Fundação Universidade Federal do ABC
Pró reitoria de pesquisa
Av. dos Estados, 5001, Santa Terezinha, Santo André/SP, CEP 09210-580
Bloco L, 3ºAndar, Fone (11) 3356-7617
iniciacao@ufabc.edu.br

Projeto de Iniciação Científica submetido
para avaliação no Edital: 01/2024

Título do projeto: Avaliação de abordagens de ajuste de hiperparâmetros em dinâmica da digitação

Palavras-chave do projeto: biometria; aprendizado de máquina; ajuste de hiperparâmetros

Área do conhecimento do projeto: Ciências Exatas e da Terra; Ciência da Computação; Metodologia e Técnicas da Computação

Resumo

O uso de biometria para autenticação de usuários, em oposição a métodos tradicionais baseados em senhas estáticas, têm sido atrativo em razão da maior segurança que um sistema biométrico pode proporcionar. Os usuários podem ser reconhecidos com base em características físicas ou comportamentais ao invés de precisarem se lembrar de uma senha ou a ter em mãos um cartão inteligente ou *token*. Dentre as diversas modalidades biométricas existentes, há a *dinâmica de digitação*, que reconhece as pessoas com base em seu ritmo de digitação. Diversos algoritmos de classificação podem ser usados nesse contexto. Esses algoritmos possuem hiperparâmetros, que precisam ser ajustados adequadamente. O ajuste de hiperparâmetros pode ser realizado para cada usuário individualmente ou pode ser global, isso é, os hiperparâmetros assumirão o mesmo valor para todos os usuários. O objetivo deste projeto é explorar diferentes abordagens de ajuste de hiperparâmetros em algoritmos para classificação para dinâmica de digitação.

Sumário

1 Introdução

Nos últimos anos, a popularização de serviços oferecidos única e exclusivamente pela da Internet tem crescido de forma extraordinária. Muitas pessoas realizam pagamentos em sites de lojas diretamente de suas casas e gerenciam suas contas bancárias sem a necessidade de irem até os bancos (**??**). Dessa forma, surge a necessidade sistemas de autenticação e identificação cada vez mais sofisticados.

Entretanto, na perspectiva de **??**), a maioria dos sistemas de autenticação comumente usados são baseados em senhas estáticas, ou uma combinação entre uma senha e um cartão inteligente ou *tokens*. **??**) destacam que essa abordagem possui uma série de limitações. Por exemplo, senhas compostas por palavras comuns, sequências numéricas ou alfanuméricas simples que sejam passíveis de memorização são consideradas "fracas" no que se refere ao nível de segurança que elas oferecem, pois podem ser descobertas por ataques do tipo "força bruta".

??) ainda afirmam que é recomendado que os usuários criem senhas maiores e compostas por uma combinação de números, letras e caracteres especiais para serem menos vulneráveis a ataques. Em adição a isso, as senhas deveriam ser diferentes para cada conta nos sites em que acessam. Todavia, muitas vezes os usuários ainda optam pela utilização de uma mesma senha para todas as suas contas, o que gera um risco de segurança: se apenas um dos sistemas for invadido e tiver os seus dados roubados, o acesso aos dados do indivíduo em todos os outros serviços estará comprometido.

Segundo **??**), muitas dessas limitações associadas ao uso de senhas podem ser contornadas pela incorporação de métodos de autenticação melhores. Nesse sentido, a biometria, como uma forma de estabelecer a identidade por meio de características físicas ou comportamentais dos indivíduos, surge como uma alternativa. Há diversas modalidades biométricas (**??**).

Neste projeto, será dado o enfoque na vertente comportamental, em particular, na *dinâmica de digitação* (**??**). Essa modalidade biométrica envolve analisar a forma com que um indivíduo digita em um teclado (**??**). Os usuários são então reconhecidos com base no seu ritmo de digitação.

Conforme definido em (**??**), sistemas biométricos são sistemas de reconhecimento de padrões que extraem características de dados biométricos e então comparam as características extraídas com uma referência biométrica em um banco de dados (**??**). Essa comparação das características extraídas a partir dos dados biométricos com a referência biométrica no banco de dados frequentemente resulta em uma pontuação (*score*) (**??**). Assumindo que seja uma pontuação indicando a similaridade, a classificação pode ser realizada aplicando um limiar de corte (*threshold*). Se a pontuação for maior que o limiar, o dado biométrico é classificado como genuíno e, caso contrário, como sendo de um impostor.

