eduroamIA: Inteligência Artificial Aplicada à Previsão de Eventos de Autenticação Federada no eduroam*

Lucas R. Frank[1], **Rodrigo T. Rego**[1], **Calebe Reis**[2], **Luciano Rocha**[2], **Edelberto Franco Silva**[1]

[1]Laboratório NetLab
Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Departamento de Ciência da Computação (DCC)
Universidade Federal de Juiz de Fora (UFJF)
Juiz de Fora/MG – Brasil

[2]Diretoria Adjunta de Gestão de Serviços
Rede Nacional de Ensino e Pesquisa (RNP)
Brasília/DF – Brasil

Abstract. *This work evaluates the use of critical event prediction in a large-scale federated identity management service, eduroam. The purpose of this research is to indicate to the administrator about possible anomalies related to user authentication and institution authentication behavior. Machine learning algorithms are applied - both offline and online. Thus, with real logs from RADIUS service at the eduroam federation level of roaming users, it was possible to determine an prediction model for the profile of each institution. the results show promising progress in the area of big data analysis for the identity management environment in a wireless federated environment with a RMSE of 3.66 in the bestcase.*

Resumo. *Este trabalho avalia a utilização de predição de eventos críticos em um serviço de gestão de identidade federada em larga escala, o eduroam. O objetivo desta pesquisa é indicar ao administrador sobre possíveis anomalias relacionadas à autenticação de usuários e comportamento de autenticação das instituições. Algoritmos de aprendizado de máquina são aplicados - tanto em modo offline quanto online. A partir dos registros reais do serviço RADIUS no nível da federação eduroam de usuários em roaming foi possível determinar um modelo de predição para o perfil de cada instituição. Os resultados mostram promissor avanço na área de análise de dados de grande massa para a área de gestão de identidade em um ambiente federado sem fio com RMSE de 3.66 no melhor caso.*

1. Introdução

As redes sem fio IEEE 802.11 são os principais meio de acesso dos usuários finais, nos mais diversos ambiente e setores [Medeiros 2019]. Os dados de gerenciamento e monitoramento dessas redes contém dados que podem gerar conhecimento sobre características

*Este trabalho faz parte do projeto de mesmo nome selecionado na chamada do Desafio RNP e Microsoft em Inteligência Artificial. Os autores agradecem todo o suporte e financiamento.

relevantes do usuário ou da própria instituição envolvida. Um exemplo de utilização dessas redes institucionais globais de acesso sem fio é o projeto eduroam [Saade et al. 2013], que está presente em 101 países e que já realizou mais de 1 bilhão de autenticações. Com este volume de clientes fica claro que, além dos dados gerados pelas tentativas de acesso/autenticação, a saúde do serviço de *roaming* federado deve ser prioridade.

Por sua vez, tratar, processar e extrair informações úteis dos dados coletados são necessárias técnicas automáticas, estatísticas e de inteligência computacional. Portanto, é fundamental identificar quais estratégias são mais apropriadas para serem utilizadas no processamento de grande massa de dados [Lopez et al. 2018].

Assim, neste trabalho, serão apontadas técnicas e ferramentas desenvolvidas para as predições de eventos críticos no serviço eduroam. O foco esta na identificação de um modelo aplicável para este fim no projeto eduroamIA da chamada Desafio Microsoft e RNP.

2. Proposta

A maioria dos produtos de *software* gera registros/*logs* que podem ser usados para análise da causa e para identificação de uma possível solução de problemas. Embora esses registros ofereçam informações úteis sobre o desempenho em tempo real, é difícil (e muitas vezes inviável) analisá-los manualmente. Seus dados, em geral, não têm estrutura e, muitas vezes, não contêm informações analíticas suficientes.

