikační hodnota. Kvůli velkému množství metod, které vycházejí z určité skupiny algoritmů, byly zpočátku

vybrány reprezentace jednotlivých typů těchto algoritmů. V případě, že by jedna metoda dosahovala znatelně lepších výsledků než ostatní, by byly následně uplatněny i jí podobné verze a kódy, pokud by jejich uplatnění mohlo vést k lepším výsledkům. Z analýzy jednotlivých typů byly vybrány tyto metody reprezentující svojí skupinu:

- Naive Bayes - skupina založena na Bayesově teorému podmíněné pravděpodobnosti
- IBl - skupina založena na učení se pomocí jednotlivých instancí, tzv. lazy learning
- ADtree - skupina algoritmů založena na principu rozhodovacích stromů

Jako testovací protokoly jsem použil křížovou validaci pro dosažení co nejlepších výsledků.

5.4.1 Naive Bayes

Algoritmus Naive Bayes patří mezi nejpoužívanější a nejefektivnější metody strojového učení a je proto uplatňován i pro detekci cheaterů ve hrách [12]. Metoda je založena na podmíněné pravděpodobnosti jevů vyjádřené následujícím vzorcem, tzv. Bayesovým teorémem:

$$p(C|A_1..A_n) = \frac{p(C)p(A_1..A_n)}{p(A_1..A_n)} \tag{3}$$

Kde $p(C)$ je pravděpodobnost že hráč je cheater,$p(C|A_n)$ pravděpodobnost že hráč je cheater při určitém jevu $A_n, p(A_n|C)$ pravděpodobnost jevuA_n, pokud je hráč cheater a nakonec $p(A_n$ pravděpodobnost určitého jevu. Naivní bayes je pak charakteristický tím, že bere jednotlivé jevy jako nezávislé, takže daný vzorec lze přepsat do následující podoby:

$$p(C|A_1..A_n) = \frac{1}{X}p(C)\prod_{i=1}^{n}(A_i|C) \tag{4}$$

Kde X je konstantní hodnota závislá na počtu jednotlivých známých atributů a jejich pravděpodobnostního rozložení mezi třídami. Tabulka 6 ukazuje zjištěné výsledky pomocí hodnot precission a recall (viz sekce 5.2) při různých ověřovacích postupech.

5.4.2 IBl

IBl je základní ze skupiny instance base learning metod, které místo aby data generalizovala a až poté klasifikovala jednotlivé instance, porovnává každou novou instanci se všemi předchozími. Výhodou této metody je stabilita vůči šumu a nepotřeba provádět znovu učící algoritmus při nových datech. V podstatě se jedná o jednoduchou verzi metody nejbližšího souseda. Tento algoritmus byl vybrán, protože jednoduchým způsobem prověří, zda-li mají cheateři nějaké společné vlastnosti, podle kterých by je bylo možné identifikovat. Algoritmus klasifikace je založen na následujícím vzorci:

$$pobobnost(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)} \tag{5}$$

Kde n je počet atributů,x testovaná instance, y porovnávaná instance a $\sqrt{\sum_{i=1}^{n} f(x_i, y_i)}$ určuje euklidovskou vzdálenost těchto instancí přes množinu atributů. Testovaný subjekt je pak klasifikován do stejné skupiny jako instance s nejvyšší podobností. Tabulka výsledků pro algoritmus IBL (viz tabulka 7) je k nalezení mezi ostatními výsledky v kapitole 7.

