# Review 27: Meta-AAD: Active Anomaly Detection with Deep Reinforcement Learning

**Paper: https://arxiv.org/abs/2009.07415v1**

פינת הסוקר:

המלצת קריאה מעדן ומייק: מומלץ לאנשים שאוהבים Anomaly Detection ו-Reinforcement Learning

בהירות כתיבה:

ידע מוקדם: Reinforcement Learning, Online Algorithms

יישומים פרקטיים אפשריים: SOC - Security Operation Center System

פרטי מאמר:

פורסם בתאריך: 16.9.20, בארקיב.

הוצג בכנס: ICDM 2020.

תחומי מאמר:

זיהוי אנומליות

למידה באמצעות חיזוקים (RL - Reinforcement Learning)

כלים מתמטיים, מושגים וסימונים:

למידה באמצעות חיזוקים (Reinforcement Learning)

תהליך החלטות מרקובי (Markov decision process)

Proximal Policy Optimization - PPO

למידה עמוקה באמצעות חיזוקים (DRL -Deep Reinforcement Learning)

מבוא:

תחומים רבים נעזרים בטכניקות לזיהוי אנומליות כדי לפתור בעיות שונות, למשל בתחום הסייבר - זיהוי התקפות סייבר, בתחום הרפואה - זיהוי גידולים סרטניים וכדומה. זיהוי אנומליות הינו משימה שבה המודל מתבקש לזהות דפוסים חריגים במידע, כאלו השונים באופן מובהק מרוב המידע. גישה זו לרוב מיושמת על ידי אלגוריתמים לא מפוקחים (Unsupervised). לפעמים קיים דאטה מתויג וניתן למנף אותו לזיהוי אנומליות ספציפיות. אלגוריתמים לזיהוי אנומליות מניחים הנחות על דפוסי ההתנהגות של אותו מידע חריג.

כל שיטה לזיהוי אנומליות מתבססת על סט הנחות המגדירות מהי אנומליה. למשל LOF מניחה כי מידע תקין יהיה לרוב מרוכז באשכול, כלומר כל נקודה תהיה קרובה לנקודות אחרות. לעומת זאת, אנומליה שונה משאר המידע ולכן תהיה באזור פחות מרוכז, שהאלגוריתם יוכל לזהות. הנחות אלו עלולות להביא לחוסר התאמה בין מה שהאלגוריתם מגדיר בתור אנומליה לבין מה שהמשתמש באלגוריתם מגדיר כאנומליה.

מערכת העושה שימוש באלגוריתמים כאלו עלולה לסבול מאחוז גבוה של False Positive - FP. לכן, לעיתים קרובות שיטות לזיהוי אנומליות משתמשות בגורם אנושי הבוחן את הדוגמאות שהוגדרו כאנומליות, ומחליט האם מה שהוגדר כאנומליה, אכן הוא אנומליה. לצורך כך, נשאף למקסם את אחוז האנומליות שהאלגוריתם ישלח לתיוג במסגרת תקציב שנקבע מראש. זאת בשונה מסתם שליחה של דוגמאות אקראיות בניסיון לשפר את זיהוי האנומליות. ההבדל הוא שבגישה זו, חשוב לחשוף את האלגוריתם גם לדגימות תקינות בנוסף לאנומליות, כדי שיוכל להפריד בין שני הסוגים.

עם זאת, שימוש בגורם אנושי הינו יקר ולכן נרצה לנצל את זמנו בצורה אופטימלית, כלומר לא לבזבז אותו על דוגמאות אשר יתבררו כ-FP. כאן באה לידי ביטוי הבעיה שכותבי המאמר מנסים לפתור: מזעור אחוז ה-FP, או לחלופין מקסום אחוז ה-TP, בקרב האנומליות שהועברו לבוחן. ההנחה כאן כי תהליך זיהוי אנומליות מערב בוחן אנושי אשר יסווג את האנומליות מתוך רשימה של אנומליות המופקת על ידי המערכת באופן אוטומטי, והמערכת תחשב ציון אנומלי (anomaly score) לכל דגימה בהתבסס על המשוב מהבוחן. בתורו, הבוחן ינסה לחקור כמה שיותר אנומליות במסגרת זמן קצובה ("התקציב") ולזהות אלו מהן הן אכן אנומליות, ואילו לא. סיטואציה זו פותחת דלת לתרחיש מעניין שבו המערכת עשויה למנף את הפידבק מהגורם האנושי באופן הבא:

אנומליות הדומות לאלו שהבוחן סימן כאמיתיות, יקבלו ציון גבוה יותר באיטרציה הבאה וכך יגדל הסיכוי שלהן להיות מזוהות כאנומליות.

