Agrupamento de Usuários para Verificação de Viabilidade de Distinção Comportamental de Uso

Anais do Computer on the Beach

Marcelo Marques Ribas
marcelomarques@ufpr.br
Departamento de Informática
Universidade Federal do Paraná
Curitiba, Paraná, Brasil

Ulisses Penteado
ulisses@bluepex.com.br
Bluepex – Centro de P&D
Limeira, SP, Brasil

André Grégio
gregio@ufpr.br
Departamento de Informática
Universidade Federal do Paraná
Curitiba, Paraná, Brasil

Paulo Lisboa de Almeida
paulorla@ufpr.br
Departamento de Informática
Universidade Federal do Paraná
Curitiba, Paraná, Brasil

Abstract

Authentication mechanisms are still the standard way to allow access to systems and devices within an organization. Through credentials (login and password) and other associated methods (multi-factor authentication, such as tokens, biometrics, or one-time passwords sent to additional devices), access control is implemented, and user activity across different necessary systems is recorded. However, organizations are concerned that access control may be bypassed due to the loss or theft of authentication information/devices, potentially leading to intellectual property breaches through industrial espionage. In this context, User and Entity Behavior Analytics (UEBA) has been studied and applied to profile users and identify anomalous patterns that could, for example, block a user from accessing another account. However, achieving this level of protection in real-world systems may be unfeasible. This article examines the feasibility of distinguishing user behavior in organizations based on the most frequently used applications and their usage time. To this end, a real dataset was collected, consisting of data from over 700 organizations and nearly 60,000 users between March and September 2024. The results discuss the techniques used, the possibility of detecting real intruder users, and the false alarm rates observed in the dataset, paving the way for future research in the field.

Keywords

Comportamento de Usuário, Segurança, Agrupamento de Dados, Aprendizado Não Supervisionado.

1 Introdução

É esperado que usuários de uma mesma corporação, ou que trabalham em um mesmo segmento (ex.: pessoas que trabalham com contabilidade) utilizem um grupo em comum de softwares durante suas jornadas de trabalho. Por exemplo, espera-se que os engenheiros civis que trabalham em determinada organização utilizem um grupo de softwares que incluem planilhas eletrônicas e aplicações de Desenho Assistido por Computador – *Computer Aided Design* (CAD). Espera-se ainda que esses usuários utilizem as mesmas versões desses softwares, já que o trabalho em equipe exige que os arquivos compartilhados sejam compatíveis, e as licenças para uso de software geralmente são compradas em lote.

Um dos mecanismos mais comumente utilizados para prover um nível mínimo de segurança da informação em todos os tipos de organizações é a autenticação [1]. A forma mais simples de autenticação em um determinado sistema é a criação de credenciais (nome de usuário e senha), a fim de prover acesso ao mesmo. Além disso, há uma preocupação justificada sobre proteção de propriedade intelectual por parte de corporações que lidam com inovação. Embora existam inúmeros mecanismos de autenticação para proteção (por exemplo, *passphrases*, tokens físicos e biometria), é crescente o uso de soluções baseadas em aprendizado de máquina, como a *User and Entity Behavior Analytics* (UEBA) [2].

Soluções baseadas em UEBA tentam suprir uma lacuna amplamente conhecida, mas de difícil solução no âmbito da segurança e autenticação de usuários das organizações: ainda que se aplique múltiplas camadas de proteção contra entidades externas, pouco se pode fazer contra ameaças internas (também conhecidos como *insiders*). Apesar de existirem normas e políticas que, se aplicadas, diminuem o risco de ataques às custas da usabilidade, é interessante abordar o problema de uma corporação como um organismo que pode ser perfilizado. Se isso for verdade, considerando que os usuários comportem-se de maneira a se estabelecer um padrão em suas respectivas corporações, então seria possível identificar um ente externo (talvez adversário) atuando internamente em uma organização a qual ele não deveria ter acesso.

Para perfilizar organizações, soluções baseadas em UEBA fazem uso de técnicas de aprendizado de máquina não supervisionado (como agrupamento) de forma a tentar identificar desvios significativos nas atividades de usuários, dispositivos e redes, permitindo a detecção precoce de ameaças. Sua capacidade de construir perfis comportamentais ao longo do tempo e de se adaptar dinamicamente às mudanças do ambiente tem o potencial de aumentar a capacidade da detecção por anomalias e na implementação de medidas de segurança proativas, que diminuam o tempo de resposta a incidentes causados por adversários internos [3]. Entretanto, entre uma solução ideal e a potencial heterogeneidade dos usuários das

corporações, há um abismo que deve ser explorado: o da viabilidade de se perfilizar (e distinguir) adequadamente organizações via comportamento de seus usuários.

Nesse contexto, essa pesquisa foca agrupar usuários por suas características de uso de software, para responder as seguintes perguntas de pesquisa:

(PP1) É possível discernir entre diferentes corporações a partir do uso dos softwares dos seus usuários?

(PP2) Dado um conjunto de dados de treinamento com dados coletados de usuários de diversas empresas, é possível rotular um novo usuário como pertencendo ou não a determinada empresa a partir de seu uso de softwares?

Ao responder a pergunta PP1, desejamos identificar se é possível caracterizar empresas de acordo com os hábitos de utilização de software de seus funcionários. A pergunta PP2 tem um foco especial em segurança, já que se os funcionários de uma mesma corporação tiverem comportamentos de uso de software similares entre si, pode ser possível a criação de sistemas de alerta e log disparados ao detectar usuários que fogem a esse padrão (e.g., para identificar o acesso de pessoas não autorizadas ao ambiente da corporação).