5.4.3 AD tree

Alternate decision tree patří do skupiny klasifikátorů založených na vytváření rozhodovacích uzlů, které vedou k určení třídy. Na rozdíl od jednoduchého procházení stromu, který pročesává pouze jednu větev, AD tree sbírá informace přes všechny cesty a jednotlivé vlastnosti bere jako jistou pravděpodobnost, že daná třída tuto vlastnost obsahuje. Klasifikace je pak přiřazena třídě která je dané hodnotě blíže (při dvou třídách se jedná většinou o hodnoty $(-1, 1)$, přičemž čím bližší je skóre k dané klasifikační hodnotě, tím větší je jistota že výsledek je pravdivý. Příklad daného stromu, pro lepší vizualizaci je přiložen (obrázek 6) [2]. Váhy jednotlivých rozhodnutí jsou určeny na základě boosting algoritmů, konkrétně na algoritmu ADAboost [14], který určuje váhy jednotlivých cest podle efekti-

[2] převzato z [13]

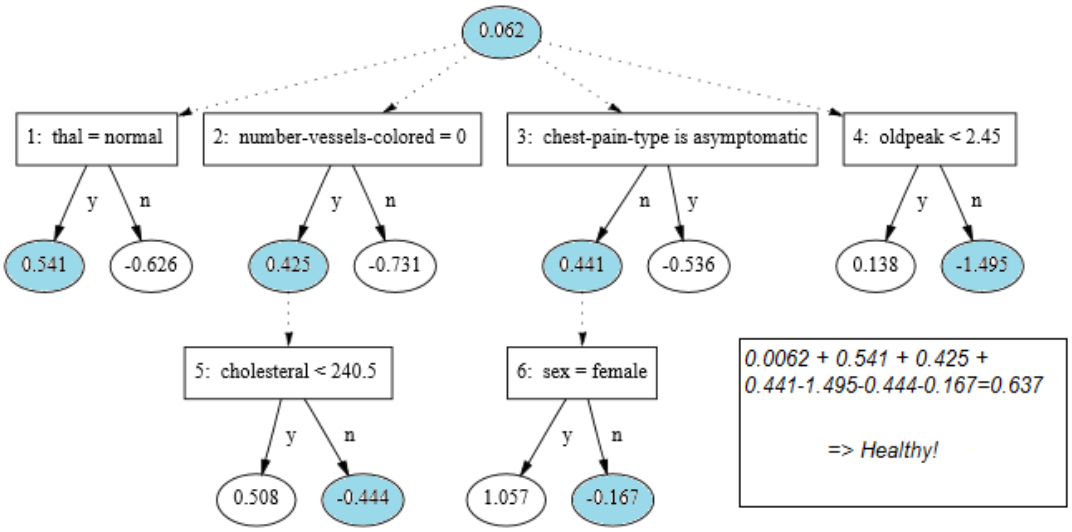


Figure 4: An alternating decision tree learned for the cleve data set (negative numbers correspond to predicted sickness).

Obrázek 6: Příklad vizualizace AD stromu

vity jejich klasifikace.Tabulka výsledků (8) AD rozhodovacího stromu pro různé iterace [3] při cross-validation:5 je přiložena v kapitole 7.

[3]Každá iterace zvedne počet uzlů stromu o 3 a počet cest o 1

6 Implementace vlastních metod

Implementace vlastního přístupu pro řešení problému hledání cheaterů ve vybrané hře se skládala z několika částí.

- Analýza dat a možných přestupků
- Vytvoření modelu aktivit možného cheatera
- Vytvoření vlastního kódu
- Aplikace kódu na reálná data

Vzhledem k tomu, že tato metoda je založena na specifických informacích poskytnutých společností Geewa a.s., jedná se tedy o metodu "ušitou"na míru hře, nebo také naivní metodu. Program bude založen na hledání vzorů v aktivitách jednotlivých uživatelů. Tato skutečnost bude brána v potaz při psaní programu, aby byla její specializace pro tuto hru co nejvíce minimalizována, tedy aby byl program co nejpřehlednější a nejsnáze modifikovatelný i pro jiné hry a logfile soubory podobné struktury.