Nesse contexto, o limiar de corte pode ser entendido com um hiperparâmetro. O trabalho de (**??**) menciona o limiar de corte como um hiperparâmetro em algoritmos que retornam uma pontuação ou uma probabilidade. Outros algoritmos usados para reconhecimento de usuários pela dinâmica da digitação podem possuir outros hiperparâmetros para serem ajustados. Os valores dos hiperparâmetros tem impacto importante no desempenho preditivo de um sistema biométrico.

O objetivo deste projeto é comparar diferentes abordagens para ajuste de hiperparâmetros em algoritmos de classificação para dinâmica de digitação. Sobre esse aspecto, há algumas questões que podem ser investigadas. A primeira é sobre realizar o ajuste de forma global ou individualizada para cada usuário no sistema biométrico. Em biometria, o limiar de corte pode ser um valor global comum para todos os usuários ou um valor específico para cada usuário (**????**). Esse conceito pode ser estendido para outros hiperparâmetros, determinando os valores de forma comum a todos os usuários ou de forma individualizada. Outra questão que pode ser estudada neste contexto é sobre a técnica de ajuste de hiperparâmetros.

As demais seções do projeto estão organizadas da seguinte forma: na Seção **??**, são introduzidos conceitos sobre dinâmica da digitação e ajuste de hiperparâmetros; na Seção **??**, são apresentados os objetivos; na Seção **??**, é descrita a metodologia; na Seção **??**, a viabilidade do projeto é discutida; e, na Seção **??**, é apresentado o cronograma deste projeto.

2 Dinâmica da digitação e ajuste de hiperparâmetros

Esta seção apresenta alguns conceitos importantes para este projeto de pesquisa envolvendo dinâmica da digitação e ajuste de hiperparâmetros.

2.1 Dinâmica da digitação

A dinâmica da digitação é um modalidade biométrica comportamental que diferencia os indivíduos com base em atributos característicos da digitação de textos, como o tempo em que o indivíduo permanece pressionando cada tecla, o intervalo de tempo entre cada ativação de tecla e padrões nos erros de digitação (**??**).

O interesse na dinâmica de digitação como uma forma de autenticação biométrica remonta o início dos anos de 1980, quando a *National Science Foundation* (NSF), agência do governo federal dos Estados Unidos da América, patrocinou um experimento conduzido pela *RAND Corporation* em que alguns digitadores profissionais tiveram que digitar um texto fixo enquanto o intervalo de tempo de ativação entre cada par de teclas era medido. O mesmo procedimento foi repetido 4 meses depois, com os mesmos digitadores e o mesmo texto. As análises estatísticas dos dados obtidos mostraram que era possível distinguir de forma satisfatória os digitadores com base nos padrões observados nas métricas obtidas, que serviam como uma "assinatura" de cada digitador (**??**).

Segundo **??**), a principal vantagem da dinâmica da digitação em relação às demais modalidades biométricas comportamentais é a sua transparência. Por exemplo, se usada em conjunto com um formulário de autenticação comum, composto por um identificador de usuário (ou nome de usuário, *username*) e uma senha, as métricas de digitação podem ser obtidas das informações que o indivíduo necessariamente deverá inserir no sistema que ele deseja acessar. Além disso, em serviços baseados na Web, muitas vezes não é viável exigir formas de autenticação por biometria, pois os usuários podem não ter acesso aos dispositivos necessários, como câmeras e sensores de impressões digitais, ou equipamentos mais sofisticados.

Para fins de reconhecimento biométrico, os dados podem ser obtidos por meio dos padrões de digitação tanto de *textos fixos* definidos previamente, que os usuários serão

requisitados a digitar para fins de identificação ou verificação, ou de *textos livres*, sem um tamanho fixo ou qualquer outra restrição. Muitos dos estudos desenvolvidos sobre autenticação pela dinâmica de digitação consideram um mecanismo baseado em texto fixo, geralmente o nome de usuário e a senha coletados previamente. Entretanto, diversos pesquisadores também aplicam algoritmos de aprendizado de máquina para desenvolver modelos capazes de autenticar os usuários de forma contínua por meio do texto digitado em um sistema durante o seu uso (**??**).

2.2 hiperparâmetros

De acordo com **??**), desenvolver um modelo de aprendizado de máquina que seja efetivo na resolução de um determinado problema é uma tarefa complexa e demorada. Ela envolve a escolha do algoritmo apropriado e a obtenção de um modelo arquitetural ótimo por meio do ajuste de hiperparâmetros. O autor explica que há dois tipos de parâmetros em modelos de aprendizado de máquina: os *parâmetros* do próprio modelo, que serão inicializados e repetidamente atualizados durante o processo de treinamento, e os chamados *hiperparâmetros*, que devem ser escolhidos antes de o modelo ser treinado. **??**) menciona os pesos dos neurônios em redes neurais como um exemplo de parâmetro de modelo e o parâmetro de penalidade C em uma *Support Vector Machine* (SVN), a taxa de aprendizado em redes neurais e o algoritmo utilizado para minimizar a função objetivo como hiperparâmetros.