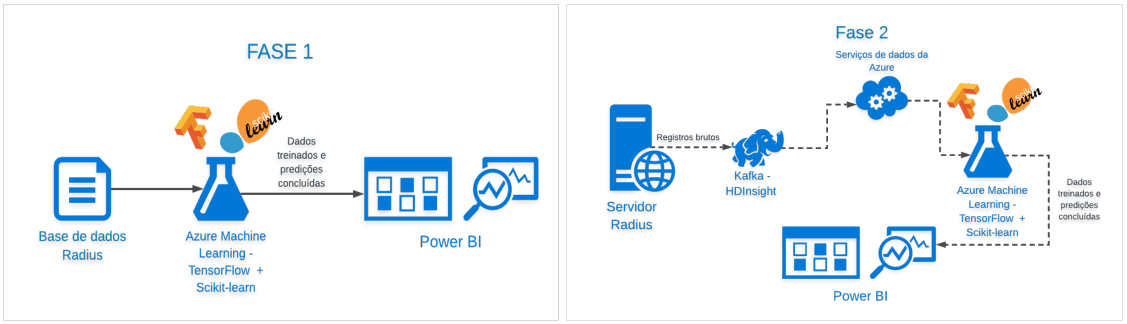
Um caminho é pensar em uma maneira eficiente de detectar anomalias desenvolvendo uma solução tecnológica que aprenda os padrões complexos e responda pela sazonalidade com uma melhor precisão do que os sistemas manuais. Como contextualização, é através da entrada de registros do serviço RADIUS no nível da federação eduroam de usuários em *roaming*, que se faz possível traçar um perfil de cada instituição e identificar um comportamento anômalo ou falhas no serviço. Por exemplo, caso uma instituição esteja rejeitando autenticações a mais do que seu padrão, é interessante gerar uma alerta ao administrador do sistema.

A intenção da avaliação proposta por esse trabalho é gerar uma interface em que seja claro para administrador do serviço verificar a saúde do mesmo com relação às autenticações e gerar relatórios de utilização, assim como a automatização de alertas. Para que os modelos da proposta sejam avaliados é necessário o tratamento e análise de registros, onde tem-se o volume atual de dados gerados diariamente somente por um servidor eduroam no nível da federação em torno de 10GB.

As figuras 1(a) e 1(b) mostram as propostas de integração dos serviços Microsoft e a solução de gestão de identidade para utilização de inteligência artificial como detecção de comportamentos anômalos no ambiente federado do serviço eduroam. Na figura 1(a) tem-se desde a coleta, pré-processamento e análise de dados a partir da solução Azure *Machine Learning* com Scikit-learn[1], gerando relatórios e alertas ao administrador com relação à análise realizada. Já a Figura 1(b) mostra a proposta de análise em fluxo de dados em tempo real, esta ainda em avaliação no escopo desta pesquisa. Nesta figura vê-se a coleta de dados em tempo real sendo tratada a partir de um serviço de *streaming* de dados, como o Apache Kafka (presente no arcabouço do *HDInsight* da Microsoft),

[1] `https://scikit-learn.org/stable/`

que por sua vez permite o armazenamento em nuvem da Microsoft Azure e os passos de análise, predição e visualização (como já demonstrado pela figura 1(a)).



(a) (b)

Figura 1. (a) Proposta para a fase 1 e análise em lote de forma offline dos dados de registros; (b) Proposta para a fase 2 e análise em streaming de forma online dos dados de registros.

Desta forma, o trabalho atual realiza a implementação e validação da chamada fase 1 do projeto, exposta pela figura 1(a).A proposta contribui ao estado da arte para análise de dados no ambiente de serviços de gestão de identidade, como veremos na próxima seção.

3. Resultados

Os resultados demonstram os avanços obtidos no desenvolvimento do projeto eduroamIA, assim como as métricas associadas ao modelos de aprendizados de máquina e regressão. Para análise foi utilizada uma massa de dados reais coletados do serviço eduroam no período de 1 ano. A implementação utilizou a linguagem Python ofertada pelo serviço Microsoft Machine Learning Studio na nuvem Microsoft Azure. O projeto ainda entrega uma interface de geração de relatório e análise de resultados baseada na ferramenta PowerBI da Microsoft.