6.1 Analýza dat a vytvoření modelu cheatera

Díky limitovanému množství informací poskytnutých v logfile souborech byly objeveny pouze podezřelé aktivity týkající se především prodeje a nákupu jednotlivých tág. Tento trend byl zpozorován díky grafům poskytnutých katedrou kybernetiky, která se hrou Geewa Pool Live Tour zabývá v rámci jiných problémů. Konkrétně se jedná o graf představující rozložení prodeje tág hráči v celém měsíci, který byl poskytnut pro analýzu (viz obrázek 7). Z grafu je vidět, že v jeden den došlo ke skokovému nárůstu prodeje tág, které se dá vysvětlit pouze dvěma jevy. Hra obdarovala hráče tágem, které bylo možné prodat, nebo došlo k nějaké formě podvodu způsobující tak vysoký prodej. Z dostupných informací bylo zjištěno, že Pool Live nevěnoval hráčům v tento ani předchozí den jakoukoli formu odměny

Figure 2: Bought cues over time.

Obrázek 7: Rozložení prodeje tág v rámci získaného měsíce

a proto můžeme předpokládat, že šlo o možný podvod. Na základě této informace bylo vypracováno několik modelů možného cheatera, které mohli nějakým způsobem souviset s danou anomálií a vycházeli z možných aktivit hráče - tedy z prodeje a nákupu tág:

- Hráč měl počet prodejů vyšší než n
- Hráč měl větší počet prodejů než nákupů za celý měsíc
- Hráč během jednoho dne prodal stejné tágo jako koupil

6.2 Vytvoření vlastního kódu

Před samotným uplatněním vlastního algoritmu byl navíc seznam podezřelých hráčů profiltrován tak, aby neobsahoval hráče kteří za celý měsíc prodali méně než dvě tága. Tento práh byl zvolen, aby se znatelně snížil počet zkoumaných subjektů a zároveň odfiltroval hráče, kteří mohli podvodu dosáhnout omylem jako je například překliknutí nebo interní

serverová chyba. Pro napsání algoritmu pak byl použit programovací jazyk Java rozšířený o java knihovny WEKA aplikace. Princip algoritmu pak stručně vypadá následovně:

```
Input: csv,logfiles,"variables"
/* Csv obsahuje zpracovaná data hráčů, logfiles jsou jednotlivé
   soubory, variables jsou proměnné pro modifikaci kódu          */
Output: Výpis ID podezřelých hráčů
begin
    while arff soubor obsahuje další ID do
        read ID file;
        while file obsahuje dalsi řádek do
            if řádek obsahuje "cue"string then
                if řádek obsahuje "cue-buy" then
                    označit tágo a datum koupě;
                if řádek obsahuje "cue-sell" then
                    if datum nákupu a prodeje tága souhlasí then
                        podezřelý prodej +1;
        if podezřelý_prodej > x then
            Zapamatovat ID hráče a počet podezřelých prodejů;
            cheater +1;
        smazat seznam označených tág a datumů;
    Vypsat cheatery podle nastaveného prahu;
```

Scenario 1: Pseudokód vlastního algoritmu, při hledání nadměrného nákupu a prodeje v časovém okně

Algoritmus postupně čte všechny soubory které jsou určeny pro analýzu. K jejich určení je použit .arff soubor vytvořený ve WEKA aplikaci a pro jejich obsluhu byla použita

knihovna weka. Důvodem je snadná modifikace pro filtraci souborů vzhledem k jejich velkému množství - kód obsahuje podmínku, která umožňuje filtraci podle hodnoty daného atributu. Po načtení souboru dochází ke zkoumání jednotlivých řádků, zda obsahují hledané atributy, v našem případě nákup a prodej tága (cue-sell, cue-buy). Algoritmus si ukládá do paměti informace o datumu prodeje jednotlivých tág (na základě jejich jména) a v případě shody datumu koupě i prodeje určitého tága je pro hráče připsaný bod za podezřelý prodej. Pokud počet těchto bodů přesáhne určenou hranici, je označen za potencionálního cheatera a jeho ID nahlášeno v konzoli pro analýzu a informace o tom, zda hráč patří mezi známé cheatery. Po projití všech souborů je vypsán počet potencionálních cheaterů i z toho zabanovaných pro porovnání.