דוגמאות שסומנו כתקינות על ידי הבוחן, יקבלו ציון נמוך יותר שיוריד את הסיכוי של דוגמאות דומות להישלח לתיוג להבא.

קיימות שיטות שבוחרות את הדוגמאות בעלות הפוטנציאל הגדול ביותר להיות אנומליות. האנליסט מתייג דוגמאות אלו, האלגוריתם מפיק מהן ידע, שוב מריצים אותו על הדאטה ובוחרים דגימות נראות אנומליות, וחוזר חלילה עד תום תקציב התיוג.

דוגמה ויזואלית ניתן לראות כאן בציור הבא:

שיטות אלו אכן מסוגלות להביא לביצועים טובים בטווח הקצר, אך עלולות לסבול מביצועים לא אופטימליים בטווח הארוך. הסיבה לכך היא קיום דוגמאות עם ציון אנומליה נמוך, אך שמגלמות בתוכן אנומליות מסוג שונה שעדיין לא נחקר (דוגמה לכך ניתן לראות כאן בפרק 3.1). בטווח הקצר השיטות הקיימות פועלות באופן חמדני, כלומר בוחרות מבין האנומליות את אלו עם הציון הגבוה ביותר, ולכן אנומליות בעלות ציון בינוני עשויות לא להישלח לתיוג כלל. כותבי המאמר טוענים שאם כן נתחשב בדוגמאות אלו, הן עלולות לגרום לתוצאות פחות טובות בטווח הקצר, אך בטווח הארוך הן יגרמו לאנומליות חדשות ומעניינות להיבחר גם כן. כך הגורם האנושי יוכל לחקור אנומליות שאחרת לא היה רואה, שכן השיטות החמדניות היו מראות לו את אותן אנומליות שוב ושוב. תרחיש זה הוא שהביא את החוקרים לחקור שיטות שמתייחסות לטווח הארוך.

מידול הטווח הארוך טומן בו מספר אתגרים:

בדרך כלל מרחב החיפוש (= סט כל הדוגמאות) הינו עצום, ונצטרך לעבור על כולו כדי לבחור את הדוגמה שתמקסם את הביצועים בטווח הארוך.

כל דאטהסט הינו שונה ויהיה קשה למצוא אסטרטגית בחירת דוגמאות גנרית עבור כולם.

ניסוח הבעיה:

כדי לנסות לפתור את האתגרים הנ"ל, כותבי המאמר מציעים את שיטת MetaAAD. הבעיה שאנו מנסים לפתור הינה בחירת הדוגמאות עם ההסתברות הגבוה ביותר להיות אנומליות תחת תקציב מסוים T, מספר הדוגמאות שהבוחן מתייג. במאמר המסוקר, הבעיה מנוסחת באופן הבא: עבור דאטהסט המכיל דוגמאות ו- המסמן את מצב התיוג של דוגמא . מצב התיוג , כאשר 1- מייצג דוגמא שנשלחה לתיוג וסומנה כאנומליה, 1 מסמן דוגמא לא אנומלית ו-0 מסמן דוגמא לא מתויגות. בהתחלה, כל הדוגמאות מאותחלות עם , כלומר אף דוגמא לא מתויגת. בכל שלב, דוגמא אחת תבחר לתיוג, כלומר מצבה ישתנה מ-0 ל- בהתאם לתשובת הבוחן. בהינתן תקציב T שאילתות שניתן לשלוח לבוחן, מטרתנו הינה ללמוד מדיניות לבחירת דוגמא שתישלח לתיוג בכל איטרציה, במטרה למקסם את מספר הדוגמאות שתויגו כאנומליה במסגרת התקציב T.

