Departamento de Informática

Devido ao volume de dados gerados pelas aplicações e pelo sistema operacional, janelas deslizantes são aplicadas a fim de avaliar um ambiente on-line, permitindo que intrusões sejam desativadas em tempo real enquanto ainda estão sendo executadas. O respectivo estudo explora o impacto que o tamanho da janela de observação tem sobre os algoritmos de uma classe de aprendizado de máquina .

Introdução

Uma variedade de estratégias pode ser usada para adaptar os dados aos modelos, uma das mais usadas é a técnica de janela deslizante, onde uma janela de tamanho n é usada para escanear os dados. Muitos pesquisadores já tentaram definir o tamanho correto de uma janela de detecção, que pode variar de cinco a onze em muitos casos. Entretanto, como muitos parâmetros ML, o tamanho das janelas de detecção pode não ser sempre o mesmo, dado que uma escolha ótima de tamanho dependerá do ambiente e do tipo de dados . Assim, o IDS não funcionaria como esperado no mundo real, razão pela qual a técnica de janela deslizante é importante .

Em teoria, quanto menor o tamanho da janela, mais rapidamente uma anomalia poderia ser detectada, com o trade-off de que estas janelas podem cortar ações maliciosas pela metade, afetando sua detecção. Finalmente, selecionar um tamanho de janela ideal é importante para detectar ataques assim que eles começam a ser executados, sem a necessidade de observar o traço inteiro. Com o objetivo de explorar o impacto do tamanho da janela em métodos ML de uma classe, nós - deixamos um conjunto de traços que foram avaliados com tamanhos diferentes, em um ambiente portuário onde, no melhor de nosso conhecimento, ainda não foi feito na literatura. A idéia principal é identificar como o tamanho de uma janela impactará o processo de identificação de uma anomalia em um ambiente, ajudando a entender como o classificador se comportará com esses dados.

Trabalho Relacionado

A literatura apresenta abordagens para detectar o tamanho de janela mais adequado para os conjuntos de dados UNM e ADFA-LD , concluindo que um tamanho entre seis e sete era o melhor para eles. Considerando as máquinas virtuais, o tamanho de janela poderia diversificar entre seis e dez . O tamanho da janela terá impacto no tamanho n-grama criado e é responsável pela divisão do traço em grupos que serão usados para treinamento e testes. Introduziram o uso de chamadas de sistema para detecção de anomalias anos atrás, onde propuseram uma abordagem utilizando uma técnica baseada em um par de look-ahead, onde cada entrada no banco de dados representava uma chamada de sistema, e uma subseqüente imediata de chamadas de sistema em uma janela de tamanho n.

A janela deslizante então se move por uma posição de cada vez para formar o banco de dados. Uma janela deslizante de tamanho fixo é usada para produzir as seqüências curtas de chamadas do sistema. Os dados são coletados a partir de uma virtualização KVM ambiente através de ferramentas de rastreamento. É utilizado um tamanho de janela de seis, com 11,1% de falsos positivos.

Alguns problemas com o conjunto de dados poderiam ser levantados, pois o período de coleta de dados supõe que o comportamento normal não será afetado por atacantes.

Proposta

Observando como esses dados podem ser recuperados e aplicados nos métodos ML, nos concentramos em como o tamanho da janela terá impacto na detecção de anomalias. A virtualização acontece em nível de Sistema Operacional , diminuindo a carga e a execução de aplicações que têm hardware gerenciado pelo contai- ner . Nosso conjunto de dados1 foi desenvolvida capturando dados de um ambiente de contêineres, observando que a coleta acontece do ponto de vista do sistema operacional. Desta forma, não há visão parcial e limitações para a captura de dados .

Isto define um conjunto de dados com comportamentos normais e anômalos que representam um ambiente de con- maquiagem. Estes dados são usados para treinar e testar dois métodos ML de uma classe. A Classificação de Uma Classe aprende de apenas uma classe, diferente de outros métodos traditio- nal que visam aprender duas ou mais classes e pode ser difícil para o classificador considerando o conjunto de dados utilizado . O processo de avaliação reflete como o crescimento do tamanho da janela irá contribuir para a detecção de ano-maly.

Este caso consiste em utilizar apenas 5% do tamanho do traço para treinar e testar os modelos. No segundo caso, o crescimento da janela não será limitado a apenas 5% do tamanho do traço, em vez disso, o crescimento acontecerá a cada vez com 5% do tamanho do traço até que todo o traço seja usado.

Experimentos

Para o menor tamanho de janela, foi utilizada uma porção entre zero e 5% do tamanho total do traço, com um crescimento de 0,5% do tamanho do traço. Esta investigação representará a maioria dos casos encontrados na literatura , considerando que 5% do tamanho do traço representará 81 chamadas ao sistema. Nosso objetivo aqui não é ficar limitado por este tamanho, mas verificar como um tamanho de traço maior se comportará na segunda abordagem considerando o crescimento da janela até o tamanho máximo do traço. O tamanho do traço crescerá 5% do tamanho total até atingir 100% do tamanho do traço.

A avaliação considerou dois algoritmos de OCC: Isolation Forest e One-Class

Tanto os dados brutos quanto os dados do filtro atingem resultados razoáveis com janela pequena e variam sobre o tamanho do crescimento até atingir F1Score próximo a 100% depois que metade da porcentagem de traços é utilizada. A variação do tamanho da janela será responsável pela queda de 1% nos resultados, considerando que isto não é um grande impacto nos resultados, supomos que seria possível evitar este comportamento com um conjunto de dados mais complexo. Embora os dados do filtro sofram mais com o tamanho da janela de crescimento, ele é o primeiro a obter um F1Score acima de 97%, com um tamanho de janela de 1,5%do traço . Os dados brutos não sofrerão o mesmo impacto com o crescimento, sendo mais estáveis com a mudança de tamanho.

O gráfico mostra uma pequena variação entre os dados brutos e de filtro para um tamanho de janela com 10% do tamanho do traço, que impactará um F1Score entre 98% e 99%, nunca alcançando um F1score de 100%. Considerando os primeiros 5%, podemos ver um alcance mais rápido para o F1score onde uma janela com menos de 0,5% do tamanho do traço já alcança um F1score de 98%. Comparando ambos os resultados, é interessante apontar que uma pequena janela tende a atingir um F1score rapidamente, mas melhores resultados serão impactados pelo algoritmo usado e as abordagens ainda poderiam ter melhores resultados usando uma porção maior do traço em seu lugar de janela pequena. Em nosso caso de estudo, os dados brutos e filtrantes apresentam variações sobre o crescimento da janela, não mostrando uma diferença maior entre o uso de um tipo de dado específico.

Crescimento da janela para o algoritmo Isola- tion Forest.