Para responder a essa pergunta, utilizamos um conjunto de dados contendo a coleta de dados de 709 corporações distintas, e 58.769 usuários. Ao todo, foram gerados 407 GB de dados de coleta e 2.784.083.334 registros de atividade. A coleta se deu por meio de um aplicativo para o monitoramento da segurança desenvolvido por uma empresa privada.

Os resultados do estudo mostram que é possível perfilizar corporações a partir do perfil de uso de softwares de seus usuários, e que é possível classificar os usuários como trabalhadores ou não de uma empresa. No entanto, isso só é possível através de técnicas que podem aumentar o número de falsos negativos (intrusos não detectados). Dessa forma, existe um custo-benefício entre a geração de falsos alarmes e a detecção de intrusos reais.

O restante deste trabalho é estruturado da seguinte forma: na Seção 2 são apresentados os trabalhos relacionados. Na Seção 3 é apresentado o protocolo experimental, que inclui a coleta de dados e a criação dos vetores de características dos usuários. Os experimentos são discutidos na Seção 4 e, finalmente, na Seção 5, são apresentadas as conclusões e respostas para as perguntas de pesquisa.

2 Conceitos e Trabalhos Relacionados

UEBA é uma solução de cibersegurança que utiliza algoritmos de aprendizado de máquina para detectar anomalias no comportamento de usuários, dispositivos e sistemas dentro de redes corporativas [4]. Ao analisar desvios em relação a padrões comportamentais estabelecidos, a UEBA identifica atividades incomuns ou suspeitas, como aumentos repentinos no número de solicitações a servidores, indicando possíveis ataques de negação de serviço distribuídos (DDoS). Diferentemente de ferramentas tradicionais de monitoramento, que focam majoritariamente em atividades humanas, a UEBA estende sua análise para entidades como dispositivos e redes, oferecendo uma visão mais ampla das ameaças potenciais [3].

As soluções de UEBA se baseiam em três componentes principais: coleta e organização de dados, integração com sistemas de segurança existentes e geração de gatilhos acionáveis para respostas a incidentes detectados, como um usuário autenticado exibindo anomalia em seu perfil comportamental no sistema. Esses componentes permitem a detecção de uma gama maior de ameaças cibernéticas, pois sua atuação auxilia na identificação de ameaças sofisticadas que podem passar despercebidas por ferramentas convencionais de segurança.

Pfleeger e Caputo apresentaram em 2012 a justificativa para se incorporar a análise do comportamento humano em soluções de seguranças a fim de torná-las mais efetivas [5]. Em vez de se concentrar exclusivamente em abordagens tecnológicas, os autores defendem a integração do entendimento do comportamento humano para aumentar a eficácia das tecnologias e processos de segurança ao se levar em conta o impacto de aspectos como carga cognitiva, viés, heurísticas e modelos comportamentais.

Em 2015, a Gartner produziu um relatório cunhando o termo UEBA, que evoluiu do conceito de UBA - *User Behavior Analytics* para incluir a análise de dispositivos e redes, refletindo a crescente complexidade das ameaças cibernéticas modernas [6]. Essa evolução agregou a necessidade de se monitorar comportamentos interconectados, frequentemente indicativos de padrões de ataque não detectáveis por meio da análise isolada de usuários. Com o objetivo de ajudar a suprir essa lacuna, o presente trabalho tem por foco o monitoramento dos comportamentos interconectados, isto é, exibidos pela massa de usuários de uma determinada organização.

A partir da introdução da UEBA, começou-se a discutir mais amplamente as limitações inerentes a sistemas de proteção baseados em assinaturas, com foco nas dificuldades em se detectar ameaças um pouco mais complexas e ataques de *zero-day*. Para lidar com tais dificuldades, Salitin e Zolaiti [3] propõem que a análise comportamental seja combinada com técnicas de aprendizado de máquina para se monitorar padrões e detectar anomalias referentes a usuários e dispositivos, após levantar as tecnologias mais usadas para se implementar UEBA.

Na mesma linha, Khaliq et al. [7] exploram a aplicação de User and Entity Behavior Analytics (UEBA) para identificar ataques internos em organizações, uma vez que os métodos de segurança convencionais falham em prover visibilidade sobre as atividades dos usuários legítimos (frequentemente possuem maiores direitos de acesso e representam um risco considerável caso sejam comprometidos). Ao comparar técnicas de aprendizado de máquina (supervisionado e não supervisionado) aplicadas à UEBA, fica clara a lacuna na literatura sobre investigação dos dados brutos de comportamento de usuário. Por exemplo, os autores mencionam que abordagens supervisionadas, além de focar em ameaças conhecidas e requererem amostras rotuladas, podem ser aplicadas à detecção de spam, *phishing* e presença de arquivos maliciosos. Por outro lado, as abordagens não supervisionadas podem fazer emergir padrões que indiquem ameaças desconhecidas e permitam a detecção de anomalias e vazamento de dados, não necessitando de dados rotulados para treinamento. Entretanto, não é feita a análise da viabilidade em se perfilizar comportamentos de usuários dentro de organizações, ou mesmo perfilizar as organizações, como é o foco do presente trabalho.

Karamolegkos et al. [8] propõe criar uma plataforma para perfilizar usuários, agrupá-los e permitir a personalização de serviços. A partir da aplicação de KMeans e *Spectral Clustering* [9] em dados de mil usuários, os autores tentam atribuir palavras-chaves que

melhor representem as preferências dos usuários. Já JinHuaXu and HongLiu [10] aborda o agrupamento de usuários na Web por meio da monitoração do acesso à URLs via navegador. Para tanto, fazem uso de um *dataset* composto por 50 usuários e 8 *Web sites*, e aplicam o algoritmo KMeans para verificar se o agrupamento é factível e escalável. Embora os autores tenham dados reais e concluam que a técnica é viável, o conjunto de dados analisado é muito limitado, tanto em número de usuários quanto em abrangência de uso, isto é, poucos *sites*.