6.3 Aplikace kódu na reálná data

6.3.1 Porovnání celkového množství prodeje a nákupu tág

Nejprve byla uplatněna zjednodušená verze algoritmu, která pouze zaznamenávala rozdíl mezi počtem prodaných a koupených tág za celý měsíc. Toto hledání mělo za úkol prověřit, zda-li se hráčům nějakým způsobem nepodařilo prolomit serverovou ochranu databáze hráčů, která obsahuje mimo jiné i data o jejich zakoupených tágách a uměle zvyšovat jejich množství. Tato metoda by byla jedním z možných vysvětlení vysokého nárůstu jejich prodeje (viz předchozí obrázek 7). Za práh podezřelého rozdílu byly zvoleny hodnoty mezi 1 - 15. Tyto hodnoty udávaly rozdíl mezi prodanými a koupenými tágy. Z tabulky 4 se

	Hodnoty														
Práh rozdílu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Počet identifikovaných	2938	1750	847	479	241	155	93	60	44	29	25	23	20	14	11

Tabulka 4: Tabulka podezřelých hráčů z podvodného prodeje

dá usoudit, že skutečně podezřelí by mohli být hráči, jejichž prodej předčil nákup o 5 tág, protože zde dochází k největšímu poklesu. Ovšem vzhledem k velice malému počtu takto podezřelých hráčů je možné, že se pouze jedná o hráče, kteří tato tága nakoupili v době,

ze které nejsou k dispozici jejich data. Pokud bychom ovšem měli k dispozici data za celou herní dobu daného hráče, mohlo by se jednat o jednu z použitelných metod jak identifikovat hráče, kteří obešli herní systém.

6.4 Analýza prodeje a nákupu stejných tág v časovém okně

Další postupem pak bylo prozkoumat nákup a prodej stejných tág (viz algoritmus 1). Smysl tohoto postupu je ten, že může poukázat na zneužití herní chyby, která mohla způsobit, že tága byla při koupi levnější než při jejich prodeji. Zneužití podobných chyb je bráno jako přestupek téměř ve všech online hrách, nehledě na jejich žánr. Časové okno tohoto prodeje bylo nastaveno na den, půlden a hodinu. Práh počtu takto prodaných tág byl nastaven na 4, abychom se vyhnuli započítávání hráčů, kteří pouze "zkoušeli"některé z tág, aby je po pár hrách opět prodali kvůli nevyhovujícím vlastnostem. Tabulky výsledků pro různé prahy a časová okna jsou zobrazeny v kapitole 7 (viz tabulky 9,10).
Z tabulky je vidět, že těchto hráčů je relativně hodně. Při analýze hráčů, jejichž prodej přesahoval 200 tág/den pak bylo zjištěno, že se skutečně jednalo o nákup tága a jeho prodeje za vyšší cenu. Mezi tato tága patřila převážně cue-bone a cue-x. Jejich nákup byl o polovinu levnější než jejich prodej, díky čemuž mohli hráči velice rychle zbohatnout. Jak je vidět, většina těchto hráčů nebyla postihnuta, především ani ti, kteří tuto herní chybu zneužívali nejvíce (ze 115 hráčů, jejichž nákup/prodej stejného tága za den překonal hranici 250 kusů, bylo zabanováno 6 hráčů. Hráč, který takto prodal nejvíce tág (přes 3000) si tímto způsobem vydělal přes 27000 herních mincí, což je v rámci hry velmi vysoké množství (hráčů s více jak 20 000 mincemi bylo pouhých 216 z 270 000). Nejedná se tedy o drobný podvod, ale v rámci hry o velice závažný.