O ajuste de hiperparâmetros é o processo de testar valores diferentes para os hiperparâmetros a fim de se obter o melhor ajuste para um modelo construído a partir de determinada base de dados. Na perspectiva de **??**), o ajuste de hiperparâmetros é uma parte fundamental da construção de modelos de aprendizado de máquina efetivos, especialmente em redes neurais artificiais e modelos baseados em árvores de decisão, que possuem diversos hiperparâmetros. **??**) explica que problemas de otimização de hiperparâmetros (*hyperparameter optimization*, HPO) exigem um entendimento profundo da relação entre as combinações de hiperparâmetros e o modelo de aprendizado de máquina resultante do processo de treinamento. Ambos dependem do algoritmo utilizado e do tipo de cada hiperparâmetro, que pode ser contínuo, discreto ou categórico.

Segundo **??**), após a escolha do algoritmo de aprendizado de máquina e dos métodos que serão utilizados para avaliar o seu desempenho, é necessário listar os hiperparâmetros que deverão ser ajustados e, então, definir os conjuntos de valores possíveis para cada um de acordo com o seu tipo. Dependendo do problema, os hiperparâmetros podem possuir restrições, isto é, não poderão assumir qualquer valor dentre todos os valores possíveis, e essas restrições impostas a um hiperparâmetro podem estar condicionadas aos valores escolhidos para outro. Além disso, para cada configuração diferente, o modelo deverá ser treinado e testado novamente, para que o seu desempenho seja medido.

Alguns motivos para aplicar técnicas de ajuste de hiperparâmetros em aprendizado de máquina são evitar a necessidade de realizar o ajuste manualmente, o aprimoramento do desempenho dos algoritmos, assim como a melhora da reprodutibilidade e justiça dos estudos realizados **????**). A próxima seção discute algumas questões sobre o ajuste de hiperparâmetros no contexto de dinâmica da digitação, que será o foco deste projeto.

2.3 Ajuste de hiperparâmetros no contexto de dinâmica da digitação

Diversos trabalhos na área de dinâmica da digitação acabam não aplicando uma técnica de ajuste de hiperparâmetros em razão da métrica usada para reportar os resultados. Isso ocorre, por exemplo, ao reportar resultados em termos de EER (*Equal Error Rate*). Ao ajustar o limiar de corte de um sistema biométrico, as taxas de falsa aceitação (impostores aceitos erroneamente) e de falsa rejeição (usuários genuínos rejeitados de forma indevida) podem mudar. De maneira geral, ao aumentar uma taxa, a outra diminui dependendo do ajuste do limiar de corte. O valor EER representa o ajuste em que as duas taxas são iguais (**??**). Para isso, os rótulos de teste (genuíno/impostor) podem ser usados para encontrar esse ajuste. Entretanto, em uma aplicação prática, o acesso aos rótulos dos dados pode não estar disponível.

Alguns trabalhos avaliaram o impacto dos hiperparâmetros. No trabalho elaborado por **??**), a técnica *Grid Search* foi aplicada para encontrar a melhor combinação de hiperparâmetros para uma implementação do algoritmo *Support Vector Machine* (SVM) usado para reconhecimento de usuários pela dinâmica da digitação. Os estudos realizados

por **??**) avaliaram diferentes arquiteturas de redes neurais e a influência de hiperparâmetros como número de filtros convolucionais, tamanho do *kernel* de convolução, número de neurônios na camada recursiva e taxa de drop out.

Uma discussão sobre ajustar o limiar de corte de forma individual e de forma global foi realizada por **??**), assim como também avaliou a adaptação de modelos ao longo do tempo. De fato, em dinâmica da digitação, o ritmo de digitação pode mudar com o tempo. Outro trabalho que avaliou o ajuste de hiperparâmetros em dinâmica da digitação foi o de **??**). Nesse trabalho, foi considerado um cenário de sistemas biométricos adaptativos (**??**), em que a referência biométrica pode ser atualizada ao longo do tempo. O trabalho de **??**) avaliou uma proposta para adaptar o limiar de corte ao longo do tempo. Os mesmos autores também discutiram essa adaptação do limitar de corte em (**??**).