Zunair Ahmed Khan et al. [11] propõem um *framework* para classificar perfis de usuários entre normais e anômalos, usando informações como endereço IP, dados de localização, organização dos usuários, URLs, entre outras. O treinamento do modelo foi feito a partir dos dados do *Insider Threat Test Dataset* [12] e envolve combinar os resultados de quatro detectores por anomalia (email, URLs, Logon/Logoff e dispositivos/arquivos). Entretanto, o *dataset* contém apenas mil usuários e foi sinteticamente produzido, não refletindo situações imprevistas e outros ruídos que a coleta de dados reais pode trazer.

Mais recentemente, Yang et al. [13] desenvolveram um *pipeline* de análise de dados para tentar modelar e prever o efeito de *churn* (perda de clientes) usando *deep learning*. O *dataset* é composto por duas semanas de dados (em agosto de 2017) de usuários de uma única aplicação (*SnapChat*). Com isso, os autores conseguiram observar novos tipos de usuários na aplicação com base em suas atividades diárias e prever a evasão destes. Embora marginalmente relacionado, o artigo é dos poucos que analisam uma grande massa de dados reais e busca agrupar os usuários por tipos, preocupando-se com a interpretabilidade dos dados.

Em geral, nota-se que a literatura carece de: (i) explorar e entender os dados das organizações de forma a perfilizá-las com base em seus usuários; (ii) coleta de dados massiva, refletindo a míriade de usuários e organizações e seu impacto em soluções para proteção; (iii) dados e cenários realistas, para investigar os problemas de falsos-positivos que podem surgir por inúmeros fatores, incluindo a representação inadequada das amostras de treinamento. O escopo do presente artigo é verificar a viabilidade em se distinguir comportamentos de organizações com base no uso de aplicações por seus usuários, com uma massa de dados considerável coletada em ambientes reais de produção.

3 Protocolo Experimental

Nesta Seção é apresentado o protocolo experimental. Na Seção 3.1 são apresentadas as informações sobre como foi realizada a coleta e armazenamento de dados. Na Seção 3.2 são apresendados os subconjuntos de dados criados para os experimentos. Finalmente, na Seção 3.3, é apresentado o método utilizado para a definição dos vetores de características dos usuários.

3.1 Coleta e Armazenamento de Dados

Os dados utilizados nos experimentos foram coletados pelos autores deste trabalho através de um aplicativo de segurança instalado consentidamente nos sistemas dos usuários de empresas, com a finalidade de monitorar as atividades desses usuários em busca de padrões suspeitos. A principal função do aplicativo é coletar a aplicação sendo utilizada em primeiro plano pelo usuário (em *foreground*), e por quanto tempo essa aplicação foi utilizada. As informações armazenadas incluem o nome do aplicativo em primeiro plano, a versão do aplicativo, a duração do uso, e um identificador do usuário e da empresa. Os dados de identificador de usuário e empresa são anonimizados de maneira consistente para se poder criar rótulos que permitam verificar se usuários/empresas podem ser agrupados por similaridade de comportamento.

A coleta de dados ocorreu entre os dias 21 de Março até 30 de Setembro de 2024, totalizando mais de seis meses de registros.

Ao todo, foram coletados dados relativos ao uso de software de 58.769 usuários distintos, gerando um total de 407 GB de dados coletados que consistem de 2.784.083.334 registros de atividade armazenados banco de dados próprios. Os registros se referem a usuários de 709 empresas distintas e somam 36.985.354 horas de uso de aplicações. Foram capturados usos de 43.857 aplicações distintas. Muitas das aplicações eram instaladores e afins, o que aumentou muito o número de aplicações únicas. Essas aplicações foram desconsideradas na criação dos conjuntos de dados, como discutido na Seção 3.2.

A fim de possibilitar a análise da quantidade massiva de dados disponibilizada, um banco de dados relacional utilizando o PostgreSQL 14.11 foi modelado.

3.2 Separação dos dados

Para realizar os experimentos, foram gerados dois conjuntos a partir dos dados coletados. O primeiro conjunto, denominado *Top 50 Aplicações*, contém os registros de uso de software dos 50 softwares mais usados por todos os usuários do conjunto original, considerando todas as empresas. O segundo conjunto, denominado *Top 20 Empresas*, contém os registros de uso de software dos 50 softwares mais usados pelos funcionários das 20 empresas com mais funcionários, ambos calculados por tempo total de uso. Caso algum usuário nunca tenha usado nenhum dos 50 softwares que estão sendo analisados em um conjunto, ele não configura uma amostra viável para o experimento e, portanto, é descartado.

Dessa forma, o conjunto *Top 50 Aplicações* contém 2.227.123.758 registros de uso de software para 57.531 usuários diferentes, e o conjunto *Top 20 Empresas* contém 467.324.330 registros de uso de software para 13.269 usuários diferentes.

O limite de 50 softwares foi criado a fim de reduzir a dimensionalidade do problema, já que para cada usuário foi gerado um vetor de características considerando os softwares utilizados (detalhes na Seção 3.3). O limite de 20 empresas para o segundo conjunto foi criado para verificar se ao se considerar apenas empresas de grande porte, que contém muitos funcionários e, portanto, muitas amostras, é possível simplificar o processo de classificar os seus funcionários de acordo com seus hábitos de uso de software[1]. A Tabela 1 resume os dados dos conjuntos utilizados onde, para fins de comparação, também estão incluídos os dados do conjunto completo coletado.

A lista dos softwares mais utilizados, de acordo com os dados analisados, incluem os navegadores Google Chrome, Mozilla Firefox e Microsoft Edge, além de aplicações como o Windows Explorer, Microsoft Excel, e Microsoft Outlook. Um lista completa está disponível no Apêndice B.