7 Výsledky úspěšnosti jednotlivých algoritmů

Jak bylo zmíněno v kapitole 4, výsledky testování existujících algoritmů byly zaznamenány hodnotami precission a recall, které vychází z chybové matice (viz. šablona 3). Celkový počet zkoumaných instancí byl **271142** a z toho **622** tvořily cheateři.Zde jsou tabulky výsledků klasifikace jednotlivých algoritmů:

Klasifikační metoda	**Precission**	**Recall**
ε:0.8,p:4	**0.0112**	**0.0871**
ε:0.6,p:5	**0.0324**	**0.0101**
ε:0.5,p:6	**0.0126**	**0.0523**
ε:0.4,p:7	**0.0132**	**0.0366**

Tabulka 5: Výsledky algoritmu DBSCAN, parametry viz. 5.3.1

Klasifikační metoda	**Precission**	**Recall**
Cross-validation:5	**0.220**	**0.012**
Cross-validation:10	**0.225**	**0.013**

Tabulka 6: Výsledky algoritmu Naive Bayes

Klasifikační metoda	**Precission**	**Recall**
Cross-validation:5	**0.202**	**0.015**
Cross-validation:10	**0.225**	**0.016**

Tabulka 7: Výsledky algoritmu IBL

Klasifikační metoda	**Precission**	**Recall**
Počet iterací:20	**0.0016**	**0.0909**
Počet iterací:18	**0.0016**	**0.0016**
Počet iterací:15	**0.0016**	**0.0833**
Počet iterací:10	**0.0032**	**0.6666**

Tabulka 8: Výsledky algoritmu AD rozhodovacího stromu

7.1 Shrnutí uplatnění existujících metod

Z výsledků zobrazených v předchozích tabulkách je patrné, že uplatnění existujících metod pro identifikaci cheaterů nepřineslo uspokojivé výsledky. Vzhledem k poměru správně identifikovaných podvodníků ku falešným můžeme říci, že aplikace těchto metod je za daných podmínek nepoužitelná. Důvodem může být nízké množství cheaterů ku poctivým hráčům, nedostatečné množství atributů způsobené omezeným množství poskytnutých informací v lofgile souborech, popřípadě špatný výběr testovaných algoritmů. Lákavou možností je, že identifikovaní hráči jsou nedetekovaní skuteční cheateři. Tato možnost je ovšem těžko ověřitelná. Posledním důvodem pak může být princip, podle kterého byli hráči společností identifikováni jako cheateři. Vzhledem k nedostatku informací na důvod banu jednotlivých hráčů v logfile souborech, se může jednat o identifikace založených na interních serverových ověření, která mohou vycházet z neznámých vlastností a akcí, nedostupných v těchto souborech. Kvůli nepřesvědčivým výsledkům metod učení s učitelem i bez něj, nebyly testovány algoritmy využívající jejich kombinaci, tedy semi-supervised metody(viz. 3.4), které z těchto metod vychází.

	Práh prodeje: 5 tág		Práh prodeje: 10 tág		Práh prodeje: 30 tág	
Časové okno	**Podezřelí**	Zabanovaní z podezřelých	**Podezřelí**	Zabanovaní z podezřelých	**Podezřelí**	Zabanovaní z podezřelých
Den	**418**	18	**323**	13	**260**	12
Hodina	**259**	11	**240**	11	**211**	10

Tabulka 9: Tabulka podezřelých hráčů z prodeje stejných tág, pro nižší prahy

	Práh prodeje: 50 tág		Práh prodeje: 100 tág		Práh prodeje: 200 tág	
Časové okno	**Podezřelí**	Zabanovaní z podezřelých	**Podezřelí**	Zabanovaní z podezřelých	**Podezřelí**	Zabanovaní z podezřelých
Den	**236**	11	**206**	10	**139**	6
Hodina	**184**	9	**136**	6	**94**	6