3 Breve descrição dos objetivos e metas

Este projeto tem o objetivo de **comparar diferentes abordagens para ajuste de hiperparâmetros de algoritmos de classificação em dinâmica da digitação**. Para isso, será avaliado o ajuste de hiperparâmetros de forma individual e global. Além disso, diferentes técnicas de ajuste de hiperparâmetros podem ser investigadas nesse contexto.

A princípio, o foco do projeto será na dinâmica de digitação de *texto fixo*, em que todos os indivíduos digitam a mesma expressão. O desempenho será avaliado por métricas como FMR, FNMR e acurácia balanceada, descritas na Seção **??**.

Os objetivos específicos do projeto são:

- Selecionar algoritmos de classificação usados em dinâmica da digitação;
- Definir abordagens de ajuste de hiperparâmetros que possam ser aplicadas em dinâmica da digitação (por exemplo: ajuste individual, ajuste global);
- Realizar experimentos comparando as diferentes abordagens de ajuste;
- Avaliar desempenho obtido pelos algoritmos com cada abordagem.

4 Metodologia

Este projeto irá comparar técnicas para ajuste de hiperparâmetros de algoritmos de classificação em dinâmica da digitação. Para isso, serão utilizados conjuntos de dados disponíveis publicamente, conforme descrito na Seção **??**. Esses dados serão divididos entre treino e teste, sendo que as amostras usadas para treinamento serão referentes a dados mais antigos em comparação com os dados usados para teste. Na Seção **??**, são descritas métricas que serão usadas para avaliação de desempenho neste trabalho.

4.1 Conjuntos de dados

Grande parte dos trabalhos que realizaram experimentos com dados de dinâmica da digitação não disponibilizaram os dados coletados (**??**). Esse fato dificulta a reprodutibilidade de estudos na área. Este projeto irá utilizar dados publicamente disponíveis. Alguns conjuntos de dados que podem ser usados são descritos a seguir:

- CMU (**??**): Este conjunto de dados [1] possui dados de 51 indivíduos que digitaram a senha ".tie5Roanl" em oito sessões de captura, com 50 amostras em cada sessão. No total, cada indivíduo digitou a senha 400 vezes.

- KeyRecs (**??**): O conjunto de dados KeyRecs [2] envolveu a captura de dinâmica da digitação de texto fixo e de texto livre. A princípio, o foco deste projeto será em texto fixo, portanto apenas essa parte do conjunto de dados deve ser utilizada. Para texto fixo, de acordo com a descrição do conjunto de dados, 99 indivíduos digitaram uma mesma senha em duas sessões, com 100 amostras em cada sessão, totalizando 200 amostras por indivíduo. Ao realizar o download da versão disponível, entretanto, observou-se que alguns usuários tem menos do que 200 amostras.

Além desses conjuntos de dados, o projeto pode eventualmente usar outros conjuntos de dados como, por exemplo, os conjuntos de dados GREYC (**??**) e também o disponibilizado por **??**).

[1] <https://www.cs.cmu.edu/~keystroke/>
[2] <https://zenodo.org/records/7886743>

4.2 Métricas

Esta seção descreve algumas métricas usadas na literatura que serão usadas para avaliação dos resultados nos experimentos realizados neste projeto de pesquisa. Essas métricas são: FMR, FNMR e acurácia balanceada (**????**). Uma breve descrição dessas métricas é apresentadas a seguir:

- FMR (*False Match Rate*, Taxa de falsa correspondência): percentual de tentativas de impostores que foram aceitas como genuínas, definida como

$$FMR = \frac{numero\ de\ tentativas\ de\ impostores\ aceitas}{total\ de\ tentativas\ de\ impostores}. \quad (1)$$

  Uma taxa relacionada é a FAR (*False Acceptance Rate*), que tem significado similar, mas considera também taxa em que o sistema biométrico falha ao obter uma amostra biométrica. Essa taxa é conhecida como FTA (*Failure to Acquire Rate*).

- FNMR (*False Non-match Rate*, Taxa de falsa não-correspondência): percentual de tentativas genuínas que foram rejeitadas como impostoras pelo sistema, definida como

$$FNMR = \frac{numero\ de\ tentativas\ genuinas\ rejeitadas}{total\ de\ tentativas\ de\ usuarios\ genuinos}. \quad (2)$$

  Uma métrica relacionada é a FRR (*False Rejection Rate*), que tem um significado similar, mas considera também a FTA.