[1]Juntas, as 20 maiores empresas representam cerca de 23% do conjunto de usuários monitorados

Tabela 1: Conjuntos criados.

Conjunto	# Horas de uso registradas	# Usuários	# Empresas
Completo	36.985.354	58.769	709
Top 50 Aplicações	19.727.895	57.531	709
Top 20 Empresas	3.943.009	13.269	20

3.3 Vetores de Características

Para cada usuário único foi gerado um vetor de características de 50 posições, onde cada posição do vetor representa um dos 50 softwares mais usados. Cada elemento do vetor é um número real indicando a média diária de uso realizada pelo usuário para cada um dos 50 softwares analisados naquele conjunto, indo do software mais utilizado no índice 0 do vetor até o menos utilizado no índice 49. Dessa forma, cada usuário é expresso em função do quanto usa, em média, os softwares analisados.

Um exemplo de vetor gerado para um usuário é dado na Figura 1, onde ao se considerar que o vetor foi gerado para o conjunto dos 50 softwares mais usados (Apêndice B), é possível notar que o usuário usou, em média, o aplicativo Google Chrome por 1.1 horas, o Windows Explorer por 0.2 horas, o Microsoft Excel por 3.7 horas, e o navegador Microsoft Edge por 0 horas.

índice	0	1	2	3	…
valor	1.1	0.2	3.7	0	…

Figura 1: Exemplo de vetor de características.

4 Experimentos

Nesta Seção são exibidos os experimentos que visam responder as perguntas de pesquisa deste trabalho. Inicialmente, na Seção 4.1, são analisados resultados qualitativos de clusters de usuários. Já nas Seções 4.2 e 4.3 constam experimentos quantitativos, onde são analisadas as acurácias de classificadores capazes de identificar as empresas de usuários desconhecidos (problema multiclasses), e se um usuário desconhecido trabalha ou não em uma empresa (problema binário).

4.1 Análise de Clusters

Inicialmente, foi realizada uma análise qualitativa do agrupamento dos usuários (*clustering*). Espera-se que, caso os comportamentos de uso de software dos usuários de uma mesma companhia sejam similares, esses usuários apareçam de maneira agrupada.

Como os vetores de características possuem 50 dimensões, é utilizado o algoritmo *T-distributed Stochastic Neighbor Embedding* (TSNE) [14] para prover um mapa bidimensional que permite a visualização dos dados. Na Figura 2, é mostrado o gráfico gerado pelo TSNE considerando o conjunto de dados que contém as 20 maiores empresas[2]. Cada ponto no gráfico indica um usuário, e as cores indicam empresas distintas. Como pode ser observado, usuários de uma mesma empresa tendem a aparecer próximos uns dos outros, apesar de muitas vezes não formarem *clusters* únicos.

A fim de colocar a hipótese do agrupamento de usuários da mesma empresa à prova, executou-se o algoritmo de agrupamento K-means [15] considerando $K = 20$ no mesmo conjunto de dados para verificar os *clusters* formados (nesse caso, foram passados apenas os vetores de características, sem os rótulos das empresas para o algoritmo). O resultado pode ser observado na Figura 3. Ao comparar as Figuras 2 e 3, que utilizam os rótulos reais, e os gerados pelo K-means, respectivamente, pode-se observar algumas similaridades entre os *clusters*. No entanto, é visível a discrepância em muitos dos *clusters*, possivelmente devido ao fato dos usuários de empresas se concentrarem em "ilhas".

A formação de múltiplos *clusters* como ilhas de usuários para uma mesma empresa, como os circulados com linhas contínuas na Figura 2, era esperada, uma vez que em uma mesma companhia, usuários que trabalham em áreas distintas podem ter comportamentos distintos, mas similares entre os seus pares que trabalham na mesma área. Por exemplo, em uma montadora de automóveis, espera-se que os usuários que trabalham na área de engenharia e projetos utilizem um conjunto limitado de softwares, enquanto usuários que trabalham com recursos humanos utilizem outro conjunto de softwares.

Por motivo semelhante, podem existir *clusters* onde usuários de múltiplas companhias podem estar presentes, como o circulado com linhas pontilhadas na Figura 2. A principal hipótese para isso acontecer é que usuários que trabalham em uma mesma função (ex.: com recursos humanos), mas em companhias diferentes, compartilham comportamentos similares de uso de software. No entanto, não podemos validar essa hipótese com os dados angariados, já que a versão atual do aplicativo de coleta de dados não fornece informações sobre a função exercida pelo usuário.

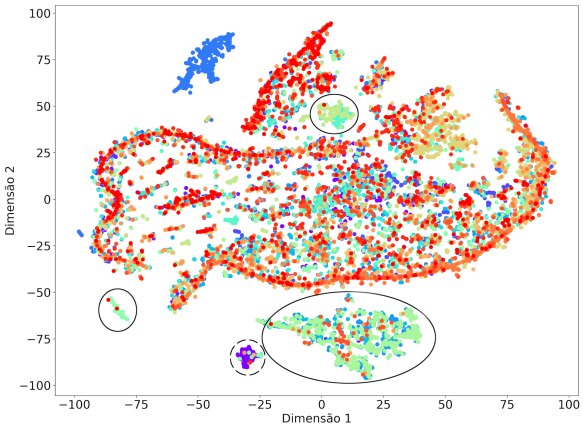

Figura 2: TSNE do conjunto *Top 20 Empresas*.

4.2 Classificação de usuários por empresas

Os resultados obtidos na Seção 4.1 mostram que parte dos usuários de determinadas empresas podem acabar por formar múltiplas

[2]Para esta análise desconsideramos o conjunto que contém todas as empresas devido ao fato de existirem 709 empresas distintas, dificultando a visualização do mapa bidimensional.