A utilização do Power BI, possibilita a comparação e análise dos dados preditos pelo Azure Machine Learning Studio, transformando os resultados preditos e dados analisados em uma importante ferramenta de monitoramento com relação a quantidade total de tentativas de autenticações com sucesso (*login* aceitos) e sem sucesso (rejeitados pelo servidor). Nesse ambiente já é possível, ainda sem utilização de nenhuma técnica de inteligência computacional, verificar qual instituição está rejeitando mais do que aceitando autenticações, como vê-se no link [2]

Com relação à predição para definição de padrões para cada instituição, foram selecionados os algoritmos de árvore de decisão e rede neural, que são modelos supervisionados. O primeiro é um modelo mais simples, que tem como objetivo criar árvores que funcionam em forma de uma fluxograma. Já o segundo, é utilizado o conceito de neurônios sendo baseado em um *Perceptron* de Multicamadas (MLP). Como nossos dados é uma série temporal, utilizamos as seguintes características para os modelos: dados históricos(dado um horário, consideramos pontos que aconteceram no mesmo horário no passado) e pontos recentes(se queremos prever 6 pontos a frente, consideramos os 5

[2]https://ibb.co/LzwdSpc

anteriores). Para uma boa avaliação dos modelos, foram utilizadas as métricas Raiz do Erro Quadrático Médio (RMSE) e Coeficiente de Determinação (R^2). O RMSE tem como objetivo dizer o quanto o valor predito varia em volta de seu valor médio, sendo assim, quanto menor este valor, melhor o modelo. Já o R^2 indica o quanto o modelo consegue explicar os valores observados, variando entre 0 e 1, apontando um bom modelo o quanto mais próximo de 1. Nos links (usp.br)[3] e pucrs.br[4], tem-se a comparação entre os resultados obtidos pelos modelos após treinamento. Nesses links é possível ver os gráficos comparativos com relação as métricas dos modelos avaliados. Tais métricas se referem a mostrar o quão próximo o valor predito está do valor real. Considerando a métrica RMSE, pode-se ver na tabela 1 que os valores na média não se distanciam significativamente do valor esperado, como destacado para `pucrs.br`.

	Arvore de Decisão		Rede Neural	
Instituições	RMSE	R^2	RMSE	R^2
pucrs.br	**3.51**	**0.87**	**3.66**	**0.86**
usp.br	5.86	0.29	5.48	0.38

Tabela 1. Resultados obtidos para as métricas RMSE e R^2 para 2 instituições.

4. Conclusões e Trabalhos Futuros

Como mencionado, esse trabalho está dividido em duas fases, onde na primeira fase tem-se como objetivo tratar os registros, realizar o pré-processamento da base de dados e treinar o modelo de predição. Nessa etapa foi identificado o modelo que melhor se adaptou ao perfil de cada instituição, de acordo com as métricas usadas. Com o modelo já treinado, incia-se a segunda fase, utilizando tal modelo para predizer em tempo real se há alguma instituição com problemas[5]. Tais resultados mostram relevância do ponto de vista da gestão de identidade e de seus serviços associados, identificando os problemas do administrador de redes federadas eduroam, assim como do usuário de tal federação.

Referências

Lopez, M. A., Sanz, I. J., Lobato, A. G., UFF, D. M. M., and Duarte, O. C. M. (2018). Aprendizado de máquina em plataformas de processamento distribuído de fluxo: Análise e detecção de ameaças em tempo real. *Minicursos do Simpósio Brasileiro de Redes de Computadores-SBRC*, 2018:150–206.

Medeiros, D. S. V.; Cunha Neto, H. N. L. M. A. M. L. C. S. S. E. F. V. A. B. . F. N. C. . M. D. M. F. (2019). *Analise de Dados em Redes Sem Fio de Grande Porte: Processamento em Fluxo em Tempo Real, Tedências e Desafios*, chapter 4, pages 142–195. Minicursos / XXXVII Simposio Brasileiro de Redes de Computadores e Sistemas Distribuidos, 1st edition.

Saade, D. C. M., Carrano, R. C., Silva, E. F., and Magalhães, L. (2013). Eduroam: Acesso sem fio seguro para a comunidade acadêmica federada. *Rede Nacional de Ensino e Pesquisa*.

[3]https://ibb.co/gyZCxYM

[4]https://ibb.co/jTXD8v0

[5]Sejam problemas no tratamento das autenticações, seja na disponibilidade do serviço de Gestão de Identidade associado, o eduroam.