Tabulka 10: Tabulka podezřelých hráčů z prodeje stejných tág, pro vyšší prahy

7.2 Shrnutí implementace vlastního kódu

Jak vidíme z tabulek (9,10), rozdíl nalezených hráčů při relativně velkém zmenšení časového okna (na 1/24) není nijak signifikantní. Rozdíl navíc může být navíc částečně způsobený tím, že hráč nakupoval na rozmezí dvou hodin, což vzhledem ke zvolenému oknu nebyl algoritmus schopný rozeznat. To odpovídá hypotéze, že inkriminovaní hráči skutečně v krátkém čase nakupovali a prodávali velké množství tág pro získání finanční výhody v rámci hry. Vlastní algoritmus byl proto úspěšnou metodou, jak detekovat možné podvodné hráče, které nebyli samotnou společností identifikováni. Důvodem je příprava kódu, vycházející z analýzy dostupných dat a nalezení podezřelých aktivit přes data mining dostupných souborů. Přestože je metoda na první pohled neaplikovatelná pro jiné hry, snadná modifikovatelnost kódu pro logfiles založené na stejném (běžném) formátu tento problém sleduje. To, že jsou hráči identifikovaní programem skutečně podvodníky v případě prodeje/nákupu tág za den (viz. 6.4), bylo ověřeno přes jejich logfile soubory, které

obsahovali dostatek informací pro potvrzení těchto výsledků, resp. záznamy o nakupování a prodeje stejných tág za účelem zbohatnutí (cue-sell a cue-purchase řádky jasně ukazovaly rozdíl v ceně nákupu a prodeje tága).

8 Závěr

Podvádění v online hrách je v dnešní době stále se vyvíjející problém, který může vést k vysokým finančním ztrátám společnostem,které dané hry vlastní, frustraci uživatelů a k negativnímu vlivu na budoucí vývoj online her. Vzhledem ke své různorodosti je v podstatě nemožné tento jev eliminovat. Vlastnoruční hledání cheaterů také bývá časově náročné a neefektivní, proto je nutné snažit se problém generalizovat a hledat řešení, která lze uplatnit na široké pole herních podvodů. Mezi tato řešení patří uplatnění detekce cheaterů na základě algoritmů využívající strojového učení a serverových informací poskytnutých herními společnostmi.

Cílem této práce byla analýza nejmodernějších metod, která se dnes používají pro detekci cheaterů a vybrání nejvhodnějších kandidátů pro uplatnění na online hře Geewa Pool Live Tour, od které byly k dispozici data aktivit týkající se všech registrovaných hráčů v rámci jednoho měsíce. Tato data se stala základem pro testování existujících algoritmů využívající strojového učení i pro vytvoření vlastní metody, která by byla schopná odhalit potencionální herní cheatery na základě vytvořeného modelu z analyzovaných dat. Posledním cílem bylo diskutovat výsledky uplatněných metod a navrhnout kroky pro budoucí zlepšení detekce uplatněných metod.

Všechny jednotlivé cíle této práce byly splněny. Byly vybrány metody z různých oblastí strojového učení využívající učení bez učitele i s učitelem a ty byly poté uplatněny na reálná data poskytnutá společností Geewa a.s. Tyto metody se ukázali jako vysoce neefektivní,pravděpodobně z důvodu náhodného rozprostření podvodníků mezi běžnými uživateli při dané reprezentaci (viz obrázek 4). Algoritmy nebyly schopni nalézt nějaký vzor, podle kterého by byly schopné cheatery s jistotou identifikovat a oddělit od poctivých hráčů. Dalším možným vysvětlením je nedostatečný počet atributů charakterizující podvodnou aktivitu. Hra tak patrně není zaplavena automatickými boty, které hrají hru za hráče, nebo nebyli společností identifikování v takovém množství, aby je strojové učení efektivně využilo.