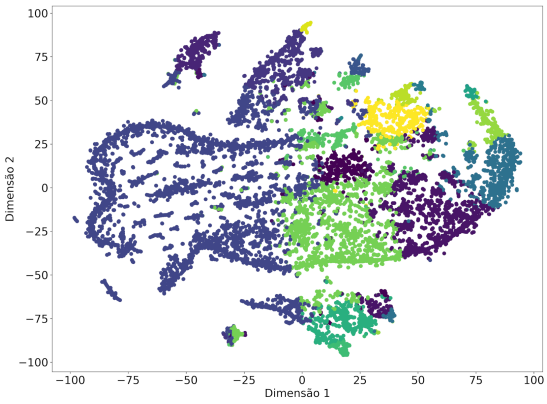

Figura 3: TSNE dos clusters calculados via k-means no conjunto *Top 20 Empresas*.

ilhas de *clusters*. Sendo assim, nesta Seção é analisada a eficácia de um classificador do tipo k Vizinhos Mais Próximos – *k-nearest Neighbors* (k-NN). Para os experimentos, foi selecionado o valor de $k = 17$ (no apêndice A são exibidos os experimentos utilizados para definir k). O princípio desse experimento é que, caso os funcionários de uma companhia formem grupos com características semelhantes, será possível categorizar corretamente um novo usuário na companhia em que ele realmente trabalha.

Categorizar o usuário como pertencente ou não a determinada empresa pode levar a soluções de segurança baseadas em anomalias. Ao se perfilizar o comportamento padrão em uma certa empresa, pode-se criar sistemas de alertas que, por exemplo, sinalizam que determinado usuário tem um comportamento dissimilar aos demais, avisando os gestores sobre potencial violação de segurança (por exemplo, acesso não autorizado) a ser investigada.

Para realizar os experimentos desta Seção, inicialmente separamos os dados em dois conjuntos disjuntos. Um de treinamento, e um de testes. Para criar os conjuntos de forma a remover possíveis vieses, separamos de maneira aleatória 50% dos usuários únicos de cada empresa para fazer parte do conjunto de treinamento, e os 50% restantes para fazer parte do conjunto de testes. Dessa forma, é possível verificar se, dado um grupo de usuários reais de uma companhia, é possível verificar se um usuário desconhecido se encaixa no mesmo padrão de comportamento dos seus pares, indicando que ele pertence a mesma companhia, ou não (nesse caso, um alerta poderia ser gerado, indicando que o usuário não se encaixa no padrão de comportamento dos seus colegas).

Os resultados estão dispostos na Tabela 2, onde são mostradas as acurácias considerando os conjuntos de dados *Top 20 Empresas* e o conjunto *Top 50 Aplicações* (onde constam todas as empresas). Na Tabela também constam resultados referentes a um modelo Top 10, onde primeiramente o k-NN é usado para computar a empresa mais provável a que o usuário pertence. Feito isso, todos os dados de treinamento dessa empresa são removidos do k-NN, e o classificador é usado mais uma vez, agora para computar a segunda empresa mais provável. O processo é repetido, até que as 10 empresas mais prováveis sejam computadas. Nesse caso, se a empresa em que o usuário realmente trabalha aparece nas Top 10 empresas geradas, é considerado um acerto. Caso contrário, um erro.

Tabela 2: Acurácias obtidas.

Conjunto	Acurácia (stdev)	Acurácia Top 10 (stdev)
Top 50 Aplicações	15,6% (0,1)	37,9% (0,2)
Top 20 Empresas	36,6% (0,6)	85,6% (0,5)

Os resultados mostram uma acurácia de 15,6% considerando o conjunto que contém todas as empresas (Top 50 Aplicações). Isso mostra que os funcionários possuem comportamentos de uso similares entre si, já que considerando que existem 709 empresas distintas no conjunto, um classificador desinformado teria uma acurácia de apenas 0,14% (assumindo que as quantidades de funcionários das empresas é aproximadamente igual). A acurácia aumenta consideravelmente, para 36,6%, quando considerado o conjunto que contém apenas as 20 maiores empresas. Apesar desse comportamento ser esperado, já que o problema se torna mais simples devido ao menor número de empresas, deve-se também levar em consideração que essas empresas possuem muitos funcionários nos conjuntos de treinamento, o que pode ser relevante para gerar melhores resultados para o algoritmo k-NN.

Levando em conta os resultados Top 10, pode-se observar um aumento considerável nos valores de acurácia, chegando a 85,6% para o conjunto das Top 20 empresas. Esse resultado mostra que, geralmente, a empresa correta está no rol das empresas mais prováveis geradas pelo classificador. Esse achado pode levar a sistemas de alerta que consideram uma baixa probabilidade de falsos alarmes ao, por exemplo, gerar um alerta para os responsáveis apenas se o usuário em questão não for classificado como pertencendo à empresa corrente em nenhuma das 10 (ou N) empresas mais prováveis.

4.3 Testes de Sistema de Alertas

Nesta Seção são apresentados os resultados considerando um sistema que indica se, dado um usuário x, se ele é classificado como sendo um usuário real da empresa, ou como um possível intruso. Como discutido anteriormente, esse tipo de classificador pode ser utilizado para, por exemplo, gerar alertas para gestores informando possíveis anomalias a serem investigadas. Para os experimentos desta Seção, os mesmos conjuntos de treino e testes definidos na Seção 4.2 foram utilizados. Os resultados são reportados como matrizes de confusão, onde, dada uma empresa e:

- Verdadeiro Positivo (VP) – usuário que trabalha na empresa e foi classificado como tal.
- Falso Positivo (FP) – usuário que não trabalha na empresa foi classificado como um trabalhador de e.
- Falso Negativo (FN) – usuário que trabalha na empresa e foi classificado como não sendo um trabalhador de e.
- Verdadeiro Negativo (VN) – usuário que não trabalha na empresa e foi classificado como não sendo um trabalhador de e.