הסבר על מושגים רלוונטים:

כותבי המאמר ממדלים את הבעיה בתור MDP (ראה נספח בסוף הסקירה). המטרה של MDP הינה ללמוד את האסטרטגיה הממקסמת את התגמול הכולל, אך איך נמדדת המדיניות הטובה ביותר? במאמר זה התועלת נוסחה כתגמול המצטבר המוזל הצפוי:

כאשר 1 > > 0 הינו פקטור ההזנחה (discount factor), המבטא ירידת ערכו של התגמול t צעדים אחרי פעולה פי פקטור של . פקטור זה הינו קבוע שמוגדר מראש.

הרעיון הכללי מאחורי השיטה:

כאשר מאמנים שיטה על דאטהסט מסוים, לא ניתן להבטיח כי היא תעבוד על דאטהסט בעל מאפיינים שונים ממנו. כתוצאה מכך יש צורך לבנות מטא-מאפיינים מספיק גנרים ורובסטיים כדי לאפיין אנומליות בדאטהסטים רבים. נזכיר כי מטרתנו הינה למצוא את מספר האנומליות הגדול ביותר במסגרת תקציב תיוג T ונרצה להשתמש במטא-מאפיינים שיתרמו לנו למטרה זו. החוקרים השתמשו במאפיינים הבאים:

מאפייני גלאים (Detectors): ציוני האנומליה שחושבו על ידי גלאי אנומליה מבוססי למידה לא מפוקחת. כל אלגוריתם גילוי אנומליות בסיסי יכול לספק לנו מאפיין שכזה.

מאפייני אנומליות: מאפיינים אלו מתארים עד כמה דוגמא ״דומה״ לאנומליות המתויגות בדאטהסט (ככל שהשיטה תרוץ יותר זמן ונקבל יותר תיוגים מהבוחן, מאגר האנומליות המתויגות יגדל). תחת קטגוריה זו נכללו 3 מאפיינים:

המרחק האוקלידי המינימלי והממוצע מהדוגמא לדוגמא אנומלית (2 מאפיינים)

מאפיין בינארי (1 או 0) המציין האם לדוגמא קיים שכן שהינו אנומליה מתוייגת כאחד מתוך K השכנים הקרובים.

מאפייני נורמליות: מאפיינים אלו זהים למאפייני האנומליה (מינימום, ממוצע ושכנים), רק שבמקום להסתכל על האנומליות המתויגות, ניקח את הדוגמאות שתויגו כתקינות.

הערות:

החוקרים מציינים שהשיטה מספיק גמישה כך שיהיה אפשר להוסיף/להחסיר מטא-מאפיינים למודל בצורה פשוטה ולבדוק כיצד זה משפיע על הביצועים.

יש צורך בחישוב המרחקים בין כל זוג דוגמאות בדאטהסט לפני תחילת האימון כדי שניתן יהיה לחשב את מאפייני האנומליות והנורמליות.

שימוש בזרם מידע:

וקטור המכיל את ששת המאפיינים האלו יציין לנו את המצב s של דוגמא. מפה עולה נקודה חשובה - כאשר מדובר באלגוריתם RL קלאסי (למשל Q-Learning, Value Iteration), יש צורך להחזיק טבלה בגודל מרחב המצבים כפול מרחב הפעולות. כל תא בטבלה ייצג פעולה ומצב. האלגוריתם שומר בכל תא עד כמה כדאי לו לבצע את אותה פעולה כאשר הוא נמצא במצב מסוים. נשים לב לכמה דברים: קודם כל, כל מצב מיוצג על ידי וקטור בגודל 6 ולא על ידי מספר בודד. שנית, מספר המצבים תלוי בגודל הדאטהסט שכן כל דוגמא מהווה מצב. כתוצאה מכך יש לשמור בכל רגע בזיכרון טבלה שיכולה להיות גדולה מדי לזיכרון החומרה. כדי להתמודד עם הבעיה, החוקרים משתמשים בזרם (stream) של מידע, כך שבכל איטרציה האלגוריתם יפגוש רק דוגמא אחת מכל המרחב ולא את המרחב כולו. כעת נתאר את הדרך שבה זה נעשה:

ניצור דאטהסט שבו כל דוגמא הינה המטא-פיצ'רים והתיוג של דוגמא מהדאטהסט המקורי, כלומר נהפוך את הבעיה לבעית למידה מפוקחת על ידי יצירת צמדים הבאים:

כאשר הינו מצב הכולל את המטא-פיצ'רים ו- הינו התיוג של אותו מצב (אנומליה או לא). בכל איטרציה נבחר בצורה רנדומלית את הדוגמא הבאה מהדאטהסט ונזין אותה למודל.

כמו שיטות למידת חיזוקים עמוקה רבות, MetaAAD משתמשת ברשת נוירונים כדי למדל את על מנת להימנע משמירת הטבלה המלאה המכילה את כל המצבים. ככל שהרשת ״תראה״ יותר פעולות ותקבל פידבק מהסביבה בדמות תגמולים, יכולת שערוך תשתפר וכך תוכל לתת מידע מדויק יותר עבור הסוכן שתוביל לשיפור המיוחל של המדיניות שלו.

MetaAAD מנסה למקסם את פונקצית המטרה הבאה:

הלוס מורכב מהחלקים הבאים:

- פונקציית הלוס של שיטת PPO (לפרטים, ראה נספח), כאשר רשת ה-״שחקן״ של PPO מקבלת כקלט מאפייני דוגמא, ומוציאה כפלט אומדן להסתברות שליחתה של הדוגמא לתיוג. בעזרת אותו אומדן דוגמים את הפעולה.

- מודדת את דיוק השערוך של רשת המבקר.

entropy - מעודד למידת מדיניות פחות דטרמיניסטית, כלומר בעלת אנטרופיה גבוהה יותר. זה מעודד את הסוכן לחקור יותר את מרחב הפעולות (exploration), תוך כדי בחירה של פעולות שכרגע תגמולן הצפוי אולי נמוך. כך לסוכן יש פחות סיכוי להיתקע במקסימום לוקאלי של ״מרחב התגמולים״. מנגנוני רגולריזציה דומים נמצאים בשימוש נרחב באלגוריתמי מדיניות (policy) של RL.

שניהם היפר פרמטרים שניתן לכוון

אימון מול מצב אמת:

במהלך פרק של אימון הסוכן יכול לבצע T פעולות, שכל אחת מהן היא שליחה או אי שליחה של דוגמא לתיוג. בזמן תהליך האימון יש צורך בדאטהסט מתויג. הסיבה לכך היא שבזמן תהליך הלמידה של הסוכן, הפידבק שיוחזר לו מהסביבה, זה שאמור להיות הפידבק מהגורם האנושי, יוחלף בתיוג האמיתי של אותה דוגמא שנבחרה להישלח לחקירה.

במצב אמת בו משתמשים בדאטהסט לא מתויג, מתבצעים השלבים הבאים:

עבור כל דוגמא בדאטהסט מחשבים את המטא-פיצ'רים ומאתחלים את מצבה כ-0 (הדוגמא לא נבחרה לתיוג), כלומר הצמד

ההסתברות שליחה לתיוג מחושבת לפי פונקציית המדיניות עבור כל דוגמא בדאטהסט לבדיקה (כלומר פעולה מספר 1).

הדוגמא בעלת ההסתברות הגבוה ביותר נשלחת לבדיקה. לאחר הפידבק של הבוחן, הסוכן מעדכן את המדיניות שלו ואת המצב להיות תוצאת התיוג.

התהליך חוזר על עצמו עד שהתקציב T נוצל במלואו.