Kromě užití těchto metod pak byl navržen i vlastní algoritmus, který hledal potencionální

podvodníky na základě vysokého prodeje a nákupu jednotlivých tág v krátkém časovém okně. Toto řešení úspěšně nalezlo hráče, kteří využili herní chyby ke svému obohacení a patřili tak mezi cheatery. Úspěch této metody je o to větší, že tito hráči nebyli samotnou společností odhaleni jako podvodníci. Algoritmus byl napsán tak, aby byl co nejsnáze modifikovatelný pro jiné hry používající podobný formát výstupních dat o jednotlivých hráčích.

Cíl práce byl tak splněn s lepšími výsledky než se předpokládalo, protože byly nalezeny nové vzory, pomocí kterých by komunitní manažeři hry mohli nalézt zatím neobjevené cheatery, pokud se daná aktivita zvýší nad určitý práh.

Zlepšení efektivity jednotlivých metod v rámci dané hry by vyžadoval podrobnější informace zprostředkované společností. Museli by se tedy upravit výstupní logfile soubory aby obsahovaly více informací o aktivitách hráčů, jako například přesnost jednotlivých šťouchů, nebo z jakého důvodu došlo k jejich banu, díky čemu by bylo možné upravit jeho proměnné pro nalezení většího množství potencionálních cheaterů.

Co se týče existujících algoritmů,jako jsou metody učení s učitelem a bez učitele, v případě Geewa Live Pool Tour nevidím jejich využití jako reálné. Jednotlivý cheateři nesdílejí tolik podobných rysů, na základě kterých by byli pomocí těchto postupů odlišitelní od ostatních hráčů, což vede k nepoužitelnosti metod na tomto principu. Zlepšení efektivity by bylo možné leda při vyšším množství identifikovaných hráčů, ve kterých by bylo možné nalézt určitý vzor. To je ovšem neovlivnitelný faktor. Algoritmy založené na strojovém učení proto pro tuto hru nedoporučuji.

Efektivita vlastního algoritmu je ovlivněna analýzou poskytnutých dat. Pokud by se podařilo nalézt nějakou další aktivitu, která by mohla být označena za podezřelou, jistě by se našlo více potencionálních cheaterů. Algoritmus by šel také rozšířit o prohledávání všech nadstandardních hodnot hráče a jeho následném označení. Společnost by tak mohla velmi rychle nalézt možné podvodníky.

Reference

[1] Forbes magazine. Earnings flash: Activision blizzard q3 'better than expected' - outlook raised, 2013.

[2] Stuart Bishop. Tabula rasa: financial disaster, 2008.

[3] Forbes magazine Andy Greenberg. Hacker attack 'kills' thousands in world of warcraft, 2012.

[4] Li-Wen Hong Kuan-Ta Chen. User identification based on game-play activity patterns, 2012.

[5] Andrew W. Moore. Clustering with gaussian mixtures, 2004.

[6] Jörg Sander Xiaowei Xu Martin Ester, Hans-Peter Kriegel. A density-based algorithm for discovering clusters in large spatial databases with noise. In Germany Oettingenstr. 67, D-80538 München, editor, *Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 1996.

[7] Nick Lalone. Dayz: Thousands of players banned over scripting drama, 2012.

[8] M.S. Sisodia Dilip Kumar Ahirwar, Sumit Kumar Saxena. Anomaly detection by naive bayes a rbf network. *International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 1, Issue 1 March 2012*, 2012.

[9] Andrea Villanes. Analytical approach for bot cheating detection in a massive multiplayer online racing game. Technical report, North Carolina State University, 2013.

[10] Marius Kloft Nico Görnitz. Toward supervised anomaly detection. 2013.

[11] Jerome H. Friedman and Usama Fayyad. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1:55–77, 1997.

[12] S. F. Yeung, John C. S. Lui, Jiangchuan Liu, and Jeff Yan. Detecting cheaters for multiplayer games: Theory, design and implementation. Technical report, 2005.

[13] Yoav Freund. The alternating decision tree learning algorithm. 2002.