Para computar os valores de VP, FP, FN e VN, considerando todas as companhias presentes nos conjuntos de dados, foi utilizado o Algoritmo 1. No algoritmo, toda vez que o classificador é

testado para uma empresa e (laço na linha 4), são selecionados do conjunto de testes te todos os usuários que realmente trabalham na empresa para para fazer parte do experimento, sendo esses os usuários positivos. Para balancear os conjuntos, uma quantidade de usuários que não trabalham na empresa (usuários negativos) são selecionados aleatoriamente, de forma que o conjunto gerado tenha o mesmo tamanho do de usuários positivos (linha 5 do Algoritmo).

Algoritmo 1: TESTA_CLASSIFICADOR(C, tr, te).

```
Input: C: classificador; tr: conjunto de treinamento; te:
       conjunto de dados de testes.
Resultado: VP, FP, FN e VN
1  treinar(C, tr)// Treinar o classificador
   // x recebe a lista de empresas individuais no
      conjunto
2  E = lista_empresas(tr)
3  VP = FP = FN = VN = 0
4  para cada empresa e ∈ E faça
       // remove aleatoriamente pessoas que não
          trabalham na empresa para balancear.
5      tb = balancear(te)
       // x é o vetor, e y a classe
6      para cada tupla (x, y) ∈ tb faça
           // ŷ é a empresa que o classificador acha
              que x pertence
7          ŷ = classificar(x, C)
8          se y == e então
               // O funcionário trabalha em e
9              se ŷ == e então
10                 VP = VP + 1
11             senão
12                 FN = FN + 1
13         senão
               // O funcionário não trabalha em e
14             se ŷ == e então
15                 FP = FP + 1
16             senão
17                 VN = VN + 1
```

Os resultados para os conjuntos Top 20 Empresas são exibidos na matriz de confusão da Tabela 3, onde e e $\overline{e}$ indicam se o usuário trabalha ou não na empresa e. Como pode ser observado, apesar das altas taxas de Verdadeiros Positivos VP e Verdadeiros Negativos VN, há também uma alta taxa de Falsos Negativos FN, o que pode ser um impeditivo para a implantação de um sistema que sinaliza a possibilidade de um invasor – nesse caso, muitos falsos negativos indicam que o sistema constantemente indica que um funcionário idôneo é um invasor, gastando recursos e gerando desconforto entre os usuários (de maneira parecida a um antivírus que constantemente coloca em quarentena softwares benignos).

No entanto, ao se considerar que o usuário é um funcionário da empresa e, se e aparece em qualquer posição na lista das Top 10 empresas mais prováveis do classificador (de maneira semelhante ao experimento da Seção 4.2), o número de Falsos Negativos FN diminui consideravelmente, como pode ser observado na Tabela 4. De maneira similar, mostramos os resultados para o classificador Top 10 considerando todas as 709 empresas na Tabela 5. Considerar o classificador Top 10, como esperado, aumenta também consideravelmente a chance de um Falso Positivo FP (um usuário que não trabalha em e confundido com um trabalhador de e). No entanto, cabe notar que comumente sistemas de alarme precisam focar no custo-benefício entre reduzir o número de falsos alarmes (i.e., não atrapalhar gestores e funcionários idôneos), mantendo uma quantidade relevante de alarmes reais). Considerando a abordagem Top 10, 3.311 dos 6.639 funcionários que não eram trabalhadores das empresas testadas foram detectados (cerca de 50%). Uma análise similar pode ser feita para a Tabela 5, no entanto, o alto valor de FP, possivelmente devido ao grande número de empresas envolvidas, indica que valores maiores de N para a abordagem Top N podem ser necessários.

Tabela 3: Matriz de Confusão do conjunto *Top 20 Empresas* com classificador k-NN

	Predito e	Predito $\overline{e}$
e	$VP = 2.398 = 18,1\%$	$FN = 4.241 = 31,9\%$
$\overline{e}$	$FP = 245 = 1,8\%$	$VN = 6.394 = 48,2\%$

Tabela 4: Matriz de Confusão do conjunto *Top 20 Empresas* com classificador k-NN Top 10

	Predito e	Predito $\overline{e}$
e	$VP = 5.670 = 42,7\%$	$FN = 969 = 7,3\%$
$\overline{e}$	$FP = 3.328 = 25,1\%$	$VN = 3.311 = 24,9\%$

Tabela 5: Matriz de Confusão do conjunto *Top 50 Aplicações* com classificador k-NN Top 10

	Predito e	Predito $\overline{e}$
e	$VP = 11079 = 19.1\%$	$FN = 17882 = 30.9\%$
$\overline{e}$	$FP = 1931 = 3.3\%$	$VN = 27030 = 46.7\%$

5 Conclusão

A fim de verificar a possibilidade de perfilizar organizações a partir dos comportamentos de uso de softwares de seus empregados, neste trabalho coletamos seis meses de registros de dados de uso de software de 58.769 usuários distintos, os quais trabalhavam em 709 empresas. A análise dos dados focou em duas perguntas de pesquisa principais:

PP1: É possível discernir entre diferentes corporações a partir do uso dos softwares dos seus usuários? Sim, mesmo em cenários onde o classificador precisava discernir se determinado usuário trabalhava em uma dentre todas as 709 empresas possíveis, uma acurácia de 15,6% foi atingida (um classificador desinformado teria uma acurácia de apenas 0,14%). Apesar do resultado mostrar

que os comportamentos de uso de softwares pelos seus empregados pode ser utilizado como uma informação para perfilizar a empresa, fica claro que essa informação sozinha pode não ser suficiente, já que existe a possibilidade de empresas que, por exemplo, operam no mesmo ramo, possuírem perfis de uso de software semelhantes.