שרטוט ויזואלי של התהליך:

הישגי מאמר:

כדי לבדוק את השיטה שלהם, החוקרים השתמשו ב-24 דאטהסטים בגדלים שונים. תקציר של הדאטהסטים ניתן לראות כאן:

החוקרים השוו מול כמה שיטות שונות:

AAD: אלגוריתם שבאותה תקופה היה נחשב לטוב ביותר בגילוי אנומליות. האלגוריתם משתמש במספר גלאים (detectors) ויודע למשקל כל גלאי בהתאם "ליכולתו לזהות אנומליות״. AAD מערב את הגורם האנושי ומשתמש בפידבק שלו לעדכון המשקלים ש הגלאים בהתאם לפידבק.

FIF: שיטה שמשתמשת בפידבק האנליסט כדי לעדכן את אלגוריתם Isolation Forest - IsF

SSDO: משלבת גלאים שונים בסביבה מפוקחת חלקית (semi-supervised).

שיטות לא מפוקחות קלאסיות לגילוי אנומליות כגון IsF. שיטות אלו לא משתמשות בפידבק מהגורם האנושי ולא מתעדכנות בזמן ריצה בהתאם לפידבק. שיטות אלו הן בעצם הבסיס המינימלי שצריך להיות מעליו.

הגלאי בו השתמשו חוקרי המאמר בשביל לחלץ את מאפייני הגלאים הינו IsF.

ניתוח תוצאות:

12 הדאטהסטים הראשונים ברשימה המוצגת לעיל נבחרו להיות סט האימון והשאר נבחרו להיות סביבת האמת שעליהם נבדקה השיטה. המטריקה שנבחרה כדי לבצע בדיקה לטיב השיטה הינה שמציירת על גרף את כמות האנומליות שהתגלו מול מספר השאילתות שהשתמשו כדי לגלות אותם. גרף עם שיפוע 1 מראה שכל השאילתות הביאו לגילוי אנומליות בעוד גרף עם שיפוע 0 מראה שבו כל השאילתות הביאו לגילוי דוגמאות רגילות ולא אנומליות. נרצה שהשיטה תביא לגרף עם שיפוע קרוב ל-1. התקציב T של שאילתות בכל פרק בו השתמשו להרצת הניסויים הינו 100:

עבור כל שיטה נבדק הדירוג הממוצע של האנומליות שנמצאו לאורך מספר שאילתות שונה:

מהתוצאות ניתן להסיק כמה מסקנות:

ניתן לראות שכלל האלגוריתמים, המשתמשים בפידבק מהמשתמש, מראים תוצאות טובות מהאלגוריתם הלא מפוקח הבסיסי לכל הדאטהסטים. מכך ניתן להבין את החשיבות שיש לשימוש בפידבק כדי להמשיך ולאמן את האלגוריתם באופן מכוון ובכך לאפשר לו להשתפר לאורך זמן.

SSDO שהוא אלגוריתם חצי מפוקח מראה תוצאות פחות טובות מאשר האלגוריתמים המשתמשים בפידבק בכל הדאטהסטים. הסיבה לכך לפי כותבי המאמר הינה פונקציית המטרה. SSDO מנסה להגיע לאופטימום פונקציית מטרה שונה מאשר האלגוריתמים המשתמשים בלמידה אקטיבית. זה יכול לגרום ל-SSDO להגיע לאופטימום מקומי אשר פחות טוב מאלגוריתמי הפידבק.

MetaAAD מראה תוצאות שהן לפחות טובות כמו שאר אלגוריתמי הפידבק הטובים ביותר. סיבה אפשרית לכך היא שב-MetaAAD המידול של הטווח הארוך נעשה באמצעות טכניקות RL. בנוסף, טבלת הדירוג מראה ש- MetaAAD מגיע לביצועים טובים יותר ככל שמספר השאילתות עולה הינו הוכחה לכך.

הערה: לא ברור מהמאמר כיצד בוצע הדירוג.

סיכום:

במאמר הוצגה שיטת MetaAAD לזיהוי אנומליות, שיטה אשר משתמשת ב-DRL כדי למקסם את מספר האנומליות שניתן למצוא בדאטה. השיטה יודעת להשתמש בגלאי אנומליות אשר כבר קיימים בארגון/מערכת ובהבנה שלהם מהי אנומליה. בנוסף, על ידי עירוב הגורם האנושי בתהליך, השיטה מנצלת את הפידבק שלו כדי להשתפר לאורך זמן בצורה מקוונת. השיטה הוכחה כטובה לא פחות ואף יותר משיטות הנחשבות לטובות ביותר בתחום גילוי האנומליות.