[14] Robert E. Schapire Yoav Freund. A short introduction to boostin. *Journal of Japanese Society for Artificial Inteligence*, 1999.

Přílohy

Název souboru	**Popis**
Adam Ficenec bp	Bakalářská práce v pdf formátu
Data	Soubor obsahující instance všech hráčů přes škálu atributů ve formátu arff
BP program	Spustitelný Java program v JAR formátu
Lib	složka obsahující knihovny pro spuštění java programu
Vlastni kód	Vlastní algoritmus v txt formátu
Chybí logfile soubory	Textový soubor obsahující odůvodnění pro nepřiložení logfile souborů

Tabulka 11: Obsah CD

Tabulka akcí z kapitoly 4, oddíl 3 - extrakce dat (4.3), je zobrazena na následující stránce

Název akce	Stručný popis	Sekundární atributy akce
lobby-start	Otevření herního lobby	X
load game	Načítání hry	X
check-version	Zkontroluje platnou hodnotu verze hry	X
connected	Oznámení připojení k serveru	X
init-texts	Inicializace textu	X
local-currency-ready	Inicializce herní měny	X
tcp-enabled	Ověření tcp připojení	X
init-view	Inicializace okna	X
user-login	Příhlášení uživatele	X
event-ready	Akce připravena k exekuci	X
invitation-info-ready	Žádost o hru odeslána	X
add-coins	Přičtení herní měny při nákupu	Množství měny
earn-coins	Připsání herní měny za turnaj	Množství připsané měny
get-packages	Připravení balíčku s herní měnou k odeslání hráči	X
coin-packages-ready	Balíček s měnou zakoupen	X
guide	zapnutí/vypnutí podpůrného míření	ON/OFF
quality	Nastavení kvality herního vzhledu	low/mid/high
sound	Zapnutí/vypnutí zvuků	ON/OFF
music	Zapnutí/vypnutí hudby	ON/OFF
ban	Zabanování hráče, pouze s nejasnou zprávou flash-cheat-memory bez dalších údajů	X
coins-remove	Odstranění hráčových mincí	Množství mincí
practice	Zapnutí trénovací hry	X
menu-enabled	Spuštění menu	X
session-start	Spuštění hry	Počet hráčových mincí
settings	Spuštění okna s herními možnostmi	X
dialog-auto-open	Otevření pop-up okna s různými informacemi	rank-up/down, out of coins, daily bonus atd.
daily bonus	Přičtení denního bonusu za přihášení	Množství připsané měny
dialog-close	Zavření pop-up okna	Typ okna: rank-up/down, out of coins, daily bonus atd.
add-gold	Přičtení speciální herní měny	Množství připsané měny
shop	Otevření okna pro nákup herních předmětů	X
shop-[podtyp]	Výběr druhu zboží z menu obchodu a jejich nákup/prodej	recommended,nation,cues,buy,buy-confirm,sell,sell-confirm,recharge,avatar
match-start	Začátek kulečníkové hry	Friendly/Play and win
shot	Šťouchnutí koule pomocí tága	generická data viz 1 včetně ID tága
fps	Navrací hodnotu FPS	Hodnota FPS
returnball	Vrácení bílé koule na stůl, pokud spadla do díry	X
match-end	Ukončení zápasu	X
cue—ID	Informace týkající se určitého tága. Může jít o prodej/nákup/počet střel/dobití atd	sell,buy,recharge,stats
match-summary	Shrnutí výsledku zápasu	match-lose, match-win + množství získaných/ztracených mincí
rematch	Potvrzení odvety	FALSE/TRUE
socket-ping-timeout	Výpadek socketu	X
socket-close	Uzavření spojení socketu	X
tcp-rejoin	Opětovné navázání TCP spojení	X
chat	Akce týkající se komunikace mezi hráči	expand,collapse,quick message, message

Tabulka 12: Tabulka možných akcí v logfile souboru