Dado um conjunto de dados de treinamento com dados coletados de usuários de diversas empresas, é possível rotular um novo usuário como pertencendo ou não a determinada empresa a partir de seu uso de softwares? Sim, apesar de ser necessária uma abordagem Top N (Top 10 nos experimentos), onde a quantidade de falsos negativos (um usuário que não trabalha na empresa é confundido com um funcionário da empresa). Nesse caso, ao utilizar uma abordagem baseada em Top N, estamos buscando um balanço entre a quantidade de falsos alarmes (funcionários idôneos confundidos com invasores) e a quantidade de alarmes reais (um usuário que não pertence à equipe da empresa).

Os resultados deste trabalho abrem portas para diversas pesquisas futuras. Por exemplo, a hipótese de os classificadores podem estar se confundindo com empresas distintas, mas que operam no mesmo segmento, pode ser testada coletando-se informações sobre os segmentos de trabalho das empresas. A hipótese de que funcionários de uma mesma área, independentemente da empresa em que trabalham, têm comportamento semelhante de uso de softwares pode ser testada coletando-se informações sobre as funções exercidas pelos usuários. Como uma implementação futura, será analisada a possibilidade de coleta desses dados. Além disso, outras estratégias de classificação poderão ser testadas no futuro, como o treinamento de classificadores do tipo *one-class* individuais para cada empresa, a fim de implementar sistemas de alarme.

Acknowledgments

Este trabalho foi apoiado pela Bluepex CyberSecurity via financiamento de projeto de Inovação da Base Industrial de Defesa – Edital MD/MCTI/FINEP/FNDCT 2022.

Referências

[1] Cong Shi, Jian Liu, Hongbo Liu, and Yingying Chen. Wifi-enabled user authentication through deep learning in daily activities. *ACM Trans. Internet Things*, 2(2), May 2021. doi: 10.1145/3448738. URL https://doi.org/10.1145/3448738.

[2] Rahma Olaniyan, Sandip Rakshit, and Narasimha Rao Vajjhala. Application of user and entity behavioral analytics (ueba) in the detection of cyber threats and vulnerabilities management. In Prasenjit Chatterjee, Dragan Pamucar, Morteza Yazdani, and Dilbagh Panchal, editors, *Computational Intelligence for Engineering and Management Applications*, pages 419–426, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-19-8493-8.

[3] Manya Ali Salitin and Ali Hussein Zolait. The role of user entity behavior analytics to detect network attacks in real time. In *2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pages 1–5, 2018. doi: 10.1109/3ICT.2018.8855782.

[4] Pierpaolo Artioli, Antonio Maci, and Alessio Magrì. A comprehensive investigation of clustering algorithms for user and entity behavior analytics. *Frontiers in Big Data*, 7, 2024. ISSN 2624-909X. doi: 10.3389/fdata.2024.1375818. URL https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1375818.

[5] Shari Lawrence Pfleeger and Deanna D. Caputo. Leveraging behavioral science to mitigate cyber security risk. *Computers & Security*, 31(4):597–611, 2012. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2011.12.010. URL https://www.sciencedirect.com/science/article/pii/S0167404811001659.

[6] Avivah Litan. User and entity behavior analytics (ueba) expands security beyond user monitoring. https://www.gartner.com/en/documents/3621357, 2015. Acessado em 06/12/2024.

[7] Salman Khaliq, Zain Ul Abideen Tariq, and Ammar Masood. Role of user and entity behavior analytics in detecting insider attacks. In *2020 International Conference on Cyber Warfare and Security (ICCWS)*, pages 1–6, 2020. doi: 10.1109/ICCWS48432.2020.9292394.

[8] Pantelis N. Karamolegkos, Charalampos Z. Patrikakis, Nikolaos D. Doulamis, and Elias Z. Tragos. User - profile based communities assessment using clustering methods. In *2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications*, pages 1–6, 2007. doi: 10.1109/PIMRC.2007.4394637.

[9] Ulrike von Luxburg. A tutorial on spectral clustering. *CoRR*, abs/0711.0189, 2007. URL http://dblp.uni-trier.de/db/journals/corr/corr0711.html#abs-0711-0189.

[10] JinHuaXu and HongLiu. Web user clustering analysis based on kmeans algorithm. In *2010 International Conference on Information, Networking and Automation (ICINA)*, volume 2, pages V2–6–V2–9, 2010. doi: 10.1109/ICINA.2010.5636772.

[11] Muhammad Zunair Ahmed Khan, Muhammad Mubashir Khan, and Junaid Arshad. Anomaly detection and enterprise security using user and entity behavior analytics (ueba). In *2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS)*, pages 1–9, 2022. doi: 10.1109/ICONICS56716.2022.10100596.

[12] Brian Lindauer. Insider Threat Test Dataset. 9 2020. doi: 10.1184/R1/12841247.v1. URL https://kilthub.cmu.edu/articles/dataset/Insider_Threat_Test_Dataset/12841247.

[13] Carl Yang, Xiaolin Shi, Luo Jie, and Jiawei Han. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 914–922, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219821. URL https://doi.org/10.1145/3219819.3219821.

[14] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[15] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

A Testes dos valores de k para o k-NN

Na Tabela são exibidos os experimentos para selecionar o valor de k para o algoritmo k-NN, utilizando o conjunto de dados Top 20 Empresas. Foram testados para k números primos entre 3 e 20.