נספחים:

תהליך קבלת החלטות מרקובי:

(MDP - Markov Decision Process) הינה מסגרת מתמטית המשמשת למדל בעיות קבלת החלטות שבהן התוצאות הן בחלקן אקראיות ובחלקן ניתנות לשליטה. מקבל החלטות, או סוכן, "חי" בסביבה, אשר משתנה באופן אקראי בתגובה לבחירות הפעולה (action) שנעשו על ידי הסוכן. ב-MDP, הסתברות המעבר והתגמול תלויים רק במצב הנוכחי ובפעולה שנבחרה על ידי הסוכן, ולא תלויה במצבי העבר ובפעולות העבר. מצב הסביבה משפיע על התגמול המיידי שמקבל הסוכן, כמו גם על ההסתברויות מעברים עתידיים בין המצבים. מטרת הסוכן היא לבחור אסטרטגיה π (סדרת פעולות) כדי למקסם את התגמול הכולל שהוא מקבל. האסטרטגיה π מגדירה את פעולת הסוכן בהינתן המצב שהוא נמצא בו.

MDP מאופיין על ידי הפרמטרים הבאים:

- סט כל המצבים האפשריים שניתן להגיע אליהם או בשם אחר מרחב המצבים.

- סט כל הפעולות האפשריות שניתן לבצע.

- ההסתברות לעבור ממצב למצב בעקבות ביצוע פעולה .

- התגמול המיידי על ביצוע פעולה ממצב ומעבר ל-

למידה עמוקה באמצעות חיזוקים:

שם כולל לאלגוריתמי למידה עמוקה בהם הרשת היא זו שקובעת את האסטרטגיה של הסוכן. בדרך כלל, משתמשים בסוכן מבוסס למידה עמוקה כדי ללמוד מדיניות עבור MDP.

Proximal Policy Optimization:

אלגוריתם RL המורכב משתי רשתות: רשת המבקר (critic) ורשת השחקן (actor):

השחקן - רשת זו אחראית להחליט איזו פעולה הכי כדאי לבצע במצב מסויים ולמעשה היא משערכת את המדיניות הסטוכסטית עבור דוגמא נתונה מהדאטה-סט.

המבקר - רשת זו אחראית לבצע קירוב של כדאיות המצב ובכך מידעת את השחקן על טיב הפעולה אותה בחר לבצע. הרשת מקבלת כקלט מצב ומוציאה כפלט מספר שמייצג בקירוב את פונקציית ערך המצב . פונקציית ערך המצב משערכת את התגמול הממוצע שניתן לקבל החל ממצב S.

הסברים יותר מעמיקים - כאן, כאן, וכאן.

נזכור ש-MDP ממודל על ידי הפרמטרים הבאים: . MetaAAD מגדירה אותם באופן הבא:

- המטא-מאפיינים של הדוגמא הנוכחית בזרם, דוגמא

- מרחב הפעולות הינו 1 או 0 כאשר 1 אומר לשלוח את הדוגמא לתיוגי ו-0 ממדל אי שליחה של דוגמא לתיוג .

- לא הוגדר במאמר. קיימים מודלים אשר מידע זה לא נתון בהם והם לומדים אותן תוך כדי אינטראקציה עם הסביבה. החוקרים לא הגדירו שזה המצב, אך לעניות דעתנו זה מה שקורה כאן.

- פונקציית תגמול

כאשר הסוכן החליט לא לשלוח דוגמא לתיוג, ניתן לו אוטומטית תגמול של 0

כאשר הסוכן החליט לשלוח הדוגמא תיוג: אם תויגה כאנומלית הסוכן מקבל תגמול של 1. אם הדוגמא תויגה כתקינה (כלומר היה עדיף לדלג עליה) התגמול הוא 0.1.