- Acurácia balanceada: média do acerto para cada classe (genuíno e impostor). Essa métrica pode ser obtida a partir da o cálculo da (HTER - *Half Total Error*, Metade do erro total)

$$HTER = \frac{FNMR + FMR}{2}, \quad (3)$$

  definida como a média entre FNMR e FMR (**??**). A partir da HTER, então é obtida a acurácia balanceada, definida como

$$BAcc = 1 - HTER. \tag{4}$$

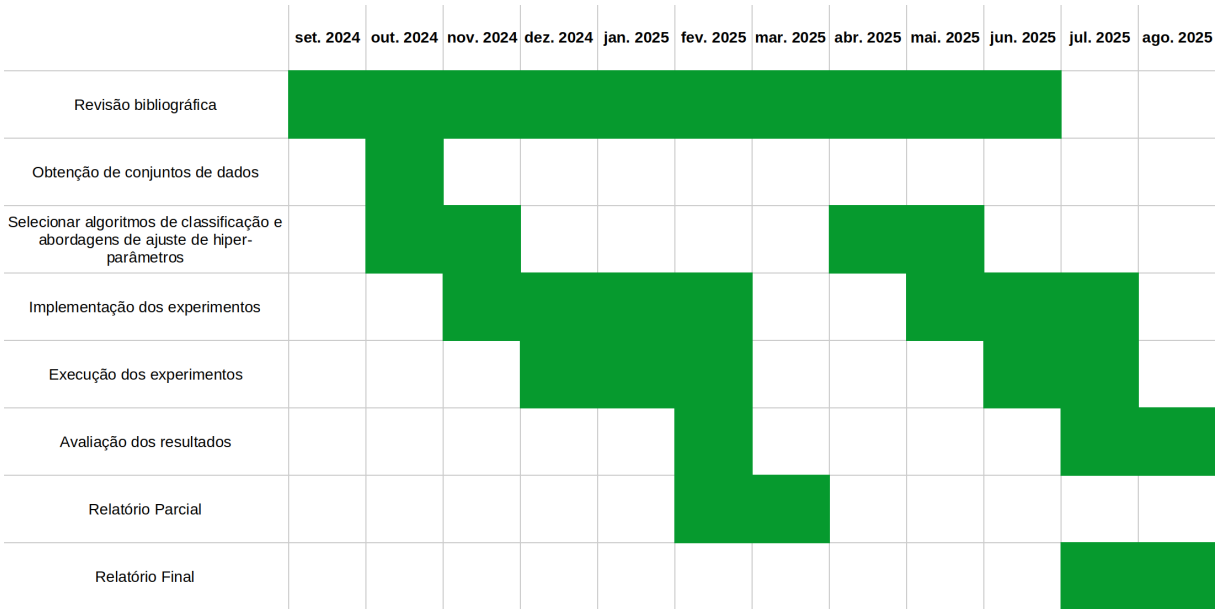
5 Descrição da viabilidade da execução do projeto

Para a realização da pesquisa bibliográfica neste projeto, será necessário acesso a artigos científicos. O Portal de Periódicos da CAPES disponível pela rede da UFABC pode ser utilizado para isso. Com relação aos experimentos, há conjuntos de dados de dinâmica da digitação disponíveis publicamente, conforme apresentado na Seção **??**. Além disso, para a implementação e execução dos experimentos, um computador/notebook é suficiente. Sobre as atividades do projeto, o cronograma para condução desta pesquisa é descrito na próxima seção.

6 Cronograma

O cronograma do projeto, dividido em 12 meses, é apresentado na Figura **??**.

Figura 1 – Cronograma do projeto.



As tarefas do cronograma são brevemente descritas a seguir:

- *Revisão bibliográfica*: pesquisa de trabalhos relacionados a dinâmica da digitação e ajuste de hiperparâmetros;
- *Selecionar algoritmos de classificação e abordagens de ajuste de hiperparâmetros*: definir quais algoritmos de aprendizado de máquina serão utilizados na classificação dos dados de dinâmica da digitação, assim como as abordagens de ajuste de hiperparâmetros que serão avaliadas;
- *Obtenção de conjuntos de dados*: obtenção de conjuntos de dados para os experimentos. Alguns conjuntos de dados que podem ser utilizados são mencionados na Sessão **??**;
- *Implementação dos experimentos*: implementação do código para realizar os experimentos;
- *Execução dos experimentos*: execução dos experimentos com diferentes abordagens de ajuste de hiperparâmetros;

- *Avaliação dos resultados*: avaliação dos resultados obtidos nos experimentos;

- *Relatório Parcial*: elaboração do relatório parcial;

- *Relatório Final*: elaboração do relatório final.

Além das atividades descritas no cronograma, artigos científicos poderão ser escritos e submetidos.