Tabela 6: Experimentos para o valor de k.

k	Acurácia
3	34.3%
5	35.8%
7	36.2%
11	36.3%
13	35.8%
17	36.4%
19	36.0%

B Softwares Analisados nos Conjuntos

Na Tabela 7 é exibida a listagem dos 50 aplicativos utilizados para gerar os vetores de características das Top 50 aplicações, e das Top 50 aplicações considerando apenas as 20 maiores empresas. As colunas de índice indicam o índice em que a aplicação aparece em cada um dos vetores (top 50 aplicações ou top 20 maiores empresas). De maneira similar, as colunas de Horas de uso mostram a contagem total de horas computadas para esses softwares, considerando todos os usuários, para os vetores das top 50 aplicações ou top 20 maiores empresas. Alguns softwares foram anonimizados para preservar a identidade das empresas.

Tabela 7: Softwares analisados nos conjuntos *Top 50 Aplicações* e *Top 20 Empresas*

Índice vetor Top 50 Apl.	Índice vetor Top 20 Empr.	Nome do Aplicativo	Descrição	Horas de Uso Top 50 Apl.	Horas de Uso Top 20 Empr.
0	0	chrome.exe	Google Chrome	7.233.796	1.361.929
1	1	explorer.exe	Windows Explorer	3.161.645	459.927
2	4	excel.exe	Microsoft Excel	1.438.572	277.411
3	3	msedge.exe	Microsoft Edge	1.368.190	292.400
4	7	outlook.exe	Microsoft Outlook	921.036	111.299
5	2	mstsc.exe	Conexão de Área de Trabalho Remota	837.291	295.166
6	5	smartclient.exe		486.992	210.291
7	9	winword.exe	Microsoft Word	448.913	56.973
8	8	ms-teams.exe	Microsoft Teams (work or school)	327.555	88.524
9	10	whatsapp.exe		310.982	54.146
10	6	navegadorsankhya.exe	NavegadorSankhya	304.525	197.041
11	13	firefox.exe	Firefox Developer Edition	240.794	39.480
12	12	simnext.exe	SIMNext	192.350	39.490
13	17	onedrive.exe	Microsoft OneDrive	167.094	27.531
14	14	acrobat.exe	Adobe Acrobat	157.482	33.478
15	-	soulmv_navegador.exe	Cent Browser	127.089	-
16	11	appcontroller.exe	AppController	124.585	45.664
17	15	soffice.bin	LibreOffice	123.170	32.979
18	18	shellexperiencehost.exe	Windows Shell Experience Host	117.496	20.268
19	-	sig_integrado.exe	*anônimo*	111.572	-
20	20	skype.exe	Skype	100.579	18.152
21	19	powerpnt.exe	Microsoft PowerPoint	94.545	18.371
22	31	javaw.exe	OpenJDK Platform binary	85.735	8.622
23	26	thunderbird.exe	Thunderbird	71.699	10.740
24	23	calculatorapp.exe	Calculator	62.437	11.926
25	41	opera.exe	Opera Internet Browser	58.437	6.450
26	29	a2start.exe	Emsisoft Security Center	57.383	9.878
27	22	olk.exe	Microsoft Outlook	57.303	13.081
28	27	notepad.exe	Bloco de notas	54.332	10.254
29	40	gg-client.exe	GraphOn GO-Global Connection	54.206	6.468
30	16	acad.exe	AutoCAD Application	52.310	27.951
31	-	presto.mes.exe	PRESTO | Manufacturing Execution System	50.851	-
32	-	srcnet.exe	Apresentação Gráfica do Sistema SERCON	50.523	-
33	39	anydesk.exe	AnyDesk	49.655	6.702
34	-	*anônimo*	*anônimo*	49.395	-
35	-	autcom.exe	Autcom	49.122	-
36	21	teams.exe	Microsoft Teams	46.575	15.634
37	-	msaccess.exe	Microsoft Access	43.646	-
38	-	ceprod.exe	CeProd	42.462	-
39	-	pcsis4116.exe		41.349	-
40	-	sgci.exe		40.665	-
41	-	mmc.exe		40.146	-
42	-	gerente.exe		39.079	-
43	-	paf.exe	Frente de Caixa (PAF-ECF)	38.743	-
44	-	vhf.exe	Visual Hotal FrontOffice	38.514	-
45	-	society2010.exe	Society ERP - Associados	34.736	-
46	24	wps.exe	WPS Office	30.967	11.454
47	46	searchapp.exe	Search application	30.937	4.731
48	-	clinux.exe		30.273	-
49	-	sghpep.exe	Prontuário Eletrônico do Paciente	30.165	-
-	25	copg.exe		-	11.247
-	28	rm.exe		-	9.915
-	30	zoom.exe	Zoom Meetings	-	9.330
-	32	wd-desk v2.exe		-	8.260
-	33	aecontabxe10.exe	AEContabXe10	-	7.931
-	34	xtop.exe	Creo 8.0.4.0 from PTC	-	7.700
-	35	acrord32.exe	Adobe Acrobat Reader	-	7.590
-	36	istagui.exe	ISTAGui	-	7.085
-	37	revit.exe	Autodesk Revit	-	7.035
-	38	sga.exe		-	6.780
-	42	viewer.exe		-	6.404
-	43	atxexec.exe		-	5.450
-	44	ivms-4200.framework.c.exe		-	5.121
-	45	startmenuexperien-cehost.exe	Windows Start Experience Host	-	4.792
-	47	code.exe	Visual Studio Code	-	4.679
-	48	setcanhotocontrolese-nha.exe	SetcanhotoControleSenha	-	4.650
-	49	pbidesktop.exe	Microsoft Power BI Desktop	